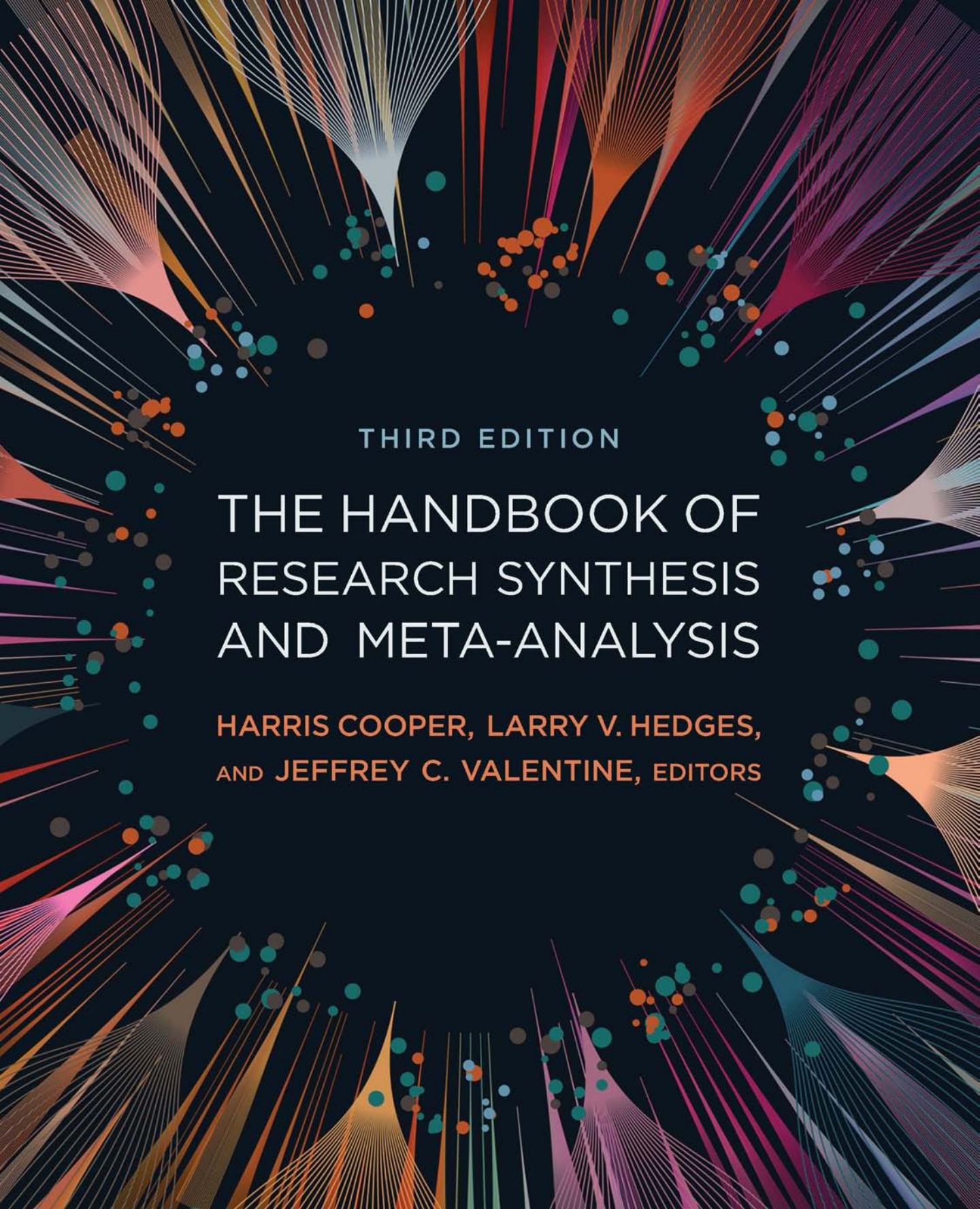
THIRD EDITION
THE HANDBOOK OF
RESEARCH SYNTHESIS
AND META-ANALYSIS
HARRIS COOPER, LARRY V. HEDGES,
AND JEFFREY C. VALENTINE, EDITORS

HANDBOOK OF RESEARCH SYNTHESIS AND META-ANALYSIS
3RD EDITION

THE HANDBOOK OF RESEARCH SYNTHESIS AND META-ANALYSIS

3RD EDITION

Edited by

HARRIS COOPER, LARRY V. HEDGES,

AND JEFFREY C. VALENTINE

RUSSELL SAGE FOUNDATION NEW YORK

THE RUSSELL SAGE FOUNDATION

The Russell Sage Foundation, one of the oldest of America's general purpose foundations, was established in 1907 by Mrs. Margaret Olivia Sage for "the improvement of social and living conditions in the United States." The foundation seeks to fulfill this mandate by fostering the development and dissemination of knowledge about the country's political, social, and economic problems. While the foundation endeavors to assure the accuracy and objectivity of each book it publishes, the conclusions and interpretations in Russell Sage Foundation publications are those of the authors and not of the foundation, its trustees, or its staff. Publication by Russell Sage, therefore, does not imply foundation endorsement.

Library of Congress Cataloging-in-Publication Data

Names: Cooper, Harris M., editor. | Hedges, Larry V., editor. | Valentine, Jeff C., editor.
Title: Handbook of research synthesis and meta-analysis / edited by Harris Cooper, Larry V. Hedges, and Jeffrey C. Valentine.
Description: 3rd edition. | New York : Russell Sage Foundation, [2019] | Includes bibliographical references and index.
Identifiers: LCCN 2018047695 (print) | LCCN 2018052201 (ebook) | ISBN 9781610448864 (ebook) | ISBN 9780871540058 (pbk. : alk. paper)
Subjects: LCSH: Research—Methodology—Handbooks, manuals, etc. | Information storage and retrieval systems—Research—Handbooks, manuals, etc. | Research—Statistical methods—Handbooks, manuals, etc.
Classification: LCC Q180.55.M4 (ebook) | LCC Q180.55.M4 H35 2019 (print) | DDC 001.4/2—dc23
LC record available at https://lccn.loc.gov/2018047695

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials. ANSI Z39.48-1992.

RUSSELL SAGE FOUNDATION
112 East 64th Street, New York, New York 10065

10 9 8 7 6 5 4 3 2 1

CONTENTS

ABOUT THE AUTHORS

Ariel M. Aloe is associate professor of educational measurement and statistics at the University of Iowa.

Betsy Jane Becker is Mode L. Stone Distinguished Professor of Educational Statistics in the College of Education at Florida State University.

Michael Borenstein is director of research at Biostat Inc., in Englewood, New Jersey.

Kathleen Coburn is a postdoctoral fellow at the University of California, Merced.

Thomas D. Cook is Joan and Serepta Harrison Professor Emeritus of Ethics and Justice at Northwestern University and research professor at the Trachtenberg School of Public Policy at George Washington University.

Harris Cooper is Hugo L. Blomquist Professor of Psychology and Neuroscience at Duke University.

Dean Giustini is biomedical librarian at the University of British Columbia Faculty of Medicine.

Julie Glanville is associate director of the York Health Economics Consortium where she manages information retrieval and reviewing projects.

Sean Grant is assistant professor in the Department of Social and Behavioral Sciences at Indiana University Richard M. Fairbanks School of Public Health.

Larry V. Hedges is Board of Trustees Professor of Statistics and Education and Social Policy at Northwestern University.

Julian P. T. Higgins is professor of evidence synthesis at the University of Bristol, United Kingdom.

Spyros Konstantopoulos is professor of measurement and quantitative methods at Michigan State University.

Huy Le is professor in the Department of Management at the University of Texas at San Antonio.

Mark W. Lipsey is research professor at the Peabody Research Institute and Department of Human and Organizational Development at Vanderbilt University.

Georg E. Matt is professor of psychology at San Diego State University.

Evan Mayo-Wilson is associate scientist in the Center for Clinical Trials and Evidence Synthesis, Department of Epidemiology at Johns Hopkins Bloomberg School of Public Health.

In-Sue Oh is Charles E. Beury Professor of Human Resource Management at Temple University.

Robert G. Orwin is senior study director with Westat, Inc.

Terri D. Pigott is associate provost for research at Loyola University Chicago.

Frank L. Schmidt is Fethke Leadership Professor Emeritus in the Tippie College of Business at the University of Iowa.

Alexander Sutton is professor of medical statistics at the University of Leicester, United Kingdom.

Rebecca M. Turner is a senior research fellow at University College London, United Kingdom.

Jeffrey C. Valentine is professor and program coordinator of the Educational Psychology, Measurement, and Evaluation Program at the University of Louisville.

Jack L. Vevea is associate professor of quantitative psychology at the University of California, Merced.

Howard D. White is professor emeritus at the College of Computing and Informatics, Drexel University.

David B. Wilson is professor of criminology at George Mason University.

Sandra Jo Wilson is principal associate in the Social and Economic Policy Division at Abt Associates.

Nicole A. M. Zelinsky is a doctoral candidate in quantitative psychology at the University of California, Merced.

PART I

INTRODUCTION

1

RESEARCH SYNTHESIS AS A SCIENTIFIC PROCESS

HARRIS COOPER
Duke University

LARRY V. HEDGES
Northwestern University

JEFFREY C. VALENTINE
University of Louisville

CONTENTS

> It is necessary, while formulating the problems of which in our further advance we are to find the solutions, to call into council the views of those of our predecessors who have declared an opinion on the subject, in order that we may profit by whatever is sound in their suggestions and avoid their errors.
>
> Aristotle, *De Anima*, Book 1, Chapter 2

1.1 INTRODUCTION

From the moment we are introduced to science, we are told it is a cooperative, cumulative enterprise. Like the artisans who construct a building from blueprints, bricks, and mortar, scientists contribute to a common knowledge structure. Theorists provide the blueprints and researchers collect the data that are the bricks. To extend the analogy further yet, we might say that research synthesists are the bricklayers and hodcarriers of the science guild. It is their job to stack the bricks according to the blueprints and apply the mortar that allows the structure to take shape.

Anyone who has attempted a research synthesis is entitled to a wry smile as the analogy continues. They know that several sets of theory-blueprints often exist, describing structures that vary in form and function, with no a priori criteria for selecting between them. They also know that our data-bricks are not all six-sided with right angles. They come in a baffling array of sizes and shapes. Making them fit, securing them with mortar, and seeing whether the resulting construction looks anything like what the blueprint suggests is a challenge worthy of the most dedicated, inspired artisan.

1.1.1 Replication and Research Synthesis

Scientific literatures are cluttered with repeated studies of the same phenomena. Multiple studies on the same problem or hypothesis arise because investigators wish to verify and extend (that is, generalize or search for influences on) previous findings. Experience has shown that even when considerable effort is made to achieve direct replication, results across studies are rarely identical at any high level of precision (Valentine et al. 2011; Open Science Collaboration 2015), even in the physical sciences (Hedges 1987). No two bricks are exactly alike. Still, the value and need for replication in the social sciences has received increased attention recently, due in part to concerns about questionable data practices, such as selective reporting of findings. For example, the journal *Perspectives on Psychological Science* published a special issue on replication (Pashler and Wagonmakers 2012).

How should scientists proceed when study results differ? First, it is clear how they should *not* proceed. They should not decide that results are not replicated simply because some results reject the null hypothesis and the others do not, in part because the outcome of null hypothesis significance tests does not imply a difference in effect size (Gelman and Stern 2006). Differences in statistical power might explain this, as well as expected sampling variation. Even results suggesting that the relation of interest is in different directions are predictable, depending on the size of the underlying effect, its sensitivity to contextual variation, and the number of times it has been tested. Certainly, scientists should not decide that one study (perhaps the most recent one, or the one they conducted, or a study chosen via some other equally arbitrary criterion) produces the correct finding and others can be ignored. If results that are expected to be similar show variability, the scientific instinct should be to account for the variability by further systematic work. This is where research synthesis comes in.

1.2 RESEARCH SYNTHESIS IN CONTEXT

1.2.1 A Definition of the Literature Review

The American Psychological Association's *PsycINFO* reference database defines a literature review as "the process of conducting surveys of previously published material" (http://psycnet.apa.org/psycinfo/1994-97192-000). Common to all definitions of literature reviews is the notion that they are "not based primarily on new facts and findings, but on publications containing such primary information, whereby the latter is digested, sifted, classified, simplified, and synthesized" (Manten 1973, 75).

Table 1.1 presents a taxonomy of literature reviews that capture six distinctions that review authors use to describe their own work (Cooper 1988). The taxonomy can be applied to literature reviews appearing throughout a broad range of both the behavioral and physical sciences. The six features and their subordinate categories permit a rich level of distinction among works of synthesis.

Table 1.1 A Taxonomy of Literature Reviews

Characteristic	Categories
Focus	Research findings
	Research methods
	Theories
	Practices or applications
Goal	Integration
	Generalization
	Conflict resolution
	Linguistic bridge-building
	Criticism
	Identification of central issues
Perspective	Neutral representation
	Espousal of position
Coverage	Exhaustive
	Exhaustive with selective citation
	Representative
	Central or pivotal
Organization	Historical
	Conceptual
	Methodological
Audience	Specialized scholars
	General scholars
	Practitioners or policy makers
	General public

SOURCE: Cooper 1988. Reprinted with permission from Transaction Publishers.

The first distinction among literature reviews concerns the *focus* of the review, the material that is of central interest to the reviewer. Most literature reviews center on one or more of four areas: the findings of individual primary studies, not necessarily but often empirical in nature; the methods used to carry out research; theories meant to explain the same or related phenomena; and the practices, programs, or treatments being used in an applied context.

The second characteristic of a literature review is its *goals*. Goals concern what the preparers of the review hope to accomplish. The most frequent goal for a review is to *integrate* past literature that is believed to relate to a common topic. Integration includes formulating general statements that characterize multiple specific instances (or research, methods, theories, or practices); resolving conflict between contradictory research results, ideas, or statements of fact by proposing a new conception that accounts for the inconsistency; and bridging the gap between concepts or theories by creating a new, common linguistic framework.

Another goal for literature reviews can be to *critically analyze the existing literature*. Unlike a review that seeks to integrate the existing work, one that involves a critical assessment does not necessarily summate conclusions or compare the covered works one with another. Instead, it holds each work up against a criterion and finds it more or less acceptable. Most often, the criterion will include issues related to the methodological quality of empirical studies; the logical rigor, completeness, or breadth of explanation if theories are involved; or comparison with the ideal treatment, when practices, policies, or applications are involved.

A third goal that often motivates literature reviews is to *identify issues central to a field*. These issues may include questions that have given rise to past work, questions that should stimulate future work, or methodological problems or problems in logic and conceptualization that have impeded progress within a topic area or field.

Of course, reviews more often than not have multiple goals. So, for example, it is rare to see a review that integrates or critical examines existing work without identifying central issues for future endeavors.

A third characteristic that distinguishes among literature reviews, *perspective*, relates to whether the reviewers have an initial point of view that might influence the discussion of the literature. The endpoints on the continuum of perspective might be called *neutral representation* and *espousal of a position*. In the former, reviewers attempt to present all arguments or evidence for and against various interpretations of the problem. The presentation is meant to be as similar as possible to those that would be provided by the originators of the arguments or evidence. At the opposite extreme of perspective, the viewpoints of reviewers play an active role in how material is interpreted and presented. The reviewers accumulate and synthesize the literature in the service of demonstrating the value of the particular point of view that they espouse. The reviewers muster arguments and evidence so that it presents their contentions in the most convincing manner.

Of course, reviewers attempting to achieve complete neutrality are likely doomed to failure. Further, reviewers who attempt to present all sides of an argument do not preclude themselves from ultimately taking a strong position based on the cumulative evidence. Similarly, reviewers can be thoughtful and fair while presenting conflicting evidence or opinions and still advocate for a particular interpretation.

The next characteristic, *coverage*, concerns the extent to which reviewers find and include relevant works in their paper. It is possible to distinguish at least four types of coverage. The first type, *exhaustive coverage*, suggests that the reviewers hope to be comprehensive in the presentation of the relevant work. An effort is made to include the entire literature and to base conclusions and discussions on this comprehensive information base. The second type of coverage also bases conclusions on entire literatures, but only a selection of works is actually described in the literature review. The authors choose a purposive sample of works to cite but claim that the inferences drawn are based on a more extensive literature. Third, some reviewers will present works that are broadly representative of many other works in a field. They hope to describe just a few exemplars that are descriptive of numerous other works. The reviewers discuss the characteristics that make the chosen works paradigmatic of the larger group. In the final coverage strategy, reviewers concentrate on works that were highly original when they appeared and influenced the development of future efforts in the topic area. These may include materials that initiated a line of investigation or thinking, changed how questions were framed, introduced new methods, engendered important debate, or performed a heuristic function for other scholars.

A fifth characteristic of literature reviews concerns a paper's *organization*. Reviews may be arranged historically, so that topics are introduced in the chronological order in which they appeared in the literature; conceptually, so that works relating to the same abstract ideas appear together; or methodologically, so that works employing similar methods are grouped together.

Finally, the intended *audiences* of reviews can vary. Reviews can be written for groups of specialized researchers, general researchers, policymakers, practitioners, or the general public. As reviewers move from addressing specialized researchers to addressing the general public, they use less technical jargon and detail and often pay greater attention to the implications of the work being covered.

1.2.1.1 Definitions of Research Synthesis The terms *research synthesis* or *research review* or *systematic review* are often used interchangeably in the social science literature, though they sometimes connote subtly different meanings. Regrettably, no consensus has been reached about what these meaningful differences might be. Therefore, we use the term *research synthesis* most frequently throughout this book. The reason for this choice is simple. In addition to its use in the context of research synthesis, the term *research review* is also used to describe the activities of evaluating the quality of research. For example, a journal editor will obtain research reviews when deciding whether to publish a manuscript. Because research syntheses often include this type of evaluative review of research, using the term *research synthesis* avoids confusion. The term *systematic review* is less often used in the context of research evaluation, though the confusion is still there, and the specification that it is the results of research that are being synthesized is missing. The Cochrane Collaboration uses *systematic review* but has moved toward using *Cochrane review* to signify the use of its distinct tools and methodology (http://community.cochrane.org). The Campbell Collaboration (http://www.campbellcollaboration.org) also uses the term *systematic review* to label all its reviews, whether quantitative or qualitative.

A research synthesis can be defined as the conjunction of a particular set of literature review characteristics. Most distinctive about research syntheses are their primary focus and goal: research syntheses attempt to integrate empirical research for the purpose of creating generalizations. Implicit in this definition is the notion that seeking generalizations also involves seeking the limits of generalizations. Also, research syntheses almost always pay attention to relevant theories, critically analyze the research they cover, try to resolve conflicts in the literature, and attempt to identify central issues for future research. According to Derek Price, research syntheses are intended to "replace those papers that have been lost from sight behind the research front" (1965, 513). Research synthesis is one of a broad array of integrative activities that scientists engage in; its intellectual heritage can be traced back at least as far as Aristotle.

Using the described taxonomy, we can make further specifications concerning the type of research syntheses that are the focus of this book. With regard to perspective, readers will note that much of the material is meant to help synthesists produce neutral statements about evidence, that is, avoid being affected by many types of bias including their own subjective outlooks. For example, the material on searching the literature for evidence is meant to help synthesists uncover all the evidence, not simply positive studies that might be overrepresented in published research, or evidence that is easy for them to find and therefore might be overly sympathetic to their point of view. The material on the reliability of extracting information from research reports and how methodologi-

cal variations in research should be handled is meant to increase transparency and interjudge reliability when these activities are carried out. The methods proposed for the statistical integration of findings are meant to ensure the same rules about data analysis are applied to the next users of data as were required of the data generators. Finally, the material on explicit and exhaustive reporting of methods is meant to assist both producers and consumers of research syntheses in evaluating if or where bias may have crept into the synthesis process and to replicate findings if they choose to do so.

Finally, the term *meta-analysis* often is used as a synonym for research synthesis. However, in this volume, it is used in its more precise and original meaning—to describe the quantitative procedures that a research synthesist may use to statistically combine the results of studies. Gene Glass coined the term *meta-analysis* to refer to "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (1976, 3). The authors of this book reserve meta-analysis to refer specifically to statistical analysis in research synthesis and not to the entire enterprise of research synthesis. Not all research syntheses are appropriate for meta-analysis.

1.3 A BRIEF HISTORY OF RESEARCH SYNTHESIS AS A SCIENTIFIC ENTERPRISE

1.3.1 Early Developments

In 1971, Kenneth Feldman published an article titled "Using the Work of Others" in which he demonstrated remarkable prescience: "Systematically reviewing and integrating . . . the literature of a field may be considered a type of research in its own right—one using a characteristic set of research techniques and methods" (86). He described four steps in the reviewing process: sampling topics and studies, developing a scheme for indexing and coding material, integrating the studies, and writing the report.

The same year, Richard Light and Paul Smith presented what they called a cluster approach to literature reviewing that was meant to redress some of the deficiencies in the existing strategies for integration (1971). They argued that, if treated properly, the variation in outcomes among related studies could be a valuable source of information rather than merely a source of consternation, as it appeared to be when treated with traditional reviewing methods.

Three years later, Thomas Taveggia struck a complementary theme. He wrote,

> A methodological principle overlooked by [reviewers] . . . is that research results are probabilistic. What this principle suggests is that, in and of themselves, the findings of any single research are meaningless—they may have occurred simply by chance. It also follows that, if a large enough number of researches has been done on a particular topic, chance alone dictates that studies will exist that report inconsistent and contradictory findings! Thus, what appears to be contradictory may simply be the positive and negative details of a distribution of findings. (1974, 397–98)

Taveggia went on to describe six common tasks in research syntheses: selecting research; retrieving, indexing, and coding information from studies; analyzing the comparability of findings; accumulating comparable findings; analyzing distributions of results, and; reporting of results.

The development of meta-analytic techniques extends back further in time but their routine use by research synthesists is also relatively recent. Where Glass gave us the term meta-analysis in 1976, in 1990 Ingram Olkin pointed out that ways to estimate effect sizes have existed since the turn to the twentieth century. For example, Karl Pearson took the average of estimates from five separate samples of the correlation between inoculation for enteric (or typhoid) fever and mortality (1904). He used this average to better estimate the typical effect of inoculation and to compare it with that of inoculation for other diseases. Early work on the methodology for combination of estimates across studies includes papers in the physical sciences by Raymond Birge (1932) and in statistics by William Cochran (1937) and Frank Yates and Cochran (1938). Although they have fallen out of use today, methods for combining probabilities across studies also have a long history (Tippett 1931; Fisher 1932; Mosteller and Bush 1954).

Still, the use of quantitative synthesis techniques in the social sciences was rare before the 1970s. Late in that decade, several applications of meta-analytic techniques captured the imagination of behavioral scientists. Included among these were: in clinical psychology, Mary Smith and Gene Glass's meta-analysis of psychotherapy research (1977); in industrial-organizational psychology, Frank Schmidt and John Hunter's validity generalization of employment tests (1977); in social psychology, Robert Rosenthal and Donald Rubin's integration of interpersonal expectancy effect research (1978); and in education,

Glass and Smith's synthesis of the literature on class size and achievement (1978).

1.3.2 Research Synthesis Comes of Age

Two papers that appeared in the *Review of Educational Research* in the early 1980s brought the meta-analytic and research synthesis-as-research perspectives together. The first, by Gregg Jackson, proposed six reviewing tasks "analogous to those performed during primary research" (1980, 441). Jackson portrayed meta-analysis as an aid to the task of analyzing primary studies but emphasized its limitations as well as its strengths. Also noteworthy about his paper was his use of a sample of thirty-six review articles from prestigious social science periodicals to examine the methods used in integrative empirical reviews. For example, Jackson reported that only one of the thirty-six reported the indexes or retrieval systems used to locate primary studies. His conclusion was that "relatively little thought has been given to the methods for doing integrative reviews. Such reviews are critical to science and social policy making and yet most are done far less rigorously than is currently possible" (459).

The first half of the 1980s also witnessed the appearance of four books primarily devoted to meta-analytic methods. The first, in 1981, by Glass, Barry McGaw, and Smith, presented meta-analysis as a new application of analysis of variance and multiple regression procedures, with effect sizes treated as the dependent variable. In 1982, Hunter, Schmidt, and Jackson introduced meta-analytic procedures that focused on comparing the observed variation in study outcomes to that expected by chance (the statistical realization of a point Taveggia made in 1974) and correcting observed effect-size estimates and their variance for known sources of bias (such as sampling error, range restrictions, unreliability of measurements). In 1984, Rosenthal presented a compendium of meta-analytic methods covering, among other topics, the combining of significance levels, effect-size estimation, and the analysis of variation in effect sizes. Rosenthal's procedures for testing moderators of variation in effect sizes were not based on traditional inferential statistics, but on a new set of techniques involving assumptions tailored specifically for the analysis of study outcomes. Finally, in 1985, with the publication of *Statistical Procedures for Meta-Analysis*, Larry Hedges and Olkin helped elevate the quantitative synthesis of research to an independent specialty within the statistical sciences. This book, summarizing and expanding nearly a decade of programmatic developments by the authors, not only covered the widest array of meta-analytic procedures but also presented rigorous statistical proofs establishing their legitimacy.

Harris Cooper drew the analogy between research synthesis and primary research to its logical conclusion and presented a five-stage model of the integrative review as a research project (1982). For each stage, he codified the research question, its primary function in the review, and the procedural differences that might cause variation in review conclusions. In addition, he applied the notion of threats-to-inferential-validity—which Donald Campbell and Julian Stanley introduced for evaluating the utility of primary research designs (1966)—to the conduct of research synthesis (also see Shadish, Cook, and Campbell 2002). Cooper identified ten threats to validity specifically associated with reviewing procedures that might undermine the trustworthiness of the findings of a research synthesis. He also suggested that other threats might exist and that any particular synthesis' validity could be threatened by consistent deficiencies in the set of studies that formed its database. Table 1.2 presents a recent revision of this schema, which proposes a seven-stage model for conducting a research synthesis, separating the original coding stage into coding and study evaluation, the analysis stage into separate analyses, and interpretation stage into two distinct stages (Cooper 2017).

Another text that appeared in 1984 also helped elevate research synthesis to a more rigorous level. In it, Light and David Pillemer focused on the use of research reviews to help decision making in the social policy domain. Their approach placed special emphasis on the importance of meshing both numbers and narrative for the effective interpretation and communication of synthesis results.

Numerous books have appeared on research synthesis and meta-analysis since the mid-1980s—in fact, too many to mention all of them. Some focus on research synthesis in general (Card 2012; Lipsey and Wilson 2001; Petticrew and Roberts 2006; Schmidt and Hunter 2015); others treat it from the perspective of particular research designs (Bohning, Kuhnert, and Rattanasiri 2008; Eddy, Hassleblad, and Schachter 1992). Still others are tied to particular software packages (Arthur, Bennett, and Huffcutt 2001; Chen and Peace 2013; Comprehensive Meta-Analysis 2015). In 1994, the first edition of this book was published; the second edition appeared in 2009. Readers interested in a popular history of the origins of meta-analysis in the social sciences can consult

Table 1.2 Research Synthesis Conceptualized as a Research Process

Step in Research Synthesis	Research Question Asked at This Stage of the Synthesis	Primary Function Served in the Synthesis	Procedural Variation That Might Produce Differences in Conclusions
Formulating the problem	What research evidence will be relevant to the problem or hypothesis of interest in the synthesis?	Define the variables and relationships of interest so that relevant and irrelevant studies can be distinguished	Variation in the conceptual breadth and distinctions within definitions might lead to differences in the research operations deemed relevant and/or tested as moderating influences
Searching the literature	What procedures should be used to find relevant research?	Identify sources (such as reference databases, journals) and terms used to search for relevant research	Variation in searched sources might lead to systematic differences in the retrieved research
Gathering information from studies	What information about each study is relevant to the problem or hypothesis of interest?	Collect relevant information about studies in a reliable manner	Variation in information gathered might lead to differences in what is tested as an influence on cumulative results, in coder training might lead to differences in entries on coding sheets, or in rules for deciding what study results are independent tests of hypotheses might lead to differences in the amount and specificity of data used to draw cumulative conclusions
Evaluating the quality of studies	What research should be included in the synthesis based on the suitability of the methods for studying the synthesis question or problems in research implementation?	Identify and apply criteria that separate studies conducted in ways that correspond with the research question from studies that do not	Variation in criteria for decisions about study methods to include might lead to systematic differences in which studies remain in the synthesis
Analyzing and integrating the outcomes of studies	What procedures should be used to condense and combine the research results?	Identify and apply procedures for combining results across studies and testing for differences in results between studies	Variation in procedures used to summarize and compare results of included studies (such as narrative, vote count, averaged effect sizes) can lead to differences in cumulative results
Interpreting the evidence	What conclusions can be drawn about the cumulative state of the research evidence?	Summarize the cumulative research evidence with regard to its strength, generality, and limitations	Variation in criteria for labeling results as important and attention to details of studies might lead to differences in interpretation of findings
Presenting the results	What information should be included in the report of the synthesis?	Identify and apply editorial guidelines and judgment to determine aspects of methods and results readers of the report will need to know	Variation in reporting might lead readers to place more or less trust in synthesis outcomes and influence others' ability to replicate results

SOURCE: Authors' compilation.

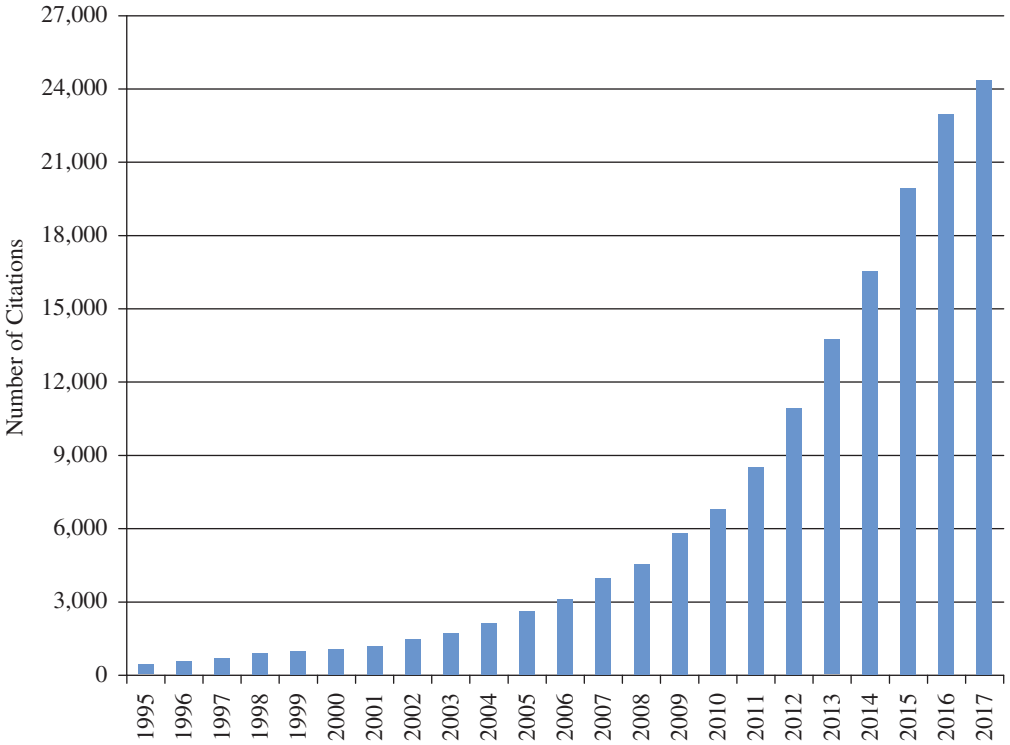

Figure 1.1 Citations to Articles Including the Terms *Research Synthesis*, *Systematic Review*, or *Meta-Analysis* in Their Titles

SOURCE: Authors' compilation based on Web of Science Core Collection (Clarivate Analytics 2018).
NOTE: Bars chart the growth in the number of titles of documents including the terms *research synthesis*, *systematic review*, or *meta-analysis* during the years following the publication of the first edition of the *Handbook of Research Synthesis.*

How Science Takes Stock: The Story of Meta-Analysis (Hunt 1997). Most recently, the journal *Research Synthesis Methods* published a special issue on the origins of modern meta-analysis (Shadish 2015).

Literally thousands of research syntheses have been published since the first edition of this book. Figure 1.1 presents some evidence of the increasing impact of research syntheses on knowledge in the sciences and social sciences. The figure is based on entries in the Web of Science Core Collection reference database (Clarivate Analytics 2018). It charts the growth in the number of document titles including the terms *research synthesis*, *systematic review*, or *meta-analysis* in their title or during the years 1995 to 2017. The figure indicates that documents in the database titles using these terms has risen every year without exception and the growth is accelerating. Clearly, the role that research syntheses play in our knowledge claims is large and growing larger.

The use of research synthesis has spread from psychology and education through many disciplines, especially in medicine, social policy analysis, and economics. Indeed, the development of scientific methods for research synthesis has its own largely independent history in the medical sciences (see Chalmers, Hedges, and Cooper 2002). A most notable event in medicine was the establishment of the UK Cochrane Center in 1992. The center was meant to facilitate the creation of an international network to prepare and maintain systematic reviews of the effects of interventions across the spectrum of health-care practices. At the end of 1993, an international network of

individuals, the Cochrane Collaboration (http://www.cochrane.org), emerged from this initiative (Chalmers 1993; Bero and Rennie 1995). By 2019, the Cochrane Collaboration was an internationally renowned initiative that counted thousands of people in more than ninety countries contributing to its work. The Cochrane Collaboration is now the leading producer of research syntheses in health care and is considered by many to be the gold standard for determining the effectiveness of different health-care interventions. Its library of systematic reviews numbers in the thousands. In 2000, an initiative known as the Campbell Collaboration was launched with similar objectives for the domain of social policy analysis, focusing initially on policies concerning education, social welfare, and crime and justice (http://www.campbellcollaboration.org).

Because of the efforts of scholars who chose to apply their skills to how research syntheses might be improved, syntheses written since the 1980s have been held to standards far more demanding than those applied to their predecessors. The process of elevating the rigor of syntheses has continued into the twenty-first century.

1.3.3 Rationale for the *Handbook*

The Handbook of Research Synthesis and Meta-Analysis is meant to be the definitive *vade mecum* for behavioral and social scientists intent on applying the synthesis craft. It distills the products of forty years of developments in how research integrations should be conducted so as to minimize the chances of conclusions that do not truly reflect the cumulated evidence. Research synthesis in the 1960s was at best an art, at worst a form of yellow journalism. Today, the summarization and integration of studies is viewed as a research process in its own right, is held to the standards of a scientific endeavor, and entails the application of data gathering and analyses techniques developed for its unique purpose.

Numerous excellent texts on research synthesis exist. However, none is as comprehensive and detailed as this volume. Some texts focus on statistical methods. These often emphasize different aspects of statistical integration (such as combining probabilities, regression-analog models, estimating population effects from sampled effects with known biases) and often approach research accumulation from different perspectives. Although these texts are complete within their domains, no single sourcebook describes and integrates all the meta-analytic approaches in most common use.

This volume incorporates quantitative statistical techniques from all the synthesis traditions. It brings the leading authorities on the various meta-analytic perspectives together in a single volume. In doing so, it is an explicit statement that all the statistical approaches share a common assumptive base. This base is not only statistical but also philosophical. Philosophically, all the approaches rest on the presupposition that research syntheses need to be held to the same standards of rigor, systematicity, and transparency as the research on which they are based. The second and later users of data must be held as accountable for the validity of their methods as were the first.

Several problems arising in the course of conducting a quantitative synthesis have not received adequate treatment in any existing text. These include nonindependence of data sets, synthesis of multivariate data sets, and sensitivity analysis, to name just a few. Every research synthesist faces these problems and has developed strategies for dealing with them. Some of their solutions are published in widely scattered journals; others are often passed on to colleagues through informal contacts. They have never received complete treatment within the same text. This *Handbook* brings these topics together in a single volume.

Further, texts focusing on the statistical aspects of integration tend to give only passing consideration to other activities of research synthesis. These activities include the unique characteristics of problem formulation in research synthesis; methods of literature search; coding and evaluation of research reports; and the meaningful interpretation and effective communication of synthesis results. The existing texts that focus on these aspects of research synthesis tend not to be comprehensive in their coverage of statistical issues. Fully half of the chapters in this volume deal with issues that are not statistical in nature, evidencing the authors' collective belief that high-quality syntheses require considerably more than simple application of quantitative procedures.

Finally, this volume is meant for those who carry out research syntheses. Discussions of theory and proof are kept to a minimum in favor of descriptions of the practical mechanics needed to apply well the synthesis craft. The chapters include multiple approaches to problem solving and discuss the strengths and weaknesses of each approach. Readers with a comfortable background in analysis of variance and multiple regression and who have access to a research library should find the chapters accessible. The *Handbook* authors want to supply working synthesists with the needed expertise to interpret their

blueprints, to wield their mortar hoe and trowel as accurately as possible.

1.4 STAGES OF RESEARCH SYNTHESIS

The description of the stages of research synthesis presented in table 1.2 provides the conceptual organization of this book. In this section, we raise the principal issues associated with each stage. This allows us to briefly introduce the content of each of the chapters that follow.

1.4.1 Formulating a Problem for a Research Synthesis

The one major constraint on problem formulation in research synthesis is that primary research on a topic must exist before a synthesis can be conducted. How much research? The methods of meta-analysis can be applied to literatures containing as few as two hypothesis tests (Valentine, Pigott, and Rothstein 2010). Under certain circumstances—for instance, researchers synthesizing a pair of replicate studies from their own lab—the use of meta-analysis in this fashion might be sensible. Yet, most scientists would argue that the benefits of such a review would be limited (and its chances for publication even more limited).

A more general answer to the "How much research?" question is that it varies depending on a number of characteristics of the problem. All else being equal, conceptually broad topics would seem to profit from a synthesis only after the accumulation of a more varied and larger number of studies than a narrowly defined topic would (see chapter 2). Similarly, literatures that contain diverse types of operations also would seem to require a relatively large number of studies before firm conclusions could be drawn from a synthesis. Ultimately, the arbiter of whether a synthesis is needed will not be numerical standards, but the fresh insights a synthesis can bring to a field. Indeed, although a meta-analysis cannot be performed without data, many social scientists see value in "empty" syntheses that point to important gaps in our knowledge. When done properly, empty syntheses should proceed through the stages of research synthesis, including careful problem formulation.

Once enough literature on a problem has collected, then the challenge, and promise, of research synthesis becomes evident. The problems that constrain primary researchers—small and homogeneous samples, limited time and money for turning constructs of interest into multiple operations—are less severe for synthesists. They can capitalize on the diversity in methods that has occurred naturally across primary studies. The heterogeneity of methods across studies may permit tests of theoretical hypotheses concerning the moderators and mediators of relations that have never been tested in any single primary study. Conclusions about the population and ecological validity of relations uncovered in primary research may also receive more thorough tests in syntheses.

Part II of this book focuses on issues in problem formulation. In chapter 2 ("Hypotheses and Problems in Research Synthesis"), Harris Cooper discusses in detail the issues just mentioned. In chapter 3 ("Statistical Considerations"), Larry Hedges looks at the implications of different problem definitions for how study results will be statistically modeled. The major issues involve the populations of people and measurements that are the target of a review's inferences; how broadly the key constructs are defined, especially in terms of whether fixed- or random-effect models are envisioned; and how choices among models influence the precision of estimates and the statistical power of meta-analytic tests.

1.4.2 Searching the Literature

The literature search is the stage of research synthesis that is most different from primary research. Still, culling through the literature for relevant studies is not unlike gathering a sample of primary data. The target of a literature search that is part of a synthesis attempting exhaustive coverage would be "all the research conducted on the topic of interest."

In contrast to the (relatively) well-defined sampling frames available to primary researchers, literature searchers confront the fact that any single source of primary reports will lead them to only a fraction of the relevant studies, and a biased fraction at that. For example, the most inclusive sources of literature are the reference databases, such as Google Scholar, Science Direct, PsycINFO, ERIC, and Medline. Still, many of these broad, nonevaluative systems exclude much of the unpublished literature. Conversely, the most exclusive literature searching technique involves accessing close colleagues and other researchers with an active interest in the topic area. Despite the obvious biases, there is no better source of unpublished and recent works. Further complicating the sampling frame problem is that the relative utility and biases associated with any single source will vary as a function of characteristics of the research problem, includ-

ing, for example, how long the topic has been the focus of study and whether it is interdisciplinary.

These problems imply that research synthesists must carefully consider multiple channels for accessing literature and how the channels they choose complement one another. The three chapters in part III are devoted to helping the synthesist consider and carry out this most unique task. In chapter 4 ("Scientific Communication and Literature Retrieval"), Howard White presents an overview of searching issues from the viewpoint of an information scientist. In chapter 5, "Searching Bibliographic Databases," Julie Glanville provides strategies for using electronic databases (such as reference databases, citation indexes, research registries) to assist researchers with finding and accessing scholarship that is relevant to their work. In chapter 6 ("Retrieving Grey Literature, Information, and Data in the Digital Age"), Dean Giustini discusses the practical considerations of how to find research that is not indexed in the usual academic databases.

1.4.3 Evaluating Study Methodology and Extracting Information from Study Reports

Part IV offers four chapters on the evaluation of the study designs and implementation and retrieving information from studies. Once the synthesists have gathered the relevant literature, they must extract from each document those pieces of information that will help answer the questions that impel research in the field. This step includes judgments about the critical aspects of each study's research design, measurements, and procedures, and how variations in these relate to the inferences the synthesists wish to make. The problems faced during data coding provide a strong test of the synthesists' knowledge of the research area, thoughtfulness, and ingenuity. The decisions made during coding will have a profound influence on the contribution of the synthesis.

The aspect of coding studies that engenders the most debate involves how synthesists should represent differences in the design and implementation of primary studies that contribute to their data. What is meant by *study quality* when we are evaluating research methods? Should studies be given more or less credibility and therefore weighted differently in a meta-analysis if they differ in quality? Should studies be excluded if they contain too many flaws? How does one rate the quality of studies described in incomplete research reports? In chapter 7 ("Incorporating Judgments About Study Quality into Research Syntheses"), Jeffrey Valentine examines the alternative approaches available to synthesists for representing primary research methodology.

But judging a studies credibility is only the beginning. Synthesists must make decisions about other classes of variables that are of potential interest to them. These can relate to variables that predict outcomes, potential moderators and mediators of effects, and the differences in how outcomes are conceptualized (and, therefore, measured). Might the type of participants and the context of the study influence its outcomes? What about characteristics of the experimental manipulation (for example, intensity, duration) and measurements (for example, reliability, timing)? If a synthesist chooses not to code a particular feature of studies, then it cannot be considered in the analysis of results.

General guidelines for what information should be extracted from primary research reports are difficult to develop beyond recommendations that are general and abstract. Instead, direction will come from the issues that have arisen in the particular literature, coupled with the synthesist's personal insights into the topic. Still, commonalities emerge about what information is important to collect and how to think about what information to retrieve from studies. Mark Lipsey, in chapter 8 ("Identifying Potentially Interesting Variables and Analysis Opportunities"), and David Wilson, in chapter 9 ("Systematic Coding for Research Synthesis"), present complementing templates for what generally should be included on coding frames.

Once decisions on what to code have been made, synthesists need to consider how to carry out the coding (for example, who will retrieve information, how will they be trained) and how to assess the trustworthiness with which the coding frame is implemented. Numerous indexes of coder reliability are available, each with different strengths and weaknesses. In chapter 10 ("Evaluating Coding Decisions"), Jack Vevea, Nicole Zelinsky, and Robert Orwin describe strategies for reducing the amount of error that enters a synthesis during the coding of the literature's features. Their description of reliability assessment focuses on three major approaches: sources of coding error, strategies for reducing coding error, and strategies for statistically assessing and quantifying coding error.

1.4.4 Statistically Describing and Combining Study Outcomes

As our brief history of research synthesis suggests, techniques for the analysis of accumulated research outcomes

is an area of statistics abundant in dramatic developments. Four decades ago, the mechanics of integrating research usually involved intuitive processes taking place inside the heads of the reviewers. Meta-analysis made these processes public and based them on explicit, shared, statistical assumptions (however well met). We would not accept as valid a primary researcher's conclusion if it were substantiated solely by the statement "I looked at the treatment and control scores and I found the treated group did better." We would demand statistical testing (for example, a simple *t*-test) to back up the claim. Likewise, we should no longer accept "I examined the study outcomes and find the treatment is effective" as sufficient warrant for the conclusion of a research synthesis.

Part V covers the components of synthesis dealing with combining study results. Chapter 11, by Michael Borenstein and Larry Hedges on effect sizes, covers methods for estimating the outcomes of studies using a common metric. Thirty years ago, Jacob Cohen defined an effect size as "the *degree* to which the phenomenon is present in the population, or the degree to which the null hypothesis is false" (1988, 9–10).

To most research synthesists, the search for influences on study results is the most interesting and rewarding part of the synthesis process. The next two chapters deal with techniques for analyzing whether and why there are differences in the outcomes of studies. As an analog to analysis of variance or multiple regression procedures, effect sizes can be viewed as dependent or criterion variables and the features of study designs as independent or predictor variables. However, because effect-size estimates do not all have the same sampling uncertainty, they cannot simply be inserted into traditional inferential statistics. In chapter 12 ("Statistically Analyzing Effect Sizes: Fixed- and Random-Effects Models"), Spyros Konstantopoulos and Larry Hedges discuss the difference between fixed- and random-effects models of effect-size homogeneity, the conceptual and statistical considerations involved in choosing an analytic model, and the statistical power of homogeneity tests. Chapter 13, by Larry Hedges, addresses recent advances in multivariate meta-analysis, in particular the use of meta-regression. This chapter also provides guidance to help reviewers avoid common mistakes when multivariate data are used in meta-analysis.

Part V delves into other approaches to the statistical combination of study results. In chapter 14, Rebecca Turner and Julian Higgins describe Bayesian meta-analysis, including Bayesian meta-regression and the advantages and limitations of this approach. Effect-size estimates may be affected by factors that attenuate their magnitudes. These may include, for example, a lack of reliability in the measurement instruments or restrictions in the range of values in the subject sample. These attenuating biases may be estimated and corrected using the procedures Frank Schmidt, Huy Le, and In-Sue Oh describe in chapter 15. In chapter 16, Betsy Becker and Ariel Aloe introduce model-based meta-analysis and how to use this approach to investigate partial effects, indirect effects (including mediation), and to address questions that have not been explicitly addressed in any individual studies.

Part VI addresses two important complications that arise when working with meta-analytic data that all research synthesists must attend to. In chapter 17, Terri Pigott takes up handling missing data. She addresses different types of missing data (missing studies, effect sizes, study descriptors), provides an overview and critique of commonly used methods, discusses model-based methods for addressing missing data, and outcome reporting biases. In chapter 18, which takes up publication bias, Jack Vevea, Kathleen Coburn, and Alexander Sutton introduce methods to identify the presence, assess the impact, and adjust results for the synthesists who want to examine whether the published literature might be a biased sample all the studies that have been conducted.

1.4.5 Interpreting Synthesis Outcomes

Estimating and averaging effect sizes and searching for moderators of their variability is how the interpretation of cumulative study results begins. However, it must be followed by other procedures that help the synthesists properly interpret what they have discovered. Proper interpretation of the results of a research synthesis requires careful use of declarative statements regarding claims about the evidence, specification of what results warrant each claim, and any appropriate qualifications to claims that need to be made.

Part VII examines two important issues in data interpretation. In chapter 19 ("Interpreting Effect Sizes"), Jeffrey Valentine, Ariel Aloe, and Sandra Jo Wilson discuss methods for interpreting effect sizes in real-world terms. In chapter 20 ("Heterogeneity in Meta-Analysis"), Michael Borenstein introduces important considerations when thinking about differences between studies, including the distinction between observed and true effects,

statistics for assessing and describing heterogeneity, the null hypothesis of effect-size homogeneity, and common mistakes in thinking about heterogeneity.

1.4.6 Presenting Synthesis Results

Presenting the background, methods, results, and meaning of a research synthesis' findings are the final challenges to the synthesists' skill and intellect. These are addressed in the summary section, part VIII. In chapter 21, Evan Mayo-Wilson and Sean Grant describe the standards for reporting meta-analysis. As is true of the coding frame, no simple reporting scheme fits all syntheses. However, certain commonalities do exist. Not too surprisingly, the organization that emerges bears considerable resemblance to that of a primary research report although, also obviously, the content differs dramatically. In chapter 22 ("Threats to the Validity of Generalized Inferences from Research Syntheses"), Georg Matt and Thomas Cook provide an overall appraisal of how inferences from research syntheses may be restricted or faulty. This chapter brings together many of the concerns expressed throughout the book by the various chapter authors. Finally, chapter 23 ("Potentials and Limitations of Research Synthesis"), Harris Cooper, Larry Hedges, and Jeffrey Valentine pay special attention to possible future developments in synthesis methodology, the feasibility and expense associated with conducting a sound research synthesis, and a broad-based definition of what makes a literature review good or bad.

No secret will be revealed by stating our conclusion in advance. If procedures for the synthesis of research are held to standards of objectivity, systematicity, and rigor, then our knowledge edifice will be made of bricks and mortar. If not, it will be a house of cards.

1.5 REFERENCES

Arthur, Winfred, Jr., Winston Bennett Jr., and Allen I. Huffcutt. 2001. *Conducting Meta-Analysis Using SAS.* Mahwah, N.J.: Lawrence Erlbaum.

Bero, Lisa, and Drummond Rennie. 1995. "The Cochrane Collaboration: Preparing, Maintaining, and Disseminating Systematic Reviews of the Effects of Health Care." *Journal of the American Medical Association* 274(24): 1935–38.

Birge, Raymond T. 1932. "The Calculation of Error by the Method of Least Squares." *Physical Review* 40(2): 207–27.

Bohning, Dankmar, Ronny Kuhnert, and Sasivimol Rattanasiri. 2008. *Meta-Analysis of Binary Data Using Profile Likelihood.* Boca Raton, Fl.: Taylor and Francis.

Campbell, Donald T., and Julian C. Stanley. 1966. *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally.

Card, Noel A. 2012. *Applied Meta-Analysis for Social Science Research*. New York: Guilford.

Chalmers, Iain. 1993. "The Cochrane Collaboration: Preparing, Maintaining and Disseminating Systematic Reviews of the Effects of Health Care." *Annals of the New York Academy of Sciences* 703:156–63.

Chalmers, Iain, Larry V. Hedges, and Harris Cooper. 2002. "A Brief History of Research Synthesis." *Evaluation and the Health Professions* 25(1): 12–37.

Chen, Ding-Geng, and Karl E. Peace. 2013. *Applied Meta-Analysis with R.* Boca Raton, Fl.: Taylor and Francis.

Clarivate Analytics. 2018. "Web of Science Core Collection." Accessed May 7, 2018. https://clarivate.com/products/web-of-science/web-science-form/web-science-core-collection.

Cochran, William G. 1937. "Problems Arising in the Analysis of a Series of Similar Experiments." Supplement, *Journal of the Royal Statistical Society* 4: 102–118.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, N.J.: Lawrence Erlbaum.

Comprehensive Meta-Analysis. 2015. *Comprehensive Meta-Analysis*. Accessed November 20, 2018. http://www.meta-analysis.com/index.php.

Cooper, Harris M. 1982. "Scientific Guidelines for Conducting Integrative Research Reviews." *Review of Educational Research* 52(2): 291–302.

———. 1988. "Organizing Knowledge Synthesis: A Taxonomy of Literature Reviews." *Knowledge in Society* 1:104–26.

———. 2017. *Research Synthesis and Meta-Analysis: A Step-by-Step Approach*, 5th ed. Thousand Oaks, Calif.: Sage Publications.

Eddy, David M., Vic Hasselblad, and Ross Schachters. 1992. *Meta-Analysis by the Confidence Profile Approach*. Boston, Mass.: Academic Press.

Feldman, Kenneth A. 1971. "Using the Work of Others: Some Observations on Reviewing and Integrating." *Sociology of Education* 4:86–102.

Fisher, Ronald A. 1932. *Statistical Methods for Research Workers*, 4th ed. London: Oliver and Boyd.

Gelman, Andrew, and Hal Stern. 2006. "The Difference Between 'Significant' and 'Not Significant' Is Not Itself Statistically Significant." *American Statistician* 60(4): 328–31.

Glass, Gene V. 1976. "Primary, Secondary, and Meta-Analysis." *Educational Researcher* 5(1): 3–8.

Glass, Gene V., Barry McGaw, and Mary L. Smith. 1981. *Meta-Analysis in Social Research*. Beverly Hills, CA: Sage Publications.

Glass, Gene V., and Mary L. Smith. 1978. *Meta-Analysis of Research on the Relationship of Class Size and Achievement*. San Francisco: Far West Laboratory for Educational Research and Development.

Hedges, Larry V. 1987. "How Hard Is Hard Science, How Soft Is Soft Science?" *American Psychologist* 42(2): 443–55.

Hedges, Larry V., and Ingram Olkin. 1985. *Statistical Methods for Meta-Analysis*. Orlando, Fl.: Academic Press.

Hunt, Morton. 1997. *How Science Takes Stock: The Story of Meta-Analysis*. New York: Russell Sage Foundation.

Hunter, John E., Frank L. Schmidt, and Gregg B. Jackson. 1982. *Meta-Analysis: Cumulating Research Findings Across Studies*. Beverly Hills, Calif.: Sage Publications.

Jackson, Gregg B. 1980. "Methods for Integrative Reviews." *Review of Educational Research* 50(3): 438–60.

Light, Richard J., and David B. Pillemer. 1984. *Summing Up: The Science of Reviewing Research*. Cambridge, Mass.: Harvard University Press.

Light, Richard J., and Paul V. Smith. 1971. "Accumulating Evidence: Procedures for Resolving Contradictions Among Research Studies." *Harvard Educational Review* 41(4): 429–71.

Lipsey, Mark W., and David B. Wilson. 2001. *Practical Meta-Analysis*. Thousand Oaks, Calif.: Sage Publications.

Manten, Arie A. 1973. "Scientific Literature Review." *Scholarly Publishing* 5:75–89.

Mosteller, Frederick, and Robert R. Bush. 1954. "Selected Quantitative Techniques." In *Handbook of Social Psychology*, vol. 1. *Theory and Method*, edited by Gardner Lindzey. Cambridge, Mass.: Addison-Wesley.

Olkin, Ingram. 1990. "History and Goals." In *The Future of Meta-Analysis*, edited by Kenneth W. Wachter and Miron L. Straf. New York: Russell Sage Foundation.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349(6251): aac4716.

Pashler, Harold, and Eric-Jan Wagonmakers. 2012. "Editor's Introduction to the Special Issue on Replicability in Psychological Science: A Crisis of Confidence?" *Perspectives on Psychological Science* 7(6): 528–30.

Pearson, Karl. 1904. "Report on Certain Enteric Fever Inoculation Statistics." *British Medical Journal* 3:1243–46.

Petticrew, Mark, and Helen Roberts. 2006. *Systematic Reviews in the Social Sciences: A Practical Guide.* Malden, Mass.: Blackwell.

Price, Derek J. de Solla. 1965. "Networks of Scientific Papers." *Science* 149(3683): 510–15.

Rosenthal, Robert. 1984. *Meta-Analytic Procedures for Social Research*. Beverly Hills, Calif.: Sage Publications.

Rosenthal, Robert, and Donald B. Rubin. 1978. "Interpersonal Expectancy Effects: The First 345 Studies." *Behavioral and Brain Sciences* 3(6): 377–86.

Schmidt, Frank L., and John E. Hunter. 1977. "Development of a General Solution to the Problem of Validity Generalization." *Journal of Applied Psychology* 62(5): 529–40.

———. 2015. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*, 3rd ed. Thousand Oaks, Calif.: Sage Publications.

Shadish, William R. 2015. "Introduction to the Special Issue on the Origins of Modern Meta-Analysis." *Research Synthesis Methodology* 6(3): 219–20.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston, Mass.: Houghton Mifflin.

Smith, Mary L., and Gene V. Glass. 1977. "Meta-Analysis of Psychotherapy Outcome Studies." *American Psychologist* 32(9): 752–60.

Taveggia, Thomas C. 1974. "Resolving Research Controversy Through Empirical Cumulation: Toward Reliable Sociological Knowledge." *Sociological Methods & Research* 2(4): 395–407.

Tippett, Leonard Henry Caleb. 1931. *The Methods of Statistics*. London: Williams and Norgate.

Valentine, Jeffrey C., Anthony Biglan, Robert F. Boruch, Filipe G. Castro, Linda M. Collins, Brian R. Flay, Sheppard Kellam, Eve K. Moscicki, and Steven P. Schinke. 2011. "Replication in Prevention Science." *Prevention Science* 12(2): 103–17.

Valentine, Jeffrey C., Theresa D. Pigott, and Hannah R. Rothstein. 2010. "How Many Studies Do You Need? A Primer on Statistical Power for Meta-Analysis." *Journal of Educational and Behavioral Statistics* 2(2): 215–47.

Yates, Frank, and William G. Cochran. 1938. "The Analysis of Groups of Experiments." *Journal of Agricultural Science* 28(4): 556–80.

PART II

FORMULATING A PROBLEM

2

HYPOTHESES AND PROBLEMS IN RESEARCH SYNTHESIS

HARRIS COOPER
Duke University

CONTENTS

2.1 INTRODUCTION

Texts on research methods often list sources of research ideas (see, for example, Christensen, Johnson, and Turner 2014). Sometimes ideas for research come from personal experiences, sometimes from pressing practical issues. Sometimes a researcher wishes to test a theory meant to help understand the roots of human behavior. Yet other times researchers find topics by reading scholarly research.

Still, sources of ideas for research are so plentiful that the universe of possibilities seems limitless. Perhaps you must be satisfied with Karl Mannheim's suggestion that the ideas researchers pursue are rooted in their social and material relations, in their existential circumstances (1936). Then, at least, you have a course of action for studying why a particular researcher chooses one topic to study rather than another.

I do not begin my exploration of problems and hypotheses in research synthesis by examining the determinants of choice. Nor do I describe the existential bounds of the researcher. Instead, I start with a simple statement: the problems and hypotheses in research syntheses are drawn from those that already have had data collected on them in primary research. By definition, a research synthesis is an integration of past research. Then, my task becomes much more manageable. In this chapter, I examine the characteristics of research problems and hypotheses that (a) make them empirically testable and (b) are similar and different for primary research and research synthesis.

That syntheses are tied to only those problems that have previously generated data does not mean research synthesis is an uncreative exercise. The process of synthesis often requires the creation of explanatory models to help make sense of related studies that produced incommensurate data. Why did this study of a reading curriculum produce twice the effect of a similar study? These schemes can be novel, having never appeared in previous theorizing or research. The cumulative results of studies are much more complex than the results of any single study. Discovering why two studies that appear to be direct replications of one another produced conflicting results presents a deliberative challenge for any research synthesist.

2.2 DEFINITIONS OF BASIC VARIABLES IN SOCIAL SCIENCE RESEARCH

In its most basic form, the statement of a research problem includes a clear delineation of what variables are of interest to the researches and how the variables can be measured empirically. The rationale for a research problem can be that some circumstance needs fuller description (as in survey research or ethnography) or the estimation of the relation between variables might be important (either associational or causal). Alternatively, the problem can contain a prediction about a particular link between the variables—based on theory or previous observation. This type of prediction is called a hypothesis.

Primary research or research syntheses can be undertaken regardless of whether a study's rationale is a description of events, the association between variables, or a causal relationship between variables (more on these distinctions follows). For example, you might be interested in how doctors' training correlates with how they perform diagnoses. Here, you may know the problem is important but may have no precise theory that leads to a hypothesis about whether and how training is related to diagnoses. Your research problem is more exploratory and might begin with rich verbal descriptions by doctors of how they question patients about their ailments. Or, you might want to test the hypothesis that teachers' expectations concerning how a student's intelligence changes over the course of a school year will cause the students' intelligence to change in the expected direction. This hypothesis might be based on firsthand observations of how teachers treat students for whom they hold different expectations or on a well-specified model of how interpersonal communications change (such as nonverbal cues, body position, voice tone) as a function of the teachers' beliefs. Another example might be that you are interested in how students' perceptions of their instructors in college correlate with achievement in the course. Here, you may know the problem is important but have no precise theory that leads to a hypothesis about whether and how perceptions of the instructor are related to achievement in the course. You might look for variations in instructor's behavior and relate them to student achievement.

2.2.1 Conceptual and Operational Definitions

The variables involved in social science research must be defined in two ways. First, each variable must be given a conceptual or theoretical definition. This describes qualities of the variable that are independent of time and space but that can be used to distinguish events that are and are not relevant to the concept (Shoemaker, Tankard, and Lasorsa 2004). For instance, a broad conceptual definition of intelligence might be the ability to acquire and apply knowledge.

Conceptual definitions can differ in breadth, or in the number of events to which they refer. Thus, if you define achievement as a thing done successfully, whether by effort, courage, or skill, the concept is broader than if you confine the domain of achievement to academic tasks, or activities related to school performance in verbal and quantitative domains. The broader definition of achievement would include goals reached in physical, artistic, and social spheres of activity, as well as academic ones. So, if you are interested in the relationship between student ratings of instructors and achievement writ large you would include ratings taken in gym, health, and drama classes. If you are interested in academic achievement only, these types of classes would fall outside your conceptual definition. The need for well-specified conceptual definitions is no different in research synthesis than it is in primary research. Both primary researchers and research synthesists must clearly specify their conceptual definitions.

To relate concepts to concrete events, the variables in empirical research must also be operationally defined. An operational definition is a description of the characteristics of observable events used to determine whether the event represents an occurrence of the conceptual variable. Put differently, a concept is operationally defined by the procedures used to produce and measure it. Again, both primary researchers and research synthesists must specify the operations included in their conceptual definitions.

An operational definition of the concept of *intelligence* might first focus on scores from standardized tests meant to measure reasoning ability. This definition might be specified further to include only tests that result in intelligent quotient (IQ) scores, for example, the Stanford-Binet and the Wechsler tests. Or, the operational definition might be broadened to include the Scholastic Aptitude Test (SAT), the Graduate Record Exam (GRE), or the Miller Analogies Test (MAT). This last group might be seen as broadening the definition of intelligence because these tests may be influenced by the test takers' knowledge—what they have been taught in the past—as well as their native ability to acquire and apply knowledge.

Also needing operational definitions are the ingredients of a treatment, or more formally, the components of an intervention or training program. For example, a psychologist might devise an intervention for doctors meant to improve their ability to make accurate diagnoses. This might include training in the interpretation of x-rays and what kinds of follow-up questions to ask when patients make particular complaints.

2.2.1.1 Distinctions Between Operational Definitions in Primary Research and Research Synthesis The first critical distinction between operational definitions in primary research and research synthesis is that primary researchers cannot start data collection until the variables of interest have been given a precise operational definition, an empirical reality. Otherwise, the researcher does not know how to proceed with a treatment or experimental manipulation or what data to collect. A primary researcher studying intelligence must pick the measure or measures of intelligence they wish to study, ones that fit nicely into their conceptual definition, before the study begins.

In contrast, research synthesists need not be quite so conceptually or operationally precise, at least not initially. The literature search for a research synthesis can begin with only a broad, and sometimes fuzzy, conceptual definition and a few known operations that measure the construct. The search for studies then might lead the synthesists not only to studies that define the construct in the manners they specified, but also to research in which the same construct was studied with different operational definitions. Then, the concept and associated operations in the research synthesis can grow broader or narrower—and hopefully more precise—as the synthesists grow more familiar with how the construct has been defined and measured in the extant research. Synthesists have the comparative luxury of being able to evaluate the conceptual relevance of different operations as they grow more familiar with the literature. They can even modify the conceptual definition as they encounter related alternative concepts and operations in the literature. This can be both a fascinating and anxiety-arousing task. Does intelligence include what is measured by the SAT? Should achievement include performance in music class?

I do not want to give the impression that research synthesis permits fuzzy thinking. Typically, research synthesists begin with a clear idea of the concepts of interest and with some a priori specification of related empirical realizations. However, during a literature search, it is not unusual for synthesists to come across definitions that raise issues about concept boundaries they may not have considered or operations that they did not know existed but seem relevant to the construct being studied. This can be as simple as finding a measure of intelligence they did not know existed. It can be as complicated as discovering a whole literature on how to measure accuracy in diagnoses and what that term even means. Some of these new considerations may lead to changes in the conceptual

definition. In sum, primary researchers need to know exactly what events will constitute the domain to be sampled before beginning data collection. Research synthesists may discover unanticipated elements of the domain along the way.

Another distinction between the two types of inquiry is that primary studies will typically involve only one, and sometimes a few, operational definitions of the same construct. In contrast, research syntheses usually include many empirical realizations. For example, primary researchers may pick a single measure of intelligence, if only because the time and economic constraints of administering multiple measures of intelligence will be prohibitive. A few measures of academic achievement might be available to primary researchers, if they can access archived student data such as grade point averages and achievement test scores. On the other hand, research synthesists can uncover a wide array of operationalizations of the same construct within the same research area. For example, a synthesist may find that one researcher used peer judgments to define the accuracy of a diagnosis, and that another used cure rate, and yet another used patient satisfaction.

2.2.2 Fit Between Concepts and Operations

The variety of operationalizations that a research synthesist may uncover in the literature can be both a curse and a blessing. The curse concerns the fit between concepts and operations.

2.2.2.1 Broadening and Narrowing of Concepts Synthesists may begin a literature search with broad conceptual definitions. However, they may discover that the operations used in previous relevant research have been confined to a narrower conceptualization. For instance, if you are conducting a research synthesis about how student perceptions of instructors in college correlate with course achievement you might discover that all or nearly all past research has measured only "liking the instructor," and not the instructor's perceived "expertise in the subject matter" or "ability to communicate." Then, it might be inappropriate for you to label the conceptual variable as "student perceptions of instructors." When such a circumstance arises, you need to narrow your conceptual definition to correspond better with the existing operations, such as "liking the instructor." Otherwise, the synthesis' conclusions might appear to apply more generally, to cover more operations, than warranted by the evidence.

The opposite problem can also confront synthesists—that is, they start with narrow concepts but then find measures in the literature that could support broader definitions. For example, this might occur if you find many studies of the effects of teachers' expectations on SAT scores when you initially intended to confine your operational definition of intelligence to IQ scores. You would then face the choice of either broadening the allowable measures of intelligence, and therefore perhaps the conceptual definition, or excluding many studies that others might deem relevant.

Thus, it is not unusual for a dialogue to take place between research synthesists and the research literature. The dialogue can result in an iterative redefining of the conceptual variables that are the focus of the synthesis as well as the included empirical realizations. As the literature search proceeds, it is extremely important that synthesists take care to reevaluate the correspondence between the breadth of their concepts and the variation in ideas and operations that primary researchers have used to define them.

Before leaving the issue of concept-to-operation-fit, one more point is important to emphasize. The dialogue between synthesists and research literatures needs to proceed on the basis of trying to provide precise conceptual definitions that lead to clear linkages between concepts and operations in a way that allows meaningful interpretation of results by the audience of the synthesis. What this means is that synthesists must never let the *results* of the primary research dictate what operations will be used to define a concept. For example, it would be inappropriate for you to decide to operationally define intelligence as cognitive abilities measured by IQ scores and leave out results involving the Miller Analogies Test because the IQ tests revealed results consistent with your hypothesis and the MAT did not. The relevance of the MAT must be based on how well its contents correspond to your conceptual definition, and that alone. If this test of intelligence produces results different from other tests, then the reason for the discrepancy should be explored as part of the research synthesis, not used as a rationale for the exclusion of MAT studies.

2.2.2.2 Multiple Operations and Concept-to-Operation Fit Multiple operations also create opportunities for research synthesists. Eugene Webb and his colleagues present strong arguments for the value of multiple operations (2000). They define multiple operationism as the use of many measures that share a conceptual definition but that have different patterns of irrelevant components. Multiple operationism has positive consequences because

> once a proposition has been confirmed by two or more independent measurement processes, the uncertainty of its

> interpretation is greatly reduced. . . . If a proposition can survive the onslaught of a series of imperfect measures, with all their irrelevant error, confidence should be placed in it. Of course, this confidence is increased by minimizing error in each instrument and by a reasonable belief in the different and divergent effects of the sources of error. (Webb et al. 2000, 35)

So, the existence of a variety of operations in research literatures offers the potential benefit of stronger inferences if it allows the synthesists to rule out irrelevant sources of influence. However, multiple operations do not ensure concept-to-operation correspondence if all or most of the operations lack a minimal correspondence to the concept. For example, studies of teacher expectation effects may manipulate expectations in a variety of ways—sometimes by providing teachers with information on students in the form of tests scores, sometimes by reports from psychologists—and measure intelligence in a variety of ways. If results do not vary systematically as a function of the operations used, then the finding that expectations influence intelligence has passed a test of robustness. If the results are systematically related to operations, then qualifications to the finding are in order. For example, if you find that teacher expectations influence SAT scores but not IQ scores, you might postulate that perhaps expectations affect knowledge acquisition but not mental acuity. This inference is less strong because it requires you to develop a post hoc explanation for your findings.

2.2.2.3 Use of Operations Meant to Represent Other Concepts Literature searches of reference databases typically begin by using keywords that represent the conceptual definitions of the variables of interests. Thus, you are much more likely to begin a search by crossing the keywords "teacher expectations" with "intelligence" than you are by crossing "Test of Intellectual Growth Potential" (a made-up test that might have been used to manipulate teacher expectations) with "Stanford-Binet" and "Wechsler." It is the abstractness of the keywords that allows unexpected operationalizations to get caught in the search net.

By extending the keywords even further, synthesists may uncover research that has been cast in conceptual frameworks different from their own but that include manipulations or measures relevant to the concepts the synthesist has in mind. For instance, several concepts similar to *interpersonal expectancy effects* appear in the research literature—one is *behavior confirmation*—but come from different disciplinary traditions. Even though these concepts are labeled differently, the operations used in the studies they generate may be very similar and are certainly relevant to one another. When relevant operations associated with different constructs are identified, they should be included in the synthesis. In fact, similar operations generated by different disciplinary traditions often do not share other features of research design—for example, they may draw participants from different populations—and therefore can be used to demonstrate the robustness of results across methodological variations. Synthesists can improve their chances of finding broadly relevant studies by searching in reference databases for multiple disciplines and by using the database thesauri to identify related, broader, and narrower terms.

2.2.3 Effects of Multiple Operations on Synthesis Outcomes

Multiple operations do more than introduce the potential for more robust inferences about relationships between conceptual variables. They are an important source of variance in the conclusions of different syntheses meant to address the same topic. A variety of operations can affect synthesis outcomes in at least two ways.

First, the operational definitions covered in two research syntheses involving the same conceptual variables can be different from one another. Thus, two syntheses claiming to integrate the research on the effects of teacher expectations on intelligence can differ in the way intelligence is operationally defined. One might include SAT and GRE scores along with IQ test results and the other only IQ tests. Each synthesis may contain some operations excluded by the other, or one definition may completely contain the other.

Second, multiple operations affect outcomes of syntheses by leading to variation in the way study operations are treated *after* the relevant literature has been identified. Some synthesists pay careful attention to study operations. They identify meticulously the operational distinctions among retrieved studies and test for whether results are robust across these variations. Other synthesists pay less attention to these details.

2.2.4 Variable Definitions and the Literature Search

I mention one way in which the choice of keywords in a reference database search can influence the studies that are uncovered. I point out that entering databases with conceptual terms as keywords will permit the serendipitous

discovery of aligned areas of research, more so than entering operational descriptions will. Now I extend this recommendation further: synthesists should begin the literature search with the broadest conceptual definition in mind. While reading study abstracts, the synthesists should use the tag "potentially relevant" as liberally as possible. At later stages of the synthesis, it is okay to exclude particular operations for their lack of correspondence with the precise conceptual definition. However, in the early stages, when conceptual and operational boundaries may still be a bit fuzzy, the synthesists should err on the overly inclusive side, just as primary researchers collect some data that might not later be used in analysis.

This strategy initially creates more work for synthesists but has several long-term benefits. First, it keeps within easy reach operations that on first consideration may be seen as marginal but later jump the boundary from irrelevant to relevant. Reconstituting the search because relevant operations have been missed consumes far more resources than first putting them in the Inbox but excluding them later.

Second, a good conceptual definition speaks not only to which operations are considered relevant but also to which are irrelevant. By beginning a search with broad terms, synthesists are forced to struggle with operations that are at the margins of their conceptual definitions. Ultimately, this results in definitions with more precise boundaries. For example, if you begin a search for studies of teacher expectation effects with the keyword "intelligence" rather than "IQ", you may decide to include research using the MAT, thus broadening the intelligence construct beyond IQ, but exclude the SAT and GRE, because they are too influenced by knowledge, rather than mental acuity. But, by explicitly stating these tests were excluded, the audience of the synthesis gets a better idea of where the boundary of the definition lies, and can argue otherwise if they choose. Had you started the search with "IQ" you might never have known that others considered the SAT and GRE tests of intelligence in teacher expectation research.

Third, a broad conceptual search allows the synthesis to be carried out with greater operational detail. For example, searching for and including IQ tests only permits us to examine variations in operations, such as whether the test was group versus individually administered and timed versus untimed. Searching for, and ultimately including, a broader range of operations permits the synthesists to examine broader conceptual issues when they cluster findings according to operations and look for variation in results. Do teacher expectations produce different effects on IQ tests and the SAT? If so, are there plausible explanations for this? What does it mean for the definition of intelligence? Often, these analyses produce the most interesting results in a research synthesis.

2.3 TYPES OF RESEARCH PROBLEMS AND HYPOTHESES

After researchers have defined their concepts and operations, they need to define the problem or hypothesis of interest. Does the problem relate simply to the prevalence or level of a phenomenon in a population? Does it relate to how a person acting in the environment interprets their experience? Does it suggest that the variables of interest relate to one another as a simple association or with a causal connection? Does it refer to a process that operates at the level of an individual unit—how it changes over time—or to general tendencies within and between groups of units, and how they might differ from one another?

The distinctions embodied in these questions have critical implications for the choice of appropriate research designs. I assume you are familiar with research designs and their implications for drawing inferences. Still, a brief discussion of the critical features of research problems as they relate to research design is needed to understand the role of design variations in research synthesis.

2.3.1 Three Questions About Research Problems and Hypotheses

Researchers need to answer three questions about their research problems or hypotheses to be able to determine the appropriateness of different research designs for addressing them (for a fuller treatment of these questions, see Cooper 2006):

- Should the results of the research be expressed in numbers or narrative?
- Is the problem or hypothesis meant to uncover a description of an event, an association between events, or an explanation of an event?
- Does the problem or hypothesis seek to understand how a process unfolds *within* an individual unit over time, or what is associated with or explains variation *between* units or groups of units?

If the answer to the first question is narrative, the synthesists will likely focus on qualitative research, or more

accurately, interpretive research (some forms of interpretive research are steeped in quantitative analysis, for example, discourse analysis). Carla Willig suggests that these approaches to research are most appropriate when the researchers want to uncover the impetus to behavior that exists for the actors themselves rather than test the explanatory power of their own (or someone else's) theory or perspective (2012).

For the remainder of this chapter, I assume that the first question was answered "numbers." However, the synthesis of narrative, or qualitative, research is an important area of methodology and scholarship (for detailed examinations of approaches to synthesizing qualitative research, see Sandelowski and Barroso 2007; Pope, Mays, and Popay 2007).

2.3.1.1 Descriptions, Associations, and Explanations First, a research problem might be largely descriptive. Such problems typically ask, "What is happening?" As in survey research, the desired description could focus on a few specific characteristics of the event or events of interest and broadly sample those few aspects across multiple event occurrences (Fowler 2014). Thus you might ask a sample of undergraduates, perhaps drawn randomly across campuses and in different subject areas, how much they like their instructors. From this, you might draw an inference about how well liked instructors generally are on college campuses.

A second type of research problem might be "What events happen together?" Here, researchers ask whether characteristics of events or phenomena co-occur with one another. A correlation coefficient might be calculated to measure the degree of association. For example, your survey of undergraduates' liking for their instructors might also ask the students how well they did in their classes. Then, the two variables could be correlated to answer the question, "Are students' liking of instructors associated with their class grades?"

The third research problem seeks an explanation for an event and might be phrased, "What events cause other events to happen?" In this case, a study is conducted to isolate and draw a direct productive link between one event (the cause) and another (the effect). In the example, you might ask, "Does increasing the liking students have for their instructors cause them to get higher grades in class?"

Three classes of research designs are used most often to help make these causal inferences. I call the first modeling research. It takes simple correlational research a step further by examining co-occurrence in a multivariate framework. For example, if you wish to know whether liking an instructor causes students to get higher grades, you might construct a multiple regression equation or structural equation model that attempts to provide an exhaustive description of a network of relational linkages (Kline 2011). This model would attempt to account for, or rule out, all other co-occurring phenomena that might explain away the relationship of interest. Likely, the model will be incomplete or imperfectly specified, so any casual inferences from modeling research will be tentative at best. Studies that compare groups of participants to one another—such as men and women or different ethnic groups—and control for other co-occurring phenomena are probably best categorized as modeling studies.

The second class of explanatory research designs involves quasi-experimental research (Shadish, Cook, and Campbell 2002). Here, the researchers (or some other external agent) control the introduction of an event, often called an intervention in applied research or a manipulation in basic research, but not precisely who may be exposed to it. Instead, the researchers use some statistical control in an attempt to equate the groups receiving and not receiving the intervention. It is difficult to tell how successful the attempt at equating groups has been. For example, you might be able to train doctors to ask specific questions when diagnosing an ailment. However, you might not be able to assign doctors to receive this training (they would be free to take part or not) so you might match doctors on educational background and ignore those who are not a good match.

Finally, in experimental research both the introduction of the event (such as training in diagnosis) and who is exposed to it (which doctors) are both controlled by the researcher (or other external agent). The researcher uses a random procedure to assign subjects to conditions, leaving the assignment to chance (Christensen 2012). In the example, because the random assignment procedure minimizes average existing differences between doctors with regard to their types and amounts of education prior to the experiment, you can be most confident that any differences between the diagnoses are caused by the new training rather than other potential explanations. Of course, numerous other aspects of the design must be attended to for this strong inference to be made—for example, standardizing office visit conditions across doctors and ensuring that doctors are unaware of the experimental hypothesis. For our purposes, however, focusing on the characteristics of researcher control over the conditions of the experiment and assignment of participants

to conditions captures what is needed to proceed with the discussion.

2.3.1.2 Examining Change Within Units or Variation Across Units Some research problems reference changes that occur within a unit over time. Others relate to the average differences and variation in a characteristic between groups of units. This latter problem is best addressed using the designs just discussed. The former—the problem of change within a unit—would best be studied using the various forms of single-case designs, a research design in which units are tested at different times, typically equal intervals, during the course of a study. For example, you might ask a student in a class how much he or she liked an instructor after each weekly class meeting. You might also collect the student's grade on that week's homework assignment. A concomitant time series design would then reveal the association between change in liking and homework grades.

If the research hypothesis seeks to test for a causal relationship between liking and grades, it would call for charting a student's change in grades, perhaps on homework assignments or in-class quizzes, before and after an experimental manipulation changed the way the instructor behaved, perhaps from being distant to being friendly. Or, the change from distant to friendly instruction might involve having the instructor add a different "friendly teaching technique" with each class session, for example, leaving time for questions during session 2, adding revealing personal information during session 3, adding telling jokes during session 4. Then you would chart how each additional "friendly" teaching technique affected a particular student's grades on each session's homework assignment.

Of course, for this type of design to allow strong causal inferences other experimental controls would be needed—for example, the random withdrawal and reintroduction of teaching techniques. For our purposes, however, the key point is that it is possible that how each individual student's grades change as a function of the different teaching techniques could look very different from the moving group average of class-level homework grades. For example, the best description of how instructor friendliness affected grades based on how individual students react might be to say, "Each student's performance improved precipitously after the introduction of a specific 'friendly' teaching technique that seemed to appeal to that student while other techniques produced little or no change." However, because different students might have found different approaches appealing, averaged across the students in the class, the friendliness manipulations might be causing gradual improvements in the average homework grade. At the class level of analysis the best description of the results might be, "The gradual introduction of additional friendly teaching techniques led to a gradual improvement in the classes' average homework grades." So, the group-averaged data can be used to explain improvement in grades only at the group level; it is not descriptive of the process happening at the individual level, nor vise versa.

The designation of whether an individual or group effect is of interest depends upon the research question being asked. So, the group performance in this example is also a description of how one classroom might perform. This would be the case if an instructor wanted to raise her or his likability but was not interested in whether or how it influenced any particular student. Thus, the key to proper inference is that the unit of analysis at which the data from the study is being analyzed corresponds to the unit of interest in the problem or hypothesis motivating the primary research or research synthesis. If not, an erroneous conclusion is possible.

2.3.2 Problems and Hypotheses in Research Synthesis

2.3.2.1 Synthesizing Descriptive Research When a description of the quantitative frequency or level of an event is the focus of a research synthesis, it is possible to integrate evidence across studies by conducting what Robert Rosenthal calls aggregate analysis (1991). Floyd Fowler suggests that two types of descriptive questions are most appropriately answered using quantitative approaches (2014). The first involves estimating the frequency of an event's occurrence. For example, you might want to know, "How many college students receive a grade of A in their classes?" The second descriptive question involves collecting information about attitudes, opinions, or perceptions. For example, your synthesis concerning liking of instructors might be used to answer the question, "On average, how well do college students like their instructors?"

Conducting a synthesis of research on the question involving frequency of grades would lead you to collect from each relevant study the number of students receiving grades of A and the total number of students in each study. The question regarding opinions would lead you to

collect the average response of students on a question about "liking" of their instructor.

The primary objective of the studies you are aggregating data from might not have been to answer your descriptive questions. For example, this evidence (on frequency of grades, liking of instructor) could have been reported as part of a study examining teacher expectation effects. However, it is not unusual for such studies to report, along with the effect of interest, the distribution of grades in the classes under study and even the average liking of the instructors (perhaps to test as a moderator of the expectation effect). Further, this might be done because the primary researchers wanted to examine the data for restrictions in range ("Was there enough variation in grades and liking to permit a strong test of their association?") or to describe their sample for purposes of assessing the generality of their findings ("How were the students in this study doing in school relative to the more general population of interest?").

The difficulty of aggregating descriptive statistics across studies depends on whether frequencies or attitudes, opinions, and perceptions are at issue. With regard to frequencies, as long as the event of interest is defined similarly across studies, it is simple to collect the frequencies, add them together, and, typically, report them as a proportion of all events. But two complications remain.

First, in using the proviso "defined similarly" I am making the assumption that, for example, it is the act of getting a grade of A that interests you, not that getting an A means the same thing in every class used in every study. If instructors use different grading schemes, getting an A might mean very different things in different classes and different studies. Of course, if different studies report different frequencies of A grades, you can explore whether other features of studies (such as the year in which the study was conducted, if you are interested in grade inflation) covary with the grading curve used in their participating classrooms.

Second, it might be that the aggregation of frequencies across studies is undertaken to estimate the frequency of an event in a population. So, your motivating question might be, "What is the frequency with which the grade of A is assigned in American college class?" The value of your aggregate estimate from studies will then depend on how well the studies represented American college classes. It would be rare for classes chosen because they are convenient to be fortuitously representative of the nation as a whole. But, it might be possible to apply some sampling weights that would improve your approximation of a population value.

Aggregating attitudes, opinions, and perceptions across studies is even more challenging. This is because social scientists often use different scales to measure these variables, even when the variables have the same conceptual definition. Suppose you wanted to aggregate across studies the reported levels of liking of the instructor. It would not be unusual to find that some studies simply asked, "Do you like your instructor, yes or no?" and that others asked, "How much do you like your instructor?" but used different scales to measure liking. Some might have used 10-point scales while others used 5-point or 7-point scales. Scales with the same numeric gradations might also differ in the anchors they used for the "liking" dimension. For example, some might have anchored the positive end of the dimension with "a lot," others with "very much," and still others with "my best instructor ever." In this case, even if all the highest numerical ratings are identical, the meaning of the response is different.

Clearly, you cannot simply average these measures. One solution would be to aggregate results across studies using the identical scales and reporting results separately for each scale type. Another solution would be to report simply the percentage of respondents using one side of the scale or the other. For example, your cumulative statement might be "Across all studies, 85 percent of students reported positive liking for their instructor." This percentage is simply the number of students whose ratings are on the positive side of the scale, no matter what the scale, divided by the total number of raters. The number can be derived from the raw rating frequencies, if they are given, or from the mean and variance of the ratings (see Cooper et al. 2003, for an example).

In general, it is rare to see research syntheses in social science that seek to aggregate evidence across studies to answer descriptive questions quantitatively. However, when this occurs, it is critical that the synthesists pay careful attention to whether the measures being aggregated are commensurate. Combining measures that are incommensurate will result in gibberish.

2.3.2.2 Synthesizing Group-Level Associations and Causal Relationships The accumulation and integration of comparisons between group-level statistics for testing associations and causal relationships has been by far the most frequent objective of research syntheses in social science. In primary research, a researcher selects the most appropriate research design for investigating the

problem or hypothesis. It is rare for a single study to implement more than one design to answer the same research question, although multiple operationalizations of the constructs of interest are possible. More typical would be instances in which primary researchers intended to carry out an optimal design for investigating their problem but, due to unforeseen circumstances, end up implementing a less-than-optimal design. For example, if you are interested in asking "Does increasing the liking students have for their instructors cause students to do better in class?" you might begin by intending to carry out an experiment in which high school students are randomly assigned to friendly and distant instructors (who are also randomly assigned to teaching style). However, as the semester proceeds, some students move out of the school district, some change sections because of conflicts with other courses, and others change sections because they do not like the instructor. By the end of the semester, it is clear that the students remaining in the sections were not equivalent, on average, when the study began. To rescue the effort, you might use post hoc matching or statistical controls to re-approximate equivalent groups. Thus, the experiment has become a quasi-experiment. If students in the two conditions at the end of the study are so different on pretests that procedures to produce post hoc equivalence seem inappropriate (for example, lead to small groups of students sampled from different tails of the class distributions), you might simply correlate the students' ratings of liking (originally meant to be used as a manipulation check in the experiment) with grades. So, a study searching for a causal mechanism has become a study of association (with a known confound). The legitimate inferences allowed by the study began as strong causal inference, degraded to weak causal inference, and ended as simple association.

In contrast, research synthesists are likely to come across a wide variety of research designs that relate the concepts of interest to one another. A search for studies using the term "instructor ratings" and examining them for measures of "achievement" should identify studies that resulted in simple correlations, multiple regressions, perhaps a few structural equation models, some quasi-experiments, a few experiments, and maybe even some time series involving particular students or class averages.

What should the synthesists do with this variety of research designs? Certainly, your treatment of them depends primarily on the nature of the research problem or hypothesis. If your problem is whether there is an association between the liking students have for their instructors and grades in the course, then it seems that any and all of the research designs address the issue. The regressions, structural equation models, quasi-experiments, and experiments ask a more specific question about association—these seek a causal connection or one with other explanations ruled out—but they are tests of an association nonetheless. Thus, it seems that you would be well advised to include all the research designs in your synthesis of evidence on association.

The issue is more complex when your research question deals with causality: "Does increasing the liking students have for their instructors cause higher student grades?" Here, the different designs produce evidence with different capabilities for drawing strong inferences about your problem. Again, correlational evidence addresses a necessary but not sufficient condition for drawing a causal inference. Thus, if this were the only research design found in the literature, it would be appropriate for you to assert that the question remained untested. When an association is found, multiple regressions statistically control for some alternative explanations for the relationship, but probably not all of them. Structural equation models relate to the plausibility of causal networks but do not address causality in the generative sense, that is, manipulating one variable will produce a change in the other variable. Well-conducted quasi-experiments may permit weak causal inferences, made stronger through multiple and varied replications. Experiments permit the strongest inferences about causality.

How should synthesists interested in problems of causality treat the various designs? At one extreme, they can discard all studies but those using true experimental designs. This approach applies the logic that these are the only studies that directly test the question of interest. All other designs either address association only or do not permit strong inferences. The other approach would be to include all the research evidence while carefully qualifying inferences as the ability of the design for providing evidence for causality moves farther from the ideal. A less extreme approach would be to include some but perhaps not all designs while again carefully qualifying inferences.

Arguments support each of these approaches. The first obligation of synthesists is to clearly state the approach they have used and the rationale for it. In research areas where strong experimental designs are relatively easy to conduct and plentiful—such as research on the impact of aerobic exercise on the cognitive functioning of older adults (Smith et al. 2010)—I have been persuaded that

excluding designs that permit only weak causal inferences was an appropriate approach to the evidence.

In other instances, experiments may be difficult to conduct and rare—such as the impact of homework on achievement (Cooper et al. 2006). Here, the synthesists are forced to make a choice between two philosophically different approaches to evidence. If the synthesists believe that discretion is the better part of valor, then they might opt for including only the few experimental studies or stating simply that little credible evidence on the causal relationship exists. Alternatively, if they believe that any evidence is better than no evidence at all, then they might proceed to summarize the less-than-optimal studies, with the appropriate precautions.

Generally speaking, when experimental evidence on causal questions is lacking or sparse, a more inclusive approach is called for, assuming that the synthesists pay careful and continuous attention to the impact of research design on the conclusions that they draw. In fact, the inclusive approach can provide some interesting benefits to inferences. Returning to the example of instructor liking, you might find a small set of studies in which the likability of the instructor has been manipulated and students assigned randomly to conditions. However, to accomplish the manipulation, these studies might have been conducted in courses that involved a series of guest lecturers, used to manipulate instructor's likability. Grades were operationalized as scores on homework assignments turned in after each lecture. These studies might have demonstrated that more likable guest lecturers produced higher student grades on homework. Thus, to carry out the manipulation it was necessary to study likability in short-term instructor-student interactions. Could it be that over time—more like the real-world relationships that develop between instructors and students—likability becomes less important and perceived competence becomes more important? Could it be that the effect of likability is short lived—it appears on homework grades immediately after a class but does not affect how hard a student studies for exams and therefore has much less, if any, effect on test scores and final grades?

These issues, related to construct and external validity, go unaddressed if only the experimental evidence is permitted into the synthesis. Instead, you might use the non-experimental evidence to help you gain tentative, first approximations about how the likability of instructors plays out over time and within broader constructions of achievement. The quasi-experiments found in the literature might use end-of-term class grades as outcome measures. The structural equation models might use large, nationally representative samples of students and relate ratings of liking of instructors in general to broad measures of achievement, such as cumulative grade point averages and SAT scores.

By using these results to form a web of evidence, you can come to more or less confident interpretations of the experimental findings. If the non-experimental evidence reveals relationships consistent with the experiments, you can be more comfortable in suggesting that the experimental results generalize beyond the specific operations used in the experiments. If the evidence is inconsistent, it should be viewed as a caution to generalization.

In sum, it is critical in both primary research and research synthesis that the type of relationship between the variables of interested be clearly specified. This specification dictates whether any particular piece of primary research has used a research design appropriate for the research question. Designs appropriate to gather data on one type of problem may or may not provide information relevant for another type. Typically, in primary research only a single research design can be used in each study. In research synthesis, however, a variety of research designs relating the variables of interest are likely. When the relation of interest concerns an association between variables, designs that seek to rule out alternative explanations or establish causal links are still relevant. When the problem of interest concerns establishing a causal link between variables, designs that seek associations or seek to rule out a specified set of but not all alternative explanations do not match the question at hand. Still, if the constraints of conducting experiments place limits on these studies' ability to establish construct and external validity, a synthesis of the non-experimental work can give a tentative, first approximation of the robustness, and limits, of inferences from the experimental work.

2.3.2.3 Synthesizing Studies of Change in Units Across Time The study of how individual units change over time and what causes this change has a long history in the social sciences. Single-case research has developed its own array of research designs and data analysis strategies (see, for example, Kazden 2011). The issues I have discussed regarding descriptive, associational, and causal inferences and how they relate to different research designs play themselves out in the single-case arena in a manner similar to that in group-level research. Different time series designs are associated with different types of research problems and hypotheses. For example, the question "How does a student's liking for the instructor

change over the course of a semester?" would be most appropriately described using a simple time series design. The question "Is a student's liking for the instructor associated with the student's grades on homework as the semester progresses?" would be studied using a concomitant time series design. The question "Does a change in a student's liking for the instructor cause a change in the student's homework grade?" would be studied using any of a number of interrupted time series designs. Much like variations in group designs, variations in these interrupted time series would result in stronger or weaker inferences about the causal relationship of interest.

The mechanics of quantitative synthesis of single-case designs requires its own unique toolbox (for examples, see Shadish and Rindskopf 2007). However, the logic of synthesizing single-case research is identical to that of group-level research. It would not be unusual for synthesists to uncover a variety of time series designs relating the variables of interest to one another. If you are interested in whether a change in a student's liking of the instructor caused a change in the student's course grade, you may very well find in the literature some concomitant time series, some interrupted time series with a single "liking" intervention, and perhaps some designs in which liking is enhanced and then lessened as the semester progresses. Appropriate inferences drawn from each of these designs correspond more or less closely with your causal question. You may choose to focus on the most correspondent designs, or to include designs that are less correspondent but that do provide some information. The decision about which of these courses of action are most appropriate again should be influenced by the number and nature of studies available. Of course, regardless of the decision rule adopted concerning the inclusion of research designs, synthesists are obligated to carefully delimit their inferences based on the strengths and weaknesses of the included designs.

2.3.2.4 Problems and Hypotheses Involving Interactions At their broadest level, the problems and hypotheses that motivate most research syntheses involve main effects, the relations between two variables. Does liking of the instructor cause higher class grades? Do teachers' expectations influence intelligence test scores? This is due to the need to establish such fundamental relationships before investigating the effects of third variables on them. Of course, research can examine multiple bivariate relations at once (with the question, for example, "What are the determinants of class grades?").

The preponderance in practice of main effect questions does not mean that a focus on an interaction cannot or should not motivate a research synthesis. For example, it may be that a bivariate relationship is so well established that an interaction hypothesis has become the focus of attention in a field. That liking the instructor causes higher grades in high school may be a finding little in dispute, perhaps because a previous research synthesis has shown it to be so. Now the issue is whether the causal impact is equally strong across classes dealing with language or mathematics. So, you undertake a new synthesis to answer this interaction question: "Is the effect of liking on grades equally strong across different subject areas?"

Also, undertaking a synthesis to investigate the existence of a main effect relationship should in no way diminish the attention paid to interactions within the same synthesis project. Indeed, in my earlier example, the fact that the previous research synthesis on liking of the instructor causing higher grades in high school did not look at the moderating influence of subject matter might be considered a shortcoming of that synthesis. Typically, when main effect relationships are found to be moderated by third variables, these findings are given inferential priority and viewed as a step toward understanding the processes involved in the causal chain of events.

Examining interactions in research syntheses presents several unique opportunities and unique problems. So that both can be fully understood, I introduce yet another critical distinction in the types of evidence that can be explored in research synthesis.

2.4 STUDY-GENERATED AND SYNTHESIS-GENERATED EVIDENCE

Research synthesis can contain two sources of evidence about the research problem or hypothesis. The first is called study-generated evidence. Study-generated evidence is present when a single study generates results that directly test the relation being considered. Research syntheses also include evidence that does not come from individual studies but instead from the variations in procedures across studies. This type of evidence, called synthesis-generated evidence, is present when the results of studies using different procedures to test the same hypothesis are compared with one another.

Any research problem or hypothesis—a description, simple association, or causal link—can be examined through either study-generated or synthesis-generated evidence. However, only study-generated evidence based

on experimental research allows synthesists to make statements concerning causality.

2.4.1 Study-Generated Evidence on Main Effect Relationships

Suppose you were interested in synthesizing the research on whether teacher expectations influence IQ test scores. You conduct a literature search and discover twenty studies that randomly assigned students to one of two conditions. Teachers for one group of students were given information indicating that they could expect unusual intellectual growth during the school year from a randomly chosen sample of students in their class. They were given no out-of-the-ordinary information about the other students. These twenty studies each provide study-generated evidence regarding a simple two-variable relationship, or main effect. The cumulative results of these studies comparing the end-of-year IQ scores of students in the two groups could then be interpreted as supporting or not supporting the hypothesis that teacher expectations influence IQ.

2.4.2 Study-Generated Evidence on Interactions

Next, assume you are interested as well in whether teacher expectations manipulated through the presentation of a bogus Test of Intellectual Growth Potential has more of an impact on student IQ scores than the same manipulation created by telling teachers simply that students' past teachers predicted unusual intellectual growth based on subjective impressions. You discover twenty experimental studies used both types of expectation manipulation. These studies provide study-generated evidence regarding an interaction. Here, if you found that test scores produced a stronger expectation-IQ link than past teachers' impressions, you could conclude that the mode of expectation induction caused the difference. This is because the type of manipulation used is not confounded with other study characteristics. For example, while the grade levels of students might differ from study to study, grade level would be equally represented in each type of manipulation group. Of course, the same logic would apply to attempts to study higher order interactions—involving more than one interacting variable—although these are still rare in research syntheses.

2.4.2.1 Integrating Interaction Results Across Studies Often, the integration of interaction results in research synthesis is not as simple as combining significance levels or calculating and averaging effect sizes from each study. Figure 2.1 illustrates the problem by presenting the results of two hypothetical studies comparing the effects of manipulated (high) teacher expectations on student IQ scores. You want to examine whether the expectation effect was moderated by the number of students in the class. In study 1, involving, say, ten classes, the sizes of the class in which expectations were manipulated ranged from ten to twenty-eight. A significant interaction was found suggesting larger expectation effects in smaller class. Study 2 also involved twenty classes but the class sizes ranged only from ten to twenty students. Study 2 might have reported a significant main effect only.

You might be tempted to conclude that the two studies produced inconsistent results regarding the existence of an interaction involving class size. However, a closer examination of the two figures illustrates why this might not be an appropriate interpretation. The results of study 2 probably would have closely approximated those of study 1 had the class sizes in study 2 represented the same range of values as those in study 1. Note that the slopes for

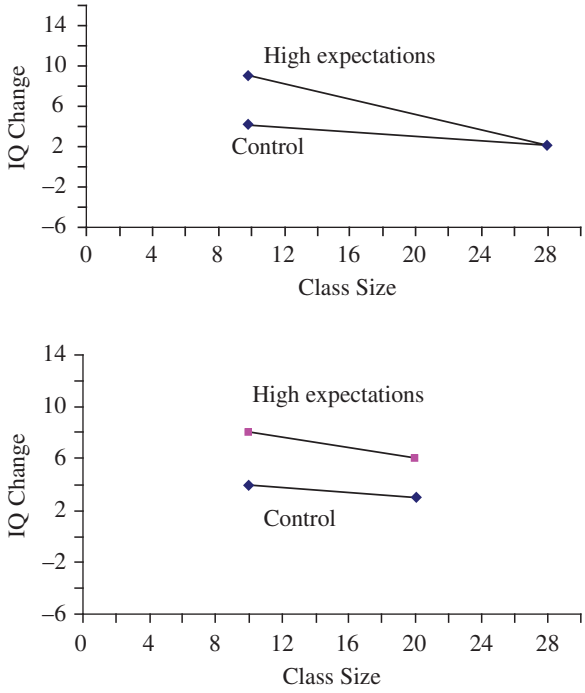

Figure 2.1 Effects of Teacher Expectations on IQ
SOURCE: Author's compilation.

the high expectations groups in study 1 and study 2 are nearly identical, as are those for the control groups.

This example demonstrates that synthesists should not assume that strengths of interaction uncovered by different studies necessarily imply inconsistent results. Synthesists need to examine the differing ranges of values of the variables employed in different studies, be they measured or manipulated. If possible, they should chart results taking the different levels into account, just as I did in figure 2.1. In this manner, one of the benefits of research synthesis is realized. Whereas one study might conclude that the effect of expectations dissipates as class size grows larger and a second study might not, the research synthesists can discover that the two results are in fact perfectly commensurate (if the range in class sizes is used as a mediating variable in a search for influences on the results of studies).

This benefit of research synthesis also highlights the importance of primary researchers presenting detailed information concerning the levels of variables used in their studies. Research synthesists cannot conduct a between-studies analysis similar to my example without this information. If the primary researchers in study 1 and study 2 neglected to specify their range of class sizes, perhaps because they simply said they compared smaller classes with larger classes, the commensurability of the results would have been impossible to uncover.

Variations in ranges of values for variables can also produce discrepancies in results involving two-variable or main effect relationships. I mention it in the case of interactions because this is the circumstance under which the problem is least likely to be recognized and is most difficult to remedy when it is discovered.

2.4.3 Synthesis-Generated Evidence on Main Effect Relationships

Earlier, I provided an example of a two-variable relationship studied with synthesis-generated evidence when I suggested that you might relate the average grade given to students by their instructors to the year in which the study was conducted. You might do so to obtain an indication of whether grades were getting higher over time. Here, you have taken two descriptive statistics from the study reports and related them to one another. This is synthesis-generated evidence for a two-variable relationship, or main effect.

It should be clear that only associations can be studied in this way because you have not randomly assigned studies to conditions. It should also be clear that looking for such relationships between study-level characteristics opens up the possibility of examining a multitude of problems and hypotheses that might never have been the principal focus of individual primary studies.

The use of synthesis-generated evidence to study main effect relationships is rarely the primary focus of research synthesis, although there is no pragmatic reason why this should be the case, other than the weak causal inferences. Study characteristics—that is, evidence at the synthesis level—most often come into play as variables that influence the magnitude of two-variable relationships examined within studies. Conceptually, the difference is simply that in the former case—interaction case—one of the two study characteristics being examined already is some expression of a two-variable relationship.

2.4.4 Synthesis-Generated Evidence on Three-Variable Interactions

Let us return to the teacher expectations and IQ example. Suppose you find no studies that manipulated the mode of expectation induction but you do discover that ten of the studies that experimentally manipulated expectations did so using test scores but ten others used the impressions of previous teachers. When you compare the magnitude of the expectation effect on IQ between the two sets of studies, you discover that the link is stronger in studies using test manipulations. You could then infer that an association exists between mode-of-expectation-induction and IQ but you could not infer a causal relation between the two. This is synthesis-generated evidence for an interaction.

When groups of effect sizes are compared within a research synthesis, regardless of whether they come from simple correlational analyses or controlled experiments using random assignment, the synthesists can establish only an association between a moderator variable—a characteristic of the studies—and the outcomes of studies. They cannot establish a causal connection. Synthesis-generated evidence is restricted to making claims only about associations and not about causal relationships because it is the ability to employ random assignment of participants that allows primary researchers to assume third variables are represented equally in the experimental conditions.

The possibility of unequal representation of third variables across study characteristics cannot be eliminated in your synthesis because you did not randomly assign

experiments to modes of expectation manipulation. For example, it might be that the set of studies using tests to manipulate IQ were also conducted in higher grades. Now, you do not know whether it was the use of tests or the age of the students that caused the difference in the relationship between teacher expectations and IQ. The synthesists cannot discern which characteristic of the studies, or perhaps some unknown other variable related to both, produced the stronger link. Thus, when study characteristics are found associated with study outcomes, the synthesists should report the finding as just that, an association, regardless of whether the included studies tested the causal effects of a manipulated variable or estimated the size of an association.

Synthesis-generated evidence cannot legitimately rule out as possible true causes other variables confounded with the study characteristic of interest. Thus, when synthesis-generated evidence reveals a relationship that would be of special interest if it were causal, the synthesists should include a recommendation that future research examine this factor using a more systematically controlled design so that its causal impact can be appraised. In the example, you would call for a primary study to experimentally manipulate the mode of expectation induction and ensure that lower- and higher-grade classes are randomly assigned the different expectation induction conditions.

2.4.5 High and Low Inference Codes of Study Characteristics

The examples of study characteristics I have used so far might all be thought of as low inference codes. Variables such as class size or the type of IQ measure require the synthesists only to locate the needed information in the research report and transfer it to the synthesis database. In some circumstances, synthesists might want to make more inferential judgments about study operations. These high inference codes typically involve attempting to infer how a contextual aspect of the studies might have been interpreted by participants.

For example, I used the mode-of-manipulation as a study-level variable that might have been examined as a potential moderator of the link between teacher expectations and change in student IQ scores. I used bogus tests and previous teachers' impressions as the two modes of manipulation. Most manipulations of these types could be classified into the two categories with relatively little inference on the part of the synthesists. However, suppose that you found in the literature a dozen types of expectation manipulations, some involving long tests taken in class, others short tests taken after school, some the impressions of past teachers, and others the clinical judgments of child psychologists. It might still be relatively simple to categorize these manipulations at an operational level but the use of the multiple categories for testing meaningful hypotheses at the synthesis level becomes more problematic. Now, you might reason that the underlying conceptual difference between study operations that interests you is the credibility of the manipulation. So, to ask the question "Does the magnitude of the teacher expectation-student IQ link vary with the credibility of the expectation manipulation?" you might want to examine each study report and score the manipulations on this dimension. This high inference code than becomes a study characteristic you relate to study outcomes. Here you are dimensionalizing the manipulations by making an inference about how much credibility the participating teachers might have placed in the information given to them.

High inference codes create a special set of problems for research synthesists. First, careful attention must be paid to the reliability of inference judgments. It is therefore important to show that these judgments are being made consistently both across and within those people making the judgments. Also, judges are being asked to play the role of a research participant—in this example the teachers being given expectation information—and the validity of role-playing methodologies has been the source of much controversy (Greenberg and Folger 1988). However, Norman Miller, Ju-Young Lee, and Michael Carlson have empirically demonstrated that high inference codes can lead to valid judgments (they did so by comparing inferences to manipulation checks) and can add a new dimension to synthesists' ability to interpret literatures and resolve controversies (1991). If high inference information can be validly extracted from articles and the benefit of doing so is clear, then this can be an important technique for exploring problems and hypotheses in research synthesis (see also chapters 8 and 9).

2.4.6 Value of Synthesis-Generated Evidence

In sum, it is critical that synthesists keep the distinction between study-generated and synthesis-generated evidence in mind. Only evidence coming from experimental manipulations within a single study can support assertions concerning causality. However, I do not want my attention to the fact that synthesis-generated evidence

cannot support causal inferences to suggest that this source of evidence should be ignored. As the examples suggest, the use of synthesis-generated evidence allows you to test relations that may have never been examined by primary researchers. In fact, typically synthesists can examine many more potential moderators of study outcomes than have appeared as interacting variables in primary research. Often, these study-level variables are of theoretical importance. For example, it is easy to see how the mode-of-expectation-induction variable relates to the credibility of the information given to teachers and how class size relates to the teachers' opportunities to communicate their expectations to students. By searching across studies for variations in the operationalizations of theoretically relevant construct, synthesists can produce the first evidence on these potentially critical moderating variables. Even though this evidence is equivocal, it is a major benefit of research synthesis and a source of potential hypotheses (and motivation) for future primary research.

2.5 CONCLUSION

A good way to summarize this chapter is to recast its major points by framing them in the context of their impact on the validity of the conclusions drawn in research syntheses. The most central decisions that synthesists make during problem and hypothesis formulation concern, first, the fit between the concepts used in variable definitions and the operational definitions found in the literature and, second, the correspondence between the inferences about relationships permitted by the evidence in the studies at hand and those drawn by the synthesists.

First, synthesists need to clearly and carefully define the conceptual variables of interest. Unlike primary researchers, research synthesists need not complete this task before the search of the literature begins. Some flexibility permits the synthesist to discover operations unknown to them before the literature search began. It also permits them to expand or contract their conceptual definitions so that, in the end, the conceptual definitions appropriately encompass operations present in the literature. If relatively broad concepts accompany claims about the generality of findings not warranted by the operationalizations in hand, then an inferential error may be made. Likewise, relatively narrow concepts accompanied by marginally related operations also may lead to inferential errors. Research synthesists need to expand or contract conceptual definitions or retain or eliminate operations so that the highest correspondence between them has been achieved.

When some discretion exists regarding whether to expand or contract definitions, the first decision rule should be that the results will be meaningful to the audience that is the target of the synthesis. The second decision rule is pragmatic: will expanding the definitions make the synthesis so large and unwieldy that the effort will outstrip the synthesists' resources? If these concerns are negligible, sometimes synthesists will opt to use narrow definitions with only a few operations to define their concepts to ensure consensus about how the concepts are related to observable events. However, there is reason to favor using broader constructs with multiple realizations. Then, numerous rival interpretations for the findings may be tested and ruled out if the multiple operations produce similar results. Also, narrow concepts may provide little information about the generality or robustness of the results. Therefore, the greater the conceptual breadth of the definitions used in a synthesis, the greater the capacity for conclusions that are more general than when narrow definitions are used.

The word *potential* is emphasized because if synthesists only cursorily detail study operations, their conclusions may mask important moderators of results. An erroneous conclusion—that research results indicate negligible differences in outcomes across studies—can occur if different results across studies are masked in the use of very broad categories.

So, holding constant the needs of the audience and assuming adequate resources are available to the synthesists, the availability of ample studies to support a fit between either narrow or broad conceptual definitions suggests that it is most desirable for syntheses to use the broadest possible conceptual definition. To test this possibility, they should begin their literature search with a few central operations but remain open to the possibility that other relevant operations will be discovered in the literature. When operations of questionable relevance are encountered, the synthesist should err toward overly inclusive decisions, at least in the early stages of the project. However, to complement this conceptual broadness, synthesists should be thorough in their attention to possibly relevant distinctions in study characteristics. Any suggestion that a difference in study results is associated with a distinction in study characteristics should be tested using moderator analyses.

The second set of validity issues introduced during problem and hypothesis formulation concerns the nature of the relationship between the variables under study. Only study-generated evidence can be used to support

causal relationships. When the relationship at issue is a causal one, synthesists must take care to categorize research designs according to their strength of causal inference. They might then exclude all but the most correspondent study-generated evidence (that is, experiments using random assignment). Or, they might separately examine less correspondent study-generated evidence and use it to obtain first approximations of the construct and external validity of a finding.

Also, synthesis-generated evidence cannot be used to make causal claims. Thus, synthesists should be careful to distinguish study-generated evidence from synthesis-generated evidence. Of course, this does not mean that synthesis-generated evidence is of little or no use. In fact, synthesis-generated evidence is new evidence, often exploring problems and hypotheses unexplored in primary research. As such, it is the first among many unique contributions that research synthesis makes to the advance of knowledge.

2.6 REFERENCES

Christensen, Larry. 2012. "Types of Designs Using Random Assignment." In *Handbook of Research Methods in Psychology*, edited by Harris Cooper, Paul Camic, Deborah Long, Abigail Panter, David Rindskopf, and Kenneth Sher. Washington, D.C.: American Psychological Association.

Christensen, Larry B., R. Burke Johnson, and Lisa A. Turner. 2014. *Research Methods: Design and Analysis.* Boston, Mass.: Pearson.

Cooper, Harris. 2006. "Research Questions and Research Designs." In *Handbook of Research in Educational Psychology*, 2nd ed., edited by Patricia A. Alexander, Philip H. Winne, and Gary Phye. Mahwah, N.J.: Lawrence Erlbaum and Associates.

Cooper, Harris, Jeffrey C. Valentine, Kelly Charlton, and April Barnett. 2003. "The Effects of Modified School Calendars on Student Achievement and School Community Attitudes: A Research Synthesis." *Review of Educational Research* 73:1–52.

Fowler, Floyd J., Jr. 2014. *Survey Research Methods*, 5th ed. Thousand Oaks, Calif.: Sage Publications.

Greenberg, Jerald, and Robert Folger. 1988. *Controversial Issues in Social Research Methods*. New York: Springer-Verlag.

Kazden, Alan, ed. 2011. *Single Case Research Designs: Methods for Clinical and Applied Settings*, 2nd ed. Oxford: Oxford University Press.

Kline, Rex B. 2011. *Principles and Practices of Structural Equation Modeling*, 3rd ed. New York: Guilford Press.

Mannheim, Karl. 1936. *Ideology and Utopia*. London: Routledge and Kegan Paul.

Miller, Norman, Ju-Young Lee, and Michael Carlson. 1991. "The Validity of Inferential Judgments When Used in Theory-Testing Meta-Analysis." *Personality and Social Psychology Bulletin* 17(3): 335–43.

Pope, Catherine, Nicholas Mays, and Jennie Popay. 2007. *Synthesizing Qualitative and Quantitative Health Evidence: A Guide to Methods.* Berkshire, UK: Open University Press.

Rosenthal, Robert. 1991. *Meta-Analytic Procedures for Social Research.* Newbury Park, Calif.: Sage Publications.

Sandelowski, Margarete, and Julie Barroso. 2007. *Handbook for Synthesizing Qualitative Research.* New York: Springer.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston, Mass.: Houghton Mifflin.

Shadish, William R., and David M. Rindskopf. 2007. "Methods for Evidence-Based Practice: Quantitative Synthesis of Single-Subject Designs." *New Directions for Evaluation* 113 (Spring): 95–109.

Shoemaker, Pamela J., James William Tankard Jr., and Dominic L. Lasorsa. 2004. *How to Build Social Science Theories.* Thousand Oaks, Calif.: Sage Publications.

Smith, Patrick, J., James. A. Blumenthal, Benson M. Hoffman, Harris Cooper, Timothy J. Strauman, Kathleen Welsh-Bohmer, Jeffrey N. Browndyke, and Andrew Sherwood. 2010. "Aerobic Exercise and Neurocognitive Performance: A Meta-Analytic Review of Randomized Clinical Trials." *Psychosomatic Medicine* 72(3): 239–52.

Webb, Eugene J., Donald T. Campbell, Richard D. Schwartz, and Lee Sechrest. 2000. *Unobtrusive Measures.* Thousand Oaks, Calif.: Sage Publications.

Willig, Carla. 2012. "Perspectives on the Epistemological Bases of Qualitative Research." In *Handbook of Research Methods in Psychology*, edited by Harris Cooper, Paul Camic, Deborah Long, Abigail Panter, David Rindskopf, and Kenneth Sher. Washington, D.C.: American Psychological Association.

3

STATISTICAL CONSIDERATIONS

LARRY V. HEDGES
Northwestern University

CONTENTS

3.1 INTRODUCTION

Research synthesis is an empirical process. As with any empirical research, statistical considerations have an influence at many points in the process. Some of these, such as how to estimate a particular effect parameter or establish its sampling uncertainty, are narrowly matters of statistical practice. They are considered in detail later in this book. Other issues, which this chapter addresses, are more conceptual and might best be considered statistical considerations that impinge on general matters of research strategy or interpretation.

3.2 PROBLEM FORMULATION

The formulation of the research synthesis problem has important implications for the statistical methods that may be appropriate and for the interpretation of results. Careful consideration of the questions to be addressed in the synthesis will also have implications for data collection, data evaluation, and presentation of results. In this section, I discuss two broad considerations in problem formulation: universe to which generalizations are referred and the number and source of hypotheses addressed.

3.2.1 Model of Generalization

A central aspect of statistical inference is making a generalization from an observed sample to a larger population or universe of generalization. Statistical methods are chosen to facilitate valid inferences to the universe of generalization. There has been a great deal of confusion about the choice between fixed- and random-effects statistical methods in meta-analysis. Although it is not always conceived in terms of inference models, the universe of generalization is the central conceptual feature that determines the difference between these methods.

The choice between fixed- and random-effects procedures has sometimes been framed as entirely a question of homogeneity of the effect-size parameters. That is, if all of the studies estimate a common effect-size parameter, then fixed-effects analyses are appropriate. However, if evidence indicates heterogeneity among the population effects estimated by the various studies, or heterogeneity remains after conditioning on covariates, then random-effects procedures should be used. Although fixed- and random-effects analyses give similar answers when in fact an effect size is common across a population, the underlying inference models remain distinct.

I argue that the most important issue in determining statistical procedure should be the nature of the inference desired, in particular, the universe to which one wishes to generalize. If the analyst wishes to make inferences only about the effect-size parameters in the set of studies observed (or to a set of studies identical to the observed studies except for uncertainty associated with the sampling of subjects into those studies), this is what I will call a *conditional* inference. One might say that conditional inferences about the observed effect sizes are intended to be robust to the consequences of sampling error associated with sampling of subjects (from the same populations) into studies. Strictly, conditional inferences apply to this collection of studies and say nothing about other studies that may be done later, could have been done earlier, or may have already been done but are not included among the observed studies. Fixed-effects statistical procedures can be appropriate for making conditional inferences.

In contrast, the analyst may wish to make a different kind of inference, one that embodies an explicit generalization beyond the observed studies. In this case, the observed studies are not the only studies that might be of interest. Indeed, the studies observed are of interest only because they reveal something about a putative population of studies that are the real object of inference. If the analyst wishes to make inferences about the parameters of a population of studies that is larger than the set of observed studies and which may not be strictly identical to them, I call this an *unconditional* inference. Random-effects (or mixed-effects) analysis procedures are designed to facilitate unconditional inferences.

3.2.1.1 Conditional Inference Model In the conditional inference model, the universe to which generalizations are made consists of ensembles of studies identical to those in the study sample except for the particular people (or primary sampling units) that appear in the studies. Thus the studies in the universe differ from those in the study sample only as a result of sampling of people into the groups of the studies. The only source of sampling error or uncertainty is therefore the variation resulting from the sampling of people into studies.

In a strict sense the universe is structured—it is a collection of identical *ensembles of studies*, each study in an ensemble corresponding to a particular study in each of the other ensembles. Each of the corresponding studies would have exactly the same effect-size parameter (population effect size). In fact, part of the definition of *identical* (in the requirement that corresponding studies in different ensembles of this universe be identical) is that they have

the same effect-size parameter. Note that the individual effect-size parameters in each ensemble need not be the same (they need not be homogeneous).

The model is called the *conditional* model because it can be conceived as one that holds fixed, or conditions on, the characteristics of studies that might be related to the effect-size parameter. The conditional model in research synthesis is in the same spirit as the usual regression model and fixed-effects analysis of variance in primary research. In the case of regression, the fixed effect refers to the fact that the values of the predictor variables are taken to be fixed (not randomly sampled). The only source of variation that enters into the uncertainty of estimates of the regression coefficients or tests of hypotheses is that due to the sampling of individuals (in particular, their outcome variable scores) *with a given ensemble of predictor variable scores*. Uncertainty due to the sampling of predictor variable scores themselves is not taken into account. To put it another way, the regression model is conditional on the particular ensemble of values of the predictor variables in the sample.

The situation is similar in the fixed-effects analysis. Here the term refers to the fact that the treatment levels in the experiment are considered fixed, and the only source of uncertainty in the tests for treatment effects is a consequence of sampling of a group of individuals (or their outcome scores) within a given ensemble of treatment levels. Uncertainty due to the sampling of treatment levels from a collection of possible treatment levels is not taken into account.

3.2.1.1.1 Inference to Other Cases. In conditional models (including regression and ANOVA) inferences are, in the strictest sense, limited to cases in which the ensemble of values of the predictor variables are represented in the sample. Of course, conditional models are widely used in primary research and the generalizations supported typically are not constrained to predictor values in the sample. For example, generalizations about treatment effects in fixed-effects ANOVA are usually not constrained to apply only to the precise levels of treatment found in the experiment, but are viewed as applying to similar treatments as well even if they were not explicitly part of the experiment. How are such inferences justified? Typically this is on the basis of an a priori (extra-empirical) decision that other levels (other treatments or ensembles of predictor values) are enough like those in the sample that their behavior will be identical. The argument has two variations. One is that a level not in the sample is enough like one or more in the sample that it is judged essentially identical to them (for example, this instance of behavior modification treatment or this twenty-week treatment is essentially the same as its counterpart in the sample). The other variation is that a level that is not in the sample "lies between" values in the sample on some implicit or explicit dimension and thus it is safe to "interpolate" between the results obtained for levels in the sample. For example, suppose we have the predictor values 10 and 20 in the sample, then a new case with predictor value 15 might reasonably have an outcome halfway between that for the two sampled values, or a new treatment might be judged between two others in intensity and therefore its outcome might reasonably be assumed to be between that of the other two. This interpolation between realized values is sometimes formalized as a modeling *assumption* (for example, in linear regression, where the assumption of a linear equation justifies the interpolation or even extrapolation to other data provided they are sufficiently similar to fit the same linear model). In either case, the generalization to levels not present in the sample requires an assumption that the levels are similar to those in the sample—one not justified by a formal sampling argument.

Inference to studies not *identical* to these in the sample can be justified in meta-analysis by the same intellectual devices used to justify the corresponding inferences in primary research. Specifically, inferences may be justified if the studies are judged a priori to be "sufficiently similar" to those in the study sample. The inference process has two distinct parts. One is the generalization from the study sample to a universe of identical studies, which is supported by a sampling theory rationale. The second is the generalization from the universe of studies that are identical to the sample to a universe of sufficiently similar but not identical studies. This second part of the generalization is not strictly supported by a sampling argument but by an extra-statistical one.

3.2.1.2 Unconditional Inference Model In the unconditional model, the study sample is presumed to be literally a sample from a hypothetical collection (or population) of studies. The universe to which generalizations are made consists of a population of studies from which the study sample is drawn. Studies in this universe differ from those in the study sample along two dimensions: study characteristics and effect-size parameters. The generalization is not, as it was in the fixed-effects case, to a universe consisting of ensembles of studies with corresponding members of the ensembles having identical characteristics and effect-size parameters. Instead, the studies in the study sample (and their effect-size parameters) differ

from those in the universe by as much as might be expected as a consequence of drawing a sample from a population. Second, the studies in the study sample differ from those in the universe because of the sampling of people into the groups of the study. This results in variation of observed effect sizes about their respective effect-size parameters.

We can conceive of these two dimensions as introducing two sources of variability into the observed (sample) effect sizes in the universe. One is due to variation in observed (or potentially observable) study effect sizes about their effect-size parameters. This variability is a result of the sampling of people into studies and is the only variability conceived as random in the conditional model. The other is variation in effect-size parameters across studies.

This model is called the unconditional model because, unlike the conditional model, it does not condition (or hold fixed) characteristics of studies that might be related to the effect-size parameter. The random-effects model in research synthesis is in the same spirit as the correlation model or the random-effects analysis of variance in primary research. In the correlation model both the values of predictor variable and those of the dependent variable are considered to be sampled from a population—in this case, one with a joint distribution. In the random-effects analysis of variance, the levels of the treatment factor are sampled from a universe of possible treatment levels (and consequently, the corresponding treatment effect parameters are sampled from a universe of treatment effect parameters). There are two sources of uncertainty in estimates and tests in random-effects analyses. One is due to sampling of the treatment effect parameters themselves and the other to the sampling of individuals (in particular, outcome scores) into each treatment.

3.2.1.2.1 Inference to Other Cases. In the unconditional model, inferences are not limited to cases with predictor variables represented in the sample. Instead, for example, inferences about the mean or variance of an effect-size parameter apply to the universe of studies from which the study sample was obtained. In effect, the warrant for generalization to other studies is via a classical sampling argument. Because the universe contains studies that differ in their characteristics, and those differences find their way into the study sample by the process of random selection, generalizations to the universe pertain to studies that are not identical to those in the study sample.

By using a sampling model of generalization, the random-effects model seems to avoid subjective difficulties that plagued the fixed-effects model in generalizations to studies not identical to the study sample. That is, we do not need to ask, "How similar is similar enough?" Instead, we substitute another question: "Is this new study part of the universe from which the study sample was obtained?" If study samples were obtained from well-defined sampling frames via overtly specified sampling schemes, this might be an easy question to answer. This however is virtually never the case in meta-analysis (and is unusual in other applications of random-effects models). The universe is usually rather ambiguously specified and consequently the ambiguity in generalization based on random-effects models is that it is difficult to know precisely what the universe is. In contrast, the universe is clear in fixed-effects models, but the ambiguity arises in deciding whether a new study might be similar enough to the studies in the study sample.

The random-effects model does provide the technical method to address an important problem not handled in the fixed-effects model, namely, the additional uncertainty introduced by the inference to studies that are not identical (except for the sample of people involved) to those in the study sample. Inference to (nonsampled) studies in the fixed-effects model occurs outside the technical framework; hence any uncertainty it contributes cannot be evaluated by technical means within the model. In contrast, the random-effects model does incorporate between-study variation into the sampling uncertainty used to compute tests and estimates.

Although the random-effects model has the advantage of incorporating inferences to a universe of studies exhibiting variation in their characteristics, the definition of the universe may be ambiguous. A tautological universe definition could be derived by using the sample of studies to define "a universe from which the study sample is representative." Such a definition remains ambiguous; furthermore, it may not be the universe definition desired for the use of the information produced by the synthesis. For example, if the study sample includes many studies of short-duration, high-intensity treatments, but the likely practical applications usually involve low-intensity, long-duration treatments, the universe defined implicitly by the study sample may not be the universe most relevant to applications.

One potential solution to this problem might be to explicitly define a structured universe in terms of study characteristics, and to consider the study sample as a stratified sample from this universe. Estimates of parameters describing this universe could be obtained by weighting each "stratum" appropriately. For example, if half of the studies in the universe are long-duration studies, but only one-third of the study sample are, the results of each

long-duration study must be weighted twice as much as the short-duration studies.

3.2.1.3 Fixed Versus Random Effects The choice between fixed (conditional) or random (unconditional) modeling strategies arises in many settings in statistics and has caused lengthy debates because it involves subtleties of how to formulate questions in scientific research and what data are relevant to answering questions. For example, the debate between R. A. Fisher and Yates versus Pearson on whether to condition on the marginal frequencies in the analysis of 2×2 tables is about precisely this issue (see Camilli 1990), as is that on whether word *stimuli* should be treated as fixed or random effects in psycho-linguistics (see Clark 1973).

Those who advocated the fixed-effects position (for example, Peto 1987) argue that the basis for scientific inference should be only the studies actually conducted and observed in the study sample. Statistical methods should be employed only to determine the chance consequences of sampling of people into these (the observed) studies. Thus they would emphasize estimation and hypothesis testing for (or conditional on) this collection of studies. If we must generalize to other studies, they would argue that this is best done by subjective or extra-statistical procedures.

Those who advocate the random-effects perspective argue that the particular studies we observe are, to some extent, an accident of chance. The important inference question is not "what is true about these studies," but "what is true about studies like these that could have been done?" They would emphasize the generalization to other studies or other situations that could have been studied and that these generalizations should be handled by formal statistical methods. In many situations where research is used to inform public policy by providing information about the likely effects of treatments in situations that have not been explicitly studied, this argument seems persuasive.

3.2.1.4 Using Empirical Heterogeneity Although most statisticians would argue that the choice of analysis procedure should be driven by the inference model, some researchers choose based on the outcome of a statistical test of heterogeneity of effect sizes. This is called a *conditionally random-effects analysis* in a study of the properties of such tests for both conditional and unconditional inferences (Hedges and Vevea 1998). The authors find that the type I error rate of conditionally random-effects analyses were in between those of fixed- and random-effects tests: slightly inferior to fixed-effects tests for conditional inferences (but better than random-effects tests) and slightly inferior to random-effects tests for unconditional inferences (but better than fixed-effects tests). Whatever the technical performance of conditionally random-effects tests, their disadvantage is that they permit the user to avoid a clear choice of inference population, or worse to allow the data to (implicitly) make that determination depending on the outcome of a statistical test of heterogeneity.

3.2.1.5 Inference Populations The concept of a population about whose parameters we wish to draw inferences (the *inference population*) is explicit in the logic of the random-effects model. The definition of that inference population is not explicit, except that the studies observed could have been a random sample from that population. However, it is not always obvious that the desired target of inference would be the effect of the typical or average study that is observed. This is one of the reasons that subsets of studies (for example, those reflecting different subject population subgroups) are often examined in research syntheses. However, in some cases, populations have rather complex structure involving many attributes, and this is difficult to reflect in a simple subgroup structure. That is, the population is defined in terms of many variables.

The problem of generalization from samples in randomized trials has begun to be studied more explicitly as a problem of matching the study sample to the inference population (see, for example, O'Muircheartaigh and Hedges 2014; Tipton 2013). These methods are in the same spirit as cross design synthesis (see Droitcour, Silberman, and Chelimsky 1993). They require explicit specification of an inference population and model-based matching of study samples to inference populations. Exactly the same methods could be applied with individual patient data to improve generalizability of the results of research syntheses. In principle, these methods could be used to improve generalizability of conventional meta-analyses involving aggregate data, although the analytic details are more complex.

3.2.2 Nature of Parameters

Another fundamental issue in problem formulation concerns the nature of the effect-size parameter to be estimated. The issue can best be described in terms of population parameters (although each parameter has a corresponding sample estimate). Consider an actual study in which the effect-size parameter represents the true or population relationship between variables measured in the study. This effect-size parameter may be systematically

affected by artifactual sources of bias such as restriction of range or measurement error in the dependent variable. Correspondingly, imagine a hypothetical study in which the biases due to artifacts were controlled or eliminated. The effect-size parameter would differ from that of the actual study because the biases from artifacts of design would not be present. A key distinction is between a theoretical effect size (reflecting a relationship between variables in a hypothetical study) and an operational effect size (the parameter that describes the population relationship between variables in an actual study) (see chapter 15). Theoretical effect sizes are often conceived as those corrected for bias or for some aspect of experimental procedure (such as a restrictive sampling plan or use of an unreliable outcome measure) that can systematically influence effect size. Operational effect-size parameters, by contrast, are often conceived as affected by whatever bias or aspects of procedure that happen to be present in a particular study.

Perhaps the most prominent example of a theoretical effect size is the population correlation coefficient corrected for attenuation due to measurement error and restriction of range. One can also conceive of this as the population correlation between true scores in a population where neither variable is subject to restriction of range. The operational effect size is the correlation parameter between observed scores in the population in which variables have restricted ranges. Because the relation between the attenuated (operational) correlation and disattenuated (theoretical) correlation is known, it is possible to convert operational effect sizes into theoretical effect sizes.

Most research syntheses use operational effect size. Theoretical effect sizes are sometimes used, however, for one of two reasons. One is to enhance the comparability and hence combinability of estimates from studies whose operational effect sizes would otherwise be influenced quite substantially (and differently) by biases or incidental features of study design or procedure. This has sometimes been characterized as "putting all of the effect sizes on the same metric" (Glass, McGaw, and Smith 1981, 116). For example, in research on personal selection, virtually all studies involve restriction of range, which attenuates correlations (see Hunter and Schmidt 2004). Moreover, the amount of restriction of range typically varies substantially across studies. Hence correction for restriction of range ensures that each study provides an estimate of the same kind of correlation—the correlation in a population having an unrestricted distribution of test scores. Because restriction of range and many other consequences of design are incidental features of the studies, disattenuation to remove their effects is sometimes called *artifact correction*.

A more controversial reason for using theoretical effect sizes is that they are considered more scientifically relevant. For example, to estimate the benefit of scientific personnel selection using cognitive tests versus selection on an effectively random basis, we would need to compare the performance of applicants with the full range of test scores (those selected at random) with that of applicants selected via the test—applicants who would have a restricted range of test scores. Although a study of the validity of a selection test would compute the correlation between test score and job performance based on the restricted sample, the correlation that reflects the effectiveness of the test in predicting job performance is the one that would have been obtained with the full range of test scores—a theoretical correlation. Another example might be the estimation of the standardized mean difference of a treatment intended for a general population of people, but which has typically been investigated with studies using more restricted groups. Because the scores in the individual studies have a smaller standard deviation than the general population, the effect sizes will be artifactually large—that is if the treatment produced the same change in raw score units, dividing by the smaller standard deviations in the study sample would make the standardized difference look artifactually large. Hence a theoretical effect size might be chosen—the effect size that would have been obtained if the outcome scores in each study had the same variation as the general population. This would lead to corrections for sampling variability. Corrections of this sort are discussed in chapter 15.

3.2.3 Number and Source of Hypotheses

More than thirty years ago, Richard Light and David Pillemer distinguished between two types of questions that might be asked in a research synthesis (1984). One kind of question concerns a hypothesis that is specified precisely in advance (for example, on average does this treatment work?). The other type of question is specified only vaguely (for example, under what conditions does the treatment work best?). This distinction in problem specification is similar to that between planned and post hoc comparisons in the analysis of variance familiar to many researchers. Although either kind of question is legitimate in research synthesis, the calculation of levels of statistical significance may be affected by whether the question was defined in advance or discovered post hoc by examination of the data.

Although it is useful to distinguish the two cases, sharp distinctions are not possible in practice. Some of the literature is surely known to the reviewer before the synthesis, thus putatively a priori hypotheses are likely to have been influenced by the data. Conversely, hypotheses derived during exploration of the data may have been conceived earlier and proposed for explicit testing because the examination of the data suggested that they might be fruitful given the data. Perhaps the greatest ambiguity arises when a very large number of hypotheses are proposed a priori for testing. In this case, it is difficult to distinguish between hypotheses selected by searching the data informally, then proposing a hypothesis a posteriori and simply proposing all possible hypotheses a priori. For this reason, it may be sensible to treat large numbers of hypotheses as if they were post hoc. Despite the difficulty in drawing a sharp distinction, we still believe that the conceptual distinction between cases in which a few hypotheses are specified in advance and those in which there are many hypotheses (or hypotheses not necessarily specified in advance) is useful.

The primary reason for insisting on this distinction is statistical. When testing hypotheses specified in advance, it is appropriate to consider that test in isolation from other tests that might have been carried out. This is often called the use of a testwise error rate in the theory of multiple comparisons, meaning that the appropriate definition of the significance level of the test is the proportion of the time this test, considered in isolation, would yield a type I error.

In contrast, when testing a hypothesis derived after exploring the data, it may not be appropriate to consider the test in isolation from other tests that might have been done. For example, by choosing to test the most "promising" of a set of study characteristics (that is, the one that appears to be most strongly related to effect size), the reviewer has implicitly used information from tests that could have been done on other study characteristics. More formally, the sampling distribution of the largest relationship is not the same as that of one relationship selected a priori. In cases where the hypothesis is picked after exploring the data, special post hoc test procedures that take account of the other hypotheses that could have been tested are appropriate (see chapter 12; Hedges and Olkin 1985; or, more generally, Miller 1981). This is often called the use of an experimentwise error rate because the appropriate definition of the statistical significance level of the test is the proportion of the time the group of tests would lead to selecting a test that made a type I error.

Post hoc test procedures are frequently much less powerful than their a priori counterparts for detecting a particular relationship that is of interest. On the other hand, the post hoc procedures can do something that a priori procedures cannot: detect relationships that are not suspected in advance. The important point is the trade-off between the ability to find relationships that are not suspected and the sensitivity to detect those thought to be likely.

3.2.4 Analytic Models and Between- Versus Within-Study Relations

The distinction between evidence generated from contrasts between effects within studies (study-generated evidence) and that derived from contrasts between effects in different studies (review-generated evidence) has been appreciated for some time (see Cooper 1982). A generalization of this distinction has been less well appreciated, but can introduce ambiguities in meta-analyses with complex data structures involving multiple effects within studies.

Multiple effect sizes within studies (such as effect sizes associated with subgroups defined by subject characteristics, treatment intensities, follow-up intervals, and so on) signify a two-level data structure (effect sizes nested within studies). In such a two-level data structure, three relations are possible between effect sizes and any covariate: the relation within studies, the relation of study means on the variables of interest, and the total relation that ignores the nesting structure of effect sizes within studies (for a tutorial on this point, see Knapp 1977). The within-study and between-study relations are independent of one another (and there is no mathematical reason for them to be consistent with each other), but the total relation is a weighted linear combination of the other two.

The important point here is that the between-study relations are necessarily confounded with between-study differences on unmeasured variables, and thus these relations have weaker validity as estimates of causal associations. Meta-regression models can (and generally should) separate between-study and within-study models. However, doing so requires careful specification of the regression models and distinct interpretation of the coefficients reflecting each type of relation (see, for example, Tanner-Smith, Tipton, and Polanin 2016).

3.3 DATA COLLECTION

Data collection in research synthesis is largely a sampling activity that raises all of the concerns attendant on any other sampling activity. Specifically, the sampling

procedure must be designed to yield studies representative of the intended universe of studies. Ideally, the sampling is carried out in a way that reveals aspects of the sampling (such as dependence of units in the sample, see chapter 13), or selection effects (such as publication bias, see chapter 18) that might influence the analytic methods chosen to draw inferences from the sample.

3.3.1 Representativeness

Given that a universe has been chosen so that we know the kind of studies about which the synthesis is to inform us, one fundamental challenge is selecting the study sample in such a way that it supports inferences to that universe. One aspect of this effort is ensuring that search criteria are consistent with the universe definition. Assuming that they are, an exhaustive sample of studies that meet the criteria is often taken to be a representative sample of studies of the universe. However, sometimes viewing this proposition skeptically is in order.

The concept of representativeness has always been a somewhat ambiguous idea (see Kruskal and Mosteller 1979a, 1979b, 1979c, 1980) but the concept is useful in helping illuminate potential problems in drawing inferences from samples.

One reason is that some types of studies in the intended universe may not have been conducted. The act of defining a universe of studies that could be conducted does not even guarantee that it will be nonempty. We hope that the sample of studies will inform us about a universe of studies exhibiting variation in their characteristics, but studies with a full range of characteristics may not have been conducted. The types of studies that have been conducted therefore limit the possibility of generalizations, whatever the search procedures. For example, the universe of studies might be planned to include studies with both behavioral observations and other outcome measures. But if no studies have been conducted using behavioral observations, no possible sample of studies can, strictly speaking, support generalizations to a universe including studies with behavioral observations. The situation need not be as simple or obvious as suggested in this example. Often the limitations on the types of studies conducted arise at the level of the joint frequency of two or more characteristics, such as categories of treatment and outcome. In such situations, the limitations of the studies available are evident only in the scarcity of studies with certain joint characteristics, not in the marginal frequency of any type of study.

A second reason that exhaustiveness of sampling may not yield a representative sample of the universe is that, although studies may have been conducted, they may not have been reported in the forums accessible to the reviewer. This is the problem of missing data, as discussed in chapter 17. Selective reporting (at least in forums accessible to synthesists) can occur in many ways: entire studies may be missing or only certain results from those studies may be missing, and missingness may or may not be correlated with study results. The principal point here is that exhaustive sampling of data bases rendered nonrepresentative by publication or reporting bias does not yield a representative sample of the universe intended.

3.3.2 Dependence

Several types of dependence may arise in the sampling of effect-size estimates. The simplest is when several estimates are computed from measures from identical or partially overlapping groups of subjects. This can occur, for example, when the different outcomes are measured on the same subjects, or when several treatment groups are compared with the same control group to produce multiple effect sizes. This form of dependence most often arises when several effect-size estimates are computed from data reported in the same study, but it can also arise when several different studies report data on the same sample of subjects. Failure to recognize this form of dependence—and to use appropriate analytic strategies to cope with it—can result in inaccurate estimates of effects and their standard errors.

A second type of dependence occurs when studies with similar or identical characteristics exhibit less variability in their effect-size parameters than the entire sample of studies does. This might happen, for example, when a single laboratory or a single group of investigators use common procedures that might not be shared by other investigators and procedural variation is related to study effects. Because this kind of dependence leads to nested groups of studies with dependency, it is often called the *hierarchal dependence* model. Such an intra-class correlation of study effects leads to misspecification of random-effects models and hence to erroneous characterizations of between-study variation in effects. This form of dependence would also suggest misspecification in fixed-effects models if the study characteristics involved were not part of the formal explanatory model for between-study variation in effects.

We usually characterize the effect-size estimate T in a study as the sum of the effect-size parameter θ and the

estimation error ε, so that $T = \theta + \varepsilon$. The first type of dependence occurs through estimation error ε, the second type through the effect-size parameters θ. Methods for handling both types of dependence are discussed in chapter 13.

3.3.3 Study Selection and Publication Bias

A particularly pernicious form of missing data occurs when the probability that a result is reported depends on the result obtained (for example, on whether it is statistically significant). Such missing data would be called non-ignorable in the Little-Rubin framework and can lead to bias in the results of a meta-analysis. The resulting biases are typically called publication bias, although the more general reporting bias is a more accurate term. Methods for detecting and adjusting for publication bias are discussed at length in chapter 18.

3.4 DATA ANALYSIS

Much of this volume deals with issues of data analysis in research synthesis. It is appropriate here to discuss three issues of broad application to all types of statistical analyses in research synthesis.

3.4.1 Heterogeneity

Heterogeneity of effects in meta-analysis introduces a variety of interpretational problems that do not exist or are simpler if effects are homogeneous. For example, if effects are homogeneous, then fixed- and random-effects analyses essentially coincide and problems of generalization are simplified. Homogeneity also simplifies interpretation by making the synthesis basically an exercise in simple triangulation, the findings from different studies behaving as simple replications.

Unfortunately, heterogeneity is a rather frequent finding in research syntheses. This compels the synthesist to find ways of representing and interpreting the heterogeneity and dealing with it in statistical analyses. Fortunately, progress has been considerable in procedures for representing heterogeneity in interpretable ways, such as theoretically sound measures of the proportion of variance in observed effect sizes due to heterogeneity (the I^2 measure) (see chapter 12, this volume). Similarly, there has been great progress in both theory and software for random- and mixed-effects analyses that explicitly include heterogeneity in analyses (see chapter 12, this volume).

Some major issues arising in connection with heterogeneity are not easily resolvable. For example, heterogeneity is sometimes a function of the choice of effect-size index. If effect sizes of a set of studies can be expressed in more than one metric, they are sometimes more consistent when expressed in one metric than another. Which then (if either) is the appropriate way to characterize the heterogeneity of effects? This question has theoretical implications for statistical analyses (for example, how necessary is homogeneity for combinability, see Cochran 1954; Radhakrishna 1965). However, it is more than a theoretical issue because empirical evidence indicates that meta-analyses using some indexes find more heterogeneity than others; for example, meta-analyses using odds ratios are often more consistent across studies than risk differences (see Engels et al. 2000).

3.4.2 Unity of Statistical Methods

Much of the literature on statistical methodology for research synthesis is conceived as statistical methods for the analysis of a particular effect-size index. Thus, much of the literature on meta-analysis provides methods for combining estimated of odds ratios, or correlation coefficients, or standardized mean differences. Even though the methods for a particular effect-size index might be similar to those for another index, they are presented in the literature as essentially different methods. There is, however, a set of underlying statistical theory (see Cochran 1954; Hedges 1983) that provides a common theoretical justification for analyses of the effect-size measures in common use (for example, the standardized mean difference, the correlation coefficient, the log odds ratio, the difference in proportions, and so on).

Essentially all commonly used statistical methods for effect-size analyses rely on two facts. The first is that the effect-size estimate (or a suitable transformation) is normally distributed in "large samples" with a mean of approximately the effect-size parameter. The second is that the standard error of the effect-size estimate is a continuous function of the within-study sample sizes, the effect size, and possibly other parameters that can be estimated consistently from within-study data. Statistical methods for different effect-size indexes appear to differ primarily because the formulas for the effect-size indexes and their standard errors differ.

I am mindful of the variety of indexes of effect size that have been found useful. Chapter 11 is a detailed treatment of the variety of effect-size indexes that can be

applied to studies with continuous or categorical outcome variables. However, in this handbook we stress the conceptual unity of statistical methods for different indexes of effect size by describing most methods in terms of a "generic" effect size statistic T, its corresponding effect size parameter θ, and a generic variance v. This permits statistical methods to be applied to a collection of any type effect-size estimates by substituting the correct formulas for the individual estimates and their standard errors. This procedure provides not only a compact presentation of methods for existing indexes of effect size, but also the basis for generalization to new indexes that are yet to be used.

3.4.3 Large Sample Approximations

Virtually all of the statistical methods described in this volume and used in the analysis of effect sizes in research synthesis are based on what are called large sample approximations. This means that, unlike some simple statistical methods such as the t-test for the differences between means or the F-test in analyses of the general linear model, the sampling theory invoked to construct hypothesis tests or confidence intervals is not exactly true in very small samples. Large sample statistical theory is not limited to meta-analysis; in fact, it is used much more frequently in applied statistics than exact (or small sample) theory is, typically because the exact theory is too difficult to develop. For example, Pearson's chi-square test for simple interactions and log linear procedures for more complex analysis in contingency tables are large sample procedures, as are most multivariate test procedures, procedures using structural equation models, item response models, or even Fisher's z-transform of the correlation coefficient.

That large sample procedures are widely used does not imply that they are always without problems in any particular setting. Indeed, one of the major questions in any application of large sample theory is whether the "large sample" approximation is accurate enough in samples of the size available to justify its use. In meta-analysis, large sample theory is primarily used to obtain the sampling distribution of the sample effect-size estimates. The statistical properties of combined estimates or tests depends on the accuracy of the (approximations) to the sampling distributions of these individual effect-size estimates. Fortunately quite a bit is known about the accuracy of these approximations to the distributions of effect-size estimates.

In the cases of the standardized mean difference, the large sample theory is quite accurate for sample sizes as small as ten per group (see Hedges 1981, 1982; Hedges and Olkin 1985). In the cases of the correlation coefficient, the large sample theory is notoriously inaccurate in samples of less than a few hundred, particularly if the population correlation is large in magnitude. However, the large sample theory for the Fisher z-transformed correlation is typically quite accurate when the sample size is twenty or more. For this reason, I usually suggest that analyses involving correlation coefficients as the effect-size index be performed using the Fisher z-transforms of the correlations.

The situation with effect-size indexes for experiments with discrete outcomes is more difficult to characterize. The large sample theory for differences in proportions and for odds ratios usually seems to be reasonably accurate when sample sizes are moderate (for example, greater than fifty) as long as the proportions involved are not too near zero or one. If they are near zero or one, larger sample sizes may be needed to ensure comparable accuracy. In cases where all of the sample sizes are very small, real caution is required. In such cases "sparse sample" methods such as Mantel-Haenszel methods should be used.

A final technical point about the notion of large sample theory in meta-analysis concerns the dual meaning of the term. The total sample size N may be thought of as the sum of the sample sizes across k studies included in the synthesis. In studies comparing a treatment group with sample size n_i^E in the ith study and a control group with sample size n_i^C in the ith study, the total sample size is

$$N = \Sigma\left(n_i^E + n_i^C \right).$$

The formal statistical theory underlying most meta-analytic methods is based on large sample approximations that hold when N is large in such a way that *all* of $n_1^E, n_1^C, n_2^E, n_2^C, \ldots, n_k^E, n_k^C$ are also large.

Formally, the large sample theory underlying most meta-analysis describes the behavior of the limiting distribution as $N \rightarrow \infty$ in such a way that n^E/N and n^C/N are fixed as N increases. Much of this theory is not true when $N \rightarrow \infty$ by letting k increase and keeping the within-study sample sizes n^E and n^C small (see Neyman and Scott 1948). In most practical situations, this distinction is not important because the within-study sample sizes are large enough to support the assumption that all n^E and n^C are "large."

New and important exceptions to the usual asymptotic methods are the robust variance methods described in chapters 12 and 13. Here the asymptotic model is one in which $k \to \infty$ with no condition on the n^E or n^C. An important part of the development of this methodology is work on understanding and improving the small sample (small k) behavior of these methods (see Tipton 2015).

3.4.4 Synthetic Secondary Data Analysis

Improvements in data storage and computation are making it increasingly possible to conduct research syntheses using the complete corpus of primary data from every study in the synthesis. In medicine, this is called individual patient data meta-analysis or individual participant data (IPD) meta-analysis. Although such analyses are currently rare outside medicine, the widespread adoption of data-sharing policies by journals and funding agencies promises that it will be increasingly possible in other areas. Such analyses offer great promise in facilitating analyses involving within-study comparisons (using individual-level covariates).

IPD meta-analyses can be carried out using conventional multilevel statistical models and therefore are actually a form of synthetic secondary analysis. In such analyses, it is important to respect the multilevel structure of the data, which might be a two-level (participants within studies) or, in the case of more than one outcome variable per participant, a three-level (participants within outcomes within studies) structure. It is also important to carefully check model assumptions such as homoscedasticity in models for multilevel general linear models for continuous data and over- or underdispersion in multilevel generalized linear models for discrete outcomes.

Although IPD analyses offer great flexibility and the advantage of using conventional statistical methods, they offer little if any increase in statistical efficiency. Most meta-analytic estimation methods are asymptotically efficient (Hedges 1983). For example, in the case of meta-analyses using the standardized mean difference, improvements in small sample efficiency must occur through savings in degrees of freedom. A degree of freedom is lost in each study to standardize the effect size, but a degree of freedom must be used in the IPD analysis to account for study fixed effects, hence the modeling uses the same number of degrees of freedom obviating the advantage of the IPD model.

Bayesian methods do not require IPD, but those that do use IPD can provide the important advantage of having known small sample properties as well as the interpretational advantages of the Bayesian framework (see chapter 14).

3.5 CONCLUSION

Statistical thinking is important in every stage of research synthesis, as it is in primary research. Statistical issues are part of the problem definition, play a key role in data collection, and are obviously important in data analysis. The careful consideration of statistical issues throughout a synthesis can help ensure its validity.

3.6 REFERENCES

Camilli, Greg. 1990. "The Test of Homogeneity for 2 × 2 Contingency Tables: A Review and Some Personal Opinions on the Controversy." *Psychological Bulletin* 108(1): 135–45.

Clark, Herbert H. 1973. "The Language-as-Fixed-Effect Fallacy: A Critique of Language Statistics in Psychological Research." *Journal of Verbal and Learning Behavior* 12(4): 335–59.

Cochran, William G. 1954. "The Combination of Estimates from Different Experiments." *Biometrics* 10(1): 101–29.

Cooper, Harris M. 1982. "Scientific Guidelines for Conducting Integrative Research Reviews." *Review of Educational Research* 52(2): 291–302.

Droitcour, Judith, George Silberman, and Eleanor Chelinsky. 1993. "Cross-Design Synthesis: A New Form of Meta-Analysis for Combining Results from Randomized Clinical Trials and Medical-Practice Databases." *International Journal of Technology Assessment in Health Care* 9(3): 440–49.

Engels, Eric A., Christopher C. Schmid, Norma Terrin, Ingram Olkin, and Joseph Lau. 2000. "Heterogeneity and Statistical Significance in Meta-Analysis: An Empirical Study of 125 Meta-Analyses." *Statistics in Medicine* 19(13): 1707–28.

Glass, Gene V., Barry McGaw, and Mary L. Smith. 1981. *Meta-Analysis in Social Research.* Beverly Hills, Calif.: Sage Publications.

Hedges, Larry V. 1981. "Distribution Theory for Glass's Estimator of Effect Size and Related Estimators." *Journal of Educational Statistics* 6(2): 107–28.

———. 1982. "Estimating Effect Size from a Series of Independent Experiments." *Psychological Bulletin* 9(2)2: 490–99.

———. 1983. "Combining Independent Estimators in Research Synthesis." *British Journal of Mathematical and Statistical Psychology* 36(1): 123–31.

Hedges, Larry V., and Ingram Olkin. 1985. *Statistical Methods for Meta-Analysis*. New York: Academic Press.

Hedges, Larry V., and Jack L. Vevea. 1998. "Fixed and Random Effects Models in Meta-Analysis." *Psychological Methods* 3(4): 486–504.

Hunter, John E., and Frank L. Schmidt. 2004. *Methods of Meta-Analysis*. Thousand Oaks, Calif.: Sage Publications.

Knapp, Thomas R. 1977. "The Unit-of-Analysis Problem in Applications of Simple Correlation Analysis to Educational Research." *Journal of Educational Statistics* 2(3): 171–86.

Kruskal, William, and Fred Mosteller. 1979a. "Representative Sampling I: Non-Scientific Literature." *International Statistical Review* 47(1): 13–24.

———. 1979b. "Representative Sampling, II: Scientific Literature, Excluding Statistics." *International Statistical Review* 47(2): 111–27.

———. 1979c. "Representative Sampling, III: The Current Statistical Literature." *International Statistical Review* 47(3): 245–65.

———. 1980. "Representative Sampling IV: The History of the Concept in Statistics 1895–1939." *International Statistical Review* 48(2): 169–95.

Light, Richard J., and David Pillemar. 1984. *Summing Up.* Cambridge, Mass.: Harvard University Press.

Miller, Rupert D. 1981. *Simultaneous Statistical Inferences*. New York: Springer Verlag.

Neymann, Jerzy, and Elizabeth L. Scott. 1948. "Consistent Estimates Based on Partially Consistent Observations." *Econometrika* 16(1): 1–32.

O'Muircheartaigh, Colm, and Larry V. Hedges. 2014. "Generalizing from Experiments with Non-Representative Samples." *Journal of the Royal Statistical Society, Series C* 63(2):195–210.

Peto, Richard 1987. "Why Do We Need Systematic Overviews of Randomized Traits (with Discussion)." *Statistics in Medicine* 6(3): 233–44.

Radhakrishna, S. 1965. "Combining the Results from Several 2×2 Contingency Tables." *Biometrics* 21(1): 86–98.

Tanner-Smith, Emily, Elizabeth Tipton, and Josh Polanin. 2016. "Handling Complex Meta-Analytic Data Structures Using Robust Variance Estimates: A Tutorial in R." *Journal of Developmental and Life-Course Criminology* 2(1): 85–112.

Tipton, Elizabeth. 2013. "Improving Generalization from Experiments Using Propensity Score Subclassification: Assumptions, Properties, and Contexts." *Journal of Educational and Behavioral Statistics* 38(3): 239–66.

———. 2015. "Small Sample Adjustments for Robust Variance Estimation in Meta-Analysis." *Psychological Methods* 20(3): 375–93.

PART III

SEARCHING THE LITERATURE

4

SCIENTIFIC COMMUNICATION AND LITERATURE RETRIEVAL

HOWARD D. WHITE
Drexel University

CONTENTS

4.1 COMMUNING WITH THE LITERATURE

A defining mission of science is the communication of warranted findings. This communication has four broad modes: informal oral, informal written, formal oral, and formal written. The first two include face-to-face or telephone conversations and emails or tweets among colleagues. They are undoubtedly important, particularly as ways of sharing news and of receiving preliminary feedback on professional work. Only through the formal modes, however, can scientists achieve their true goal—priority in making sound new knowledge claims in a cumulative enterprise. The cost to them is the effort needed to prepare the claims for public delivery, especially to critical peers.

Of all the modes, formal written communication is the most premeditated, even more so than a formal oral presentation, such as a briefing or lecture. Its typical products—papers, articles, monographs, reports, in printed or digital form—constitute the primary literatures by which scientists lay open their work to permanent public scrutiny in hopes that their claims to new knowledge will be validated and esteemed.

Reading to keep current in one's field is itself an act of communication. It is communing with the literature—with durably stored, interrelated, claims-bearing texts. However, the mass of studies on many topics is daunting, and research reviews are a partial remedy. They emerge at a late stage in formal communication, after scores or even hundreds of primary writings have appeared, and they serve both to pack literatures into brief compass and to warrant or undermine claims. As a result, they are likely to find readers not only among researchers, but among policymakers, practitioners, and members of the general public.

Among the literature reviewer's tasks, three stand out: discovering and retrieving primary works, critically evaluating them in light of theory, and distilling their essentials in syntheses that conserve the reader's time. All three tasks require the judgment of one trained in the research tradition (the theory and methodology) under study. As Patrick Wilson puts it, "The surveyor must be, or be prepared to become, a specialist in the subject matter being surveyed" (1977, 17). But the payoff, he argues, can be rich:

> The striking thing about the process of evaluation of a body of work is that, while the intent is not to increase knowledge by the conducting of independent inquiries, the result may be the increase of knowledge, by the drawing of conclusions not made in the literature reviewed but supported by the part of it judged valid. The process of analysis and synthesis can produce new knowledge in just this sense, that the attempt to put down what can be said to be known, on the basis of a given collection of documents, may result in the establishment of things not claimed explicitly in any of the documents surveyed. (11)

4.2 THE REVIEWER'S PROGRESS

In the first edition of this book, this section was called "The Reviewer's Burden" to highlight the oft-perceived difficulty of getting scientists to write literature reviews as part of their communicative duties (White 1994). The typical review involved tasks that many scientists found irksome—an ambitious literature search, obtaining of documents, extensive reading, reconciliation of conflicting claims, preparation of citations—and the bulk of effort was centered on the work of others, as if one had to write chapter 2 of one's dissertation all over again. The Committee on Scientific and Technical Communication quoted one reviewer as saying, "Digesting the material so that it could be presented on some conceptual basis was plain torture; I spent over 200 hours on that job. I wonder if 200 people spent even one hour reading it" (1969, 181). A decade later, Charles Bernier and Neil Yerkey observed, "Not enough reviews are written, because of the time required to write them and because of the trauma sometimes experienced during the writing of an excellent critical review" (1979, 48–49). A decade after that, Eugene Garfield identified a further impediment in the view among some scientists "that a review article—even one that is highly cited—is a lesser achievement than a piece of original research" (1989, 113).

Against that backdrop, the research synthesis movement was an intriguing development. It began circa 1975 with initiatives for meta-analyses of studies (Cooper 2000; Glass 2000). Far from avoiding the literature search, meta-analysts were advised to look for all empirical studies on a subject—even the unpublished ones—so as to evaluate the full range of reported statistical effects (Rosenthal 1984; Green and Hall 1984). This required a creative attack on the problem of overload. What had been merely too many things to read became a population of studies that could be treated like the respondents in survey research. Just as pollsters put questions to people, reviewers could "interview" existing studies and systematically record their attributes. The process could become a team effort, with specialization of labor prior to writing: for example, a librarian to retrieve abstracts, a project director to choose studies for review and to create the

coding sheet, graduate students to code the attributes, and a statistician to run computer analyses and prepare tables.

By the time of the second edition of this book, the research synthesis movement had become a strong counterforce to the older, "narrative" style of reviewing (White 2009; see also Hunt 1997; Petticrew 2001). Research synthesists sought to make reviews more rigorous, with at least the level of statistical sophistication found in primary research. Narrative reviewers had been able to get by on their methodological skills without extra reading; the synthesists wrote entire textbooks on meta-analysis (Wang and Bushman 1999; Lipsey and Wilson 2001; Hunter and Schmidt 2004; Littell, Corcoran, and Pillai 2008). Narrative reviewers had been mute on how they found the studies under review; the synthesists called for explicit mention of sources and search strategies. Narrative reviewers had accepted or rejected studies impressionistically; the synthesists wanted firm editorial criteria. Narrative reviewers were inconsistent in deciding which aspects of studies to discuss; the synthesists required consistent coding of attributes across studies. Narrative reviewers used ad hoc judgments as to the meaning of statistical findings; the synthesists insisted on formal comparative and summary techniques. These standards were being extended even to qualitative studies and to nonstatistical integration of findings (Bland, Meurer, and Maldonado 1995; Petticrew 2001; Hawker et al. 2002; Jones 2004; Dixon-Woods et al. 2004; Thomas et al. 2004).

The period also saw spectacular technical improvements in communication (Brown 2010). The internet advanced interpersonal messaging, expanded the resources for publication, created limitless library space, and brought swift retrieval to vast stores of documents. The full reaches of the Web and its specialized databases were accessible day or night through search engines, notably Google. Retrievals were routinely based on specific words and phrases in documents and not merely on global descriptions such as titles and subject headings. In many cases, one could pass quickly from bibliographic entries to full texts, closing a centuries-old gap.

The new technologies for literature retrieval made it much easier for fact-hungry professionals to retrieve something they especially valued: documentary evidence for claims, preferably statistical evidence from randomized controlled experimental trials (Mulrow and Cook 1998; Petitti 2000; Egger, Davey-Smith, and Altman 2008). Evidence-based practice, although not without critics (Pawson 2006), had become another movement, especially in medicine and health care but also in psychology, education, social work, and public policy studies. The various evidence-based communities were an ideal audience for research as represented by this book—that is, compactly displayed tests of evidence from multiple experiments (Cooper 2000, see appendix).

In this edition, the research synthesis movement is fully mainstream. The Cochrane Collaboration and the Campbell Collaboration, to be discussed shortly, have flourished as its institutional bases since the 1990s. The commercial website Comprehensive Meta-analysis offers software and other resources (https://www.meta-analysis.com). Specialized journals such as *Research Synthesis Methods* (2010–), *Systematic Reviews* (2012–), and *World Journal of Meta-Analysis* (2013–) have emerged. New or revised textbooks still appear (Borenstein et al. 2009; Bronson and Davis 2012; Saini and Shlonsky 2012; Fink 2014; Cheung 2015; Cooper 2017; Booth, Sutton, and Papaioannou 2016). The movement's literature continues its rapid growth, as seen in graphics to come. Reviews in some topical areas now require integrative reviews themselves (Whitlock et al. 2008; Smith et al. 2011). A paper by Hilda Bastian, Paul Glasziou, and Iain Chalmers is titled "Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up?" (2010). But conditions of overload are likely to persist, because techniques for producing reviews are becoming faster (Ananiadou et al. 2009; Ganann, Ciliska, and Thomas 2010; Khangura et al. 2012).

4.3 THE COCHRANE COLLABORATION

Although research synthesis and evidence-based practice exemplify symbiotic scientific communication in the internet era, both movements were gathering strength before the internet took off. A striking organizational improvement antedates many of the internet's technological advances: the international Cochrane Collaboration, founded in 1993. Cochrane members typically believe that finding and exploiting studies with randomized controlled trials (RCTs) is too important to be left to the older, haphazard methods. To produce systematic reviews—another name for meta-analytic integrative reviews or research syntheses—these members have built a socio-technical infrastructure for assembling relevant materials. Criteria for selecting studies to review, for statistical analysis, and for reporting results are now explicit. Teams of specialists known as review groups are named to monitor specific empirical literatures and to contribute their choices to several databases, called the Cochrane Library, that function as

large-scale instruments of quality control. The library includes databases of completed systematic reviews, of methodology reviews, of abstracts of reviews of effects, and of controlled trials available for meta-analysis. Library staff also publish an annual titled *Cochrane Methods* (2010–).

The *Cochrane Handbook for Systematic Reviews of Interventions*, viewable online and downloadable by subscribers, sets procedures for retrieving and analyzing items from the databases (Higgins and Green 2011). According to Andrea Furlan and her colleagues, for instance, retrievals for the Cochrane Back and Neck (CBN) Group should involve

> A search of the Cochrane Central Register of Controlled Trials (CENTRAL) that is included in the most recent issue of the Cochrane Library. CENTRAL includes some, but not all trials from MEDLINE and EMBASE as well as the CBN Group Specialized Trials Register.
>
> A computer-aided search of MEDLINE (for example, via PubMed) database since its inception for new reviews and since the date of the previous search for updates of reviews.
>
> Screening references listed in relevant systematic reviews and identified trials.
>
> Identification of unpublished and ongoing trials: WHO International Clinical Trials Registry Platform (http://www.who.int/ictrp/en) and the U.S. National Institutes of Health (https://clinicaltrials.gov). Including a search for unpublished trials is useful to assess the presence and magnitude of publication bias. (2015, 1662)

Recommendations such as these are complemented by *Systematic Reviews*, a Cochrane-linked monograph from the University of York's Centre for Reviews and Dissemination (2009). It, too, is strict about conducting and documenting literature searches properly (see 21–22, 249–52).

In the bad old days, Gregg Jackson surveyed thirty-six integrative reviews and found that only one stated the indexes (such as *Psychological Abstracts*) used in the search, and only three mentioned drawing on the bibliographies of previous review articles (1980). He remarked, "The failure of almost all integrative review articles to give information indicating the thoroughness of the search for appropriate primary sources does suggest that neither the reviewers nor their editors attach a great deal of importance to such thoroughness" (444). Today, by contrast, the Cochrane influence can be seen even in the abstracts of reviews, many of which describe the literature search underlying them, as in this protocol-based section of an abstract by Rob Smeets and his colleagues:

> *Method*. Systematic literature search in PUBMED, MEDLINE, EMBASE and PsycINFO until December 2004 to identify observational studies regarding deconditioning signs and high quality RCTs regarding the effectiveness of cardiovascular and/or muscle strengthening exercises. Internal validity of the RCTs was assessed by using a checklist of nine methodology criteria in accordance with the Cochrane Collaboration. (2006, 673)

This is not to say that post-Cochrane reviewers always characterize their work adequately. For instance, Vivien Bramwell and Christopher Williams report that in more than a hundred qualitative reviews in the *Journal of Clinical Oncology*, "Authors rarely gave information on methods of data identification (11.3 percent), data selection (10.4 percent) and assessment of validity (8.4 percent)" (1997, 1185). Donna Stroup and her colleagues propose better specification of details in meta-analytic observational studies (2000). David Moher and his colleagues criticize three hundred systematic reviews from MEDLINE, writing, "SRs are now produced in large numbers, and our data suggest that the quality of their reporting is inconsistent. This situation might be improved if more widely agreed upon evidence-based reporting guidelines were endorsed and adhered to by authors and journals" (2007, 447). As a follow-up, Moher and colleagues introduced PRISMA, an initialism for preferred reporting items for systematic reviews and meta-analyses (2009). Published in several medical journals and online, PRISMA is a checklist with commentary and examples of the deficiencies it seeks to correct.

The Cochrane approach upgrades reviewing in general. The stakeholders in medical and health-care research, not least the funders, are likely to want Cochrane standards imposed wherever possible—Gresham's law in reverse. Reviewers for their part seem willing to accept greater discipline, as indicated by their exploding literature. Figure 4.1 shows publications during 2006 through 2015 with "meta-analy*," "integrative review," or "systematic review" in their descriptions in the Science Citation Index Expanded of the Web of Science. As of April 2016, they exceed 131,500, with most coming from journals in medicine or allied fields. This count actually understates the literature's true size, because it excludes items not matching the search terms, as well as books and other publications not covered by SCI Expanded.

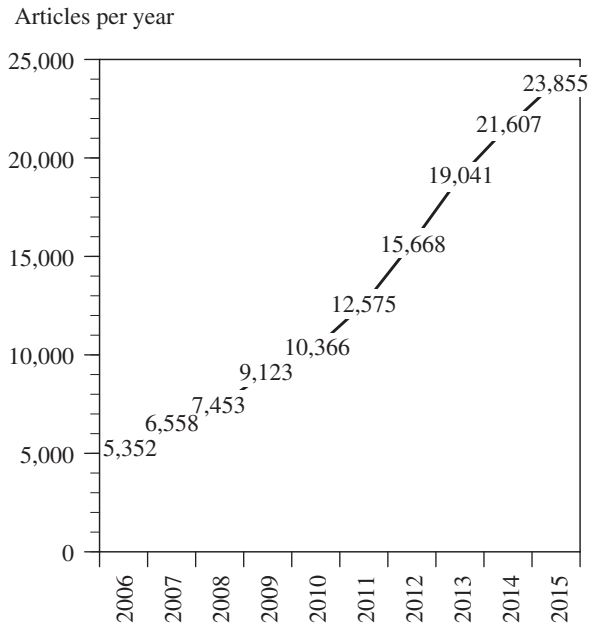

Figure 4.1 Growth of Research Synthesis Literatures in Science Citation Index Expanded, 2006–2015

SOURCE: Author's compilation.

4.4 THE CAMPBELL COLLABORATION

The goal of medical fields is beneficial intervention in crucial situations. Because crucial in this case can mean life or death, it is obvious why professionals in these fields would want to base their decisions on the best available evidence and why the Cochrane Collaboration emerged in response. But professionals outside medicine and health care also seek to intervene beneficially in people's lives, and their interventions may also have far-reaching social and behavioral consequences. Thus, drawing on the Cochrane model, the international Campbell Collaboration (C2) was formed in 1999 to facilitate systematic reviews of intervention effects in such areas as education, delinquency and criminal justice, mental health, welfare, housing, and employment (Cooper 2000). For example, the influence of C2 is seen in at least one study that relates systematic review techniques to education (Davies 2000), and in articles that relate them to social policy and practice in general (see Boaz, Ashby, and Young 2002; Petticrew and Roberts 2006; Shlonsky et al. 2011).

Eamonn Noonan and Arild Bjørndal note three ways in which C2 reviews differ from their Cochrane counterparts: research in the social sciences is less integrated with practice than in clinical medicine; the relative scarcity of randomized trials means that C2 reviews include more nonrandomized studies; and interventions are complex (2011). Mark Petticrew clarifies these points: "If researchers see 'simplicity' in an intervention, then they may be more likely to argue that randomized controlled trials (as opposed to other sorts of research) are feasible and appropriate. However if they see complexity (nonlinear pathways, multiple synergistic components, feedback loops and so on) as the key features, then by implication other types of research may be necessary for illuminating those complex processes" (2011, 397). The weighing of more disparate kinds of evidence in C2 reviews is thus a particular concern.

Like Cochrane, C2 is strong on tutorial materials—for example, this re-paragraphed description from the Campbell Collaboration website:

> The purpose of a systematic review is to sum up the best available research on a specific question. This is done by synthesizing the results of several studies. A systematic review uses transparent procedures to find, evaluate and synthesize the results of relevant research. Procedures are explicitly defined in advance, in order to ensure that the exercise is transparent and can be replicated. This practice is also designed to minimize bias. Studies included in a review are screened for quality, so that the findings of a large number of studies can be combined. Peer review is a key part of the process; qualified independent researchers control the author's methods and results. (2016)

Also like Cochrane, the Campbell Collaboration offers infrastructural databases: the C2 Social, Psychological, Education, and Criminological Trials Registry, and the C2 Reviews of Interventions and Policy Evaluations. Within these broad areas, it has organized coordinating groups to supervise review preparation. Accordingly, the Steering Group of the Campbell Collaboration calls for crisply standardized reporting, as suggested by their outline of what the methods section of a C2 protocol should contain (2015, 22):

- Characteristics of the studies relevant to the objectives of the review
- Criteria for inclusion and exclusion of studies in the review
- Search strategy for finding eligible studies

Data extraction and study coding procedures

Risk of bias

Synthesis procedures and statistical analysis

Treatment of qualitative research

Persons who perform the actual retrievals for C2 reviews can consult a monograph by Shannon Kugley and her colleagues that sets out the whole search process and its documentation in detail (2016).

Prescriptiveness along these lines has coincided with steady increases in literatures relevant to the C2 world. Figure 4.2 plots the growth of the almost thirty-six thousand articles and reviews covered by the Social Sciences Citation Index (SSCI) between 2006 and 2015. These items were retrieved with the same search strategy as those in figure 4.1. A fair number are also counted in figure 4.1 because, if an item has both medical aspects and social-behavioral aspects, SCI Expanded and SSCI both cover it. Table 4.1 ranks the fifty journals that yield the most items in SSCI. Journals in psychology, psychiatry, and medicine predominate. Numerous social science journals appear in ranks not shown.

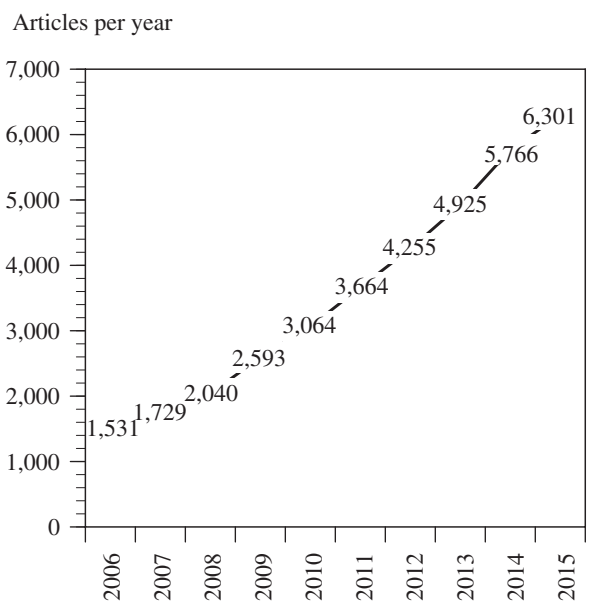

Figure 4.2 Growth of Research Synthesis Literatures in SSCI, 2006–2015

SOURCE: Author's compilation.

Although the problems of literature retrieval are never entirely solved, the Cochrane and Campbell materials make them much likelier to be addressed by research synthesists today. The idea of carefully documenting retrievals is not new, of course. Decades ago, for example, Patrick Wilson (1968) and Marcia Bates (1976) argued that the compiler of a rigorous bibliography should state its domain, scope, and selection principles:

- Domain—all the sources searched, including sources that yielded nothing although they initially seemed plausible.
- Scope—the subject headings or other topical language used to search the various sources; also the terms that imposed geographic, temporal, organizational, and language constraints on retrievals.
- Selection principles—editorial criteria used to include or exclude items on the basis of their content.

But these guidelines were published primarily for librarians, scholars, and students, who cannot prescribe bibliographic practices to scientists. Now, similar criteria have been adopted by scientists who can prescribe them to their peers and who also have an influential constituency, the evidence-based practitioners, to back them up. Since scientists communicate not only results but norms, this new consensus on bibliographic standards is a significant advance.

4.5 RECALL AND PRECISION

The field centrally concerned with retrieving documents from literatures is library and information science (LIS). Research synthesis, evidence-based practice, and LIS thus fit naturally together. The convergence is still incomplete, however, as Laura Sheble's extensive historical and bibliometric analyses attest (2014, 2016).

In the vocabulary of LIS, synthesists are unusually interested in *high recall* of documents (Beahler, Sundheim, and Trapp 2000; Conn et al. 2003; Schlosser et al. 2005). Recall is a measure used by information scientists to evaluate literature searches: it expresses (as a percentage) the ratio of relevant documents retrieved to all those in a collection that should be retrieved. For large collections the latter value is, in truth, unavailable: if we could identify all relevant documents in order to count them, we could retrieve them all, and so recall would always be 100 percent. The denominator in the recall ratio is therefore almost always an estimate. Nevertheless, recall is a useful fiction in analyzing possible demands on retrieval

Table 4.1 Top 50 SSCI Journals in Articles Relevant to Research Synthesis, 2006–2015

1,171	Value in Health
528	Cochrane Database of Systematic Reviews
516	PLoS One
319	Clinical Psychology Review
315	Journal of Affective Disorders
269	BMC Public Health
252	Schizophrenia Research
246	Psychological Bulletin
246	Journal of Advanced Nursing
238	Journal of Clinical Nursing
207	Journal of the American Geriatrics Society
196	Psychological Medicine
191	Schizophrenia Bulletin
185	International Journal of Nursing Studies
182	Psycho Oncology
179	European Psychiatry
177	BMC Health Services Research
176	Journal of Epidemiology and Community Health
166	British Journal of Psychiatry
164	Implementation Science
162	Patient Education and Counseling
161	Journal of Clinical Psychiatry
161	BMJ Open
160	Addiction
156	Obesity Reviews
156	Disability and Rehabilitation
143	American Journal of Preventive Medicine
139	Social Science Medicine
136	Annals of Behavioral Medicine
135	Journal of Applied Psychology
131	Psychology Health
129	International Psychogeriatrics
129	Acta Psychiatrica Scandinavica
123	Journal of Psychosomatic Research
120	Quality of Life Research
119	Frontiers in Psychology
116	Pharmacoeconomics
113	Aggression and Violent Behavior
110	Health Technology Assessment
103	Preventive Medicine
103	Australian and New Zealand Journal of Psychiatry
102	Psychiatry Research
98	Journal of Clinical Epidemiology
98	International Journal of Geriatric Psychiatry
97	Nursing Research
97	European Journal of Public Health
96	Pediatrics
96	Ciencia Saude Coletiva
94	American Journal of Public Health
93	Gerontologist

SOURCE: Author's compilation.

systems. Another major desideratum in retrieval systems is *high precision,* where precision expresses (as a percentage) the ratio of documents retrieved and judged relevant to all those actually retrieved. Precision measures how many irrelevant documents—false positives—one must examine to find the true positives, or hits.

Precision and recall tend to vary inversely. If one seeks high recall—complete or comprehensive retrievals—one must examine the many irrelevant documents that present technology also disgorges, which degrades precision. Alternatively, retrievals can be made highly precise so as to cut down on false positives, but at the cost of missing the relevant documents (false negatives) that the search terms fail to capture, which degrades recall.

Most literature searchers actually want high-precision retrievals, preferring relatively little bibliographic output to scan and relatively few items to read at the end of the judgment process. This is one manifestation of "least effort," an economizing behavior often seen in information seekers (Mann, 1993, 91–101). The research synthesists are distinctive in wanting—or at least accepting the need for—high recall. As Jackson puts it, "Since there is no way of ascertaining whether the set of located studies is representative of the full set of *existing* studies on the topic, the best protection against an unrepresentative set is to locate as many of the existing studies as is possible" (1978, 14). Bert Green and Judith Hall call such attempts essential, although mundane and often tedious (1984). Some meta-analysts doubt that comprehensive searches are worth the effort, but even they seem to have higher standards for uncovering studies than librarians and information specialists typically encounter (Laird 1990). The only other group likely to be as driven by a need for exhaustiveness are doctoral students in the early stages of writing their dissertations.

The point of high recall is not to track down every paper that is somehow related to the topic. Research synthesists who reject this idea are quite sensible. The point is to avoid missing a consequential paper that lies outside the synthesist's regular purview—in other words, to ensure

that habitual channels of communication will not bias the retrieval. In professional matters, most researchers find it hard to believe that subscriptions to certain journals and conversations with certain colleagues can fail to keep them well informed. But the history of science and scholarship is full of examples of mutually relevant specialties that were unaware of each other for years because their members construed their topics too narrowly and failed to ask what other researchers—perhaps with different technical vocabularies (Grupp and Heider 1975)—were doing in a similar vein.

More to the point, researchers are likely to miss writings in their own specialties. The experience of one research team shows that even known sources immediately at hand may produce an unexpectedly large yield (Greenhouse, Wachter, and Straf 1990). The further lesson is that sources worth checking may still be unknown. Research synthesists primarily concerned with substantive and methodological issues should get professional advice in literature retrieval, just as they would in statistics or computing if a problem exceeded their expertise (Petticrew 2001, 100–101). Retrieval specialists have produced many publications of their own in the research synthesis context (for example, Doig and Simpson 2003; Wong, Wilczynski, and Haynes 2006; Jesson, Matheson, and Lacey 2011; Bayliss, Davenport, and Pennant 2014).

4.6 IMPROVING THE YIELD

The most obvious way to improve recall is for researchers to avail themselves of a variety of reference databases (also called bibliographic databases or abstracting and indexing services). Harris Cooper describes fifteen ways in which fifty-seven authors of empirical research reviews actually conducted literature searches (1985, 1987). Most authors preferred to trace the references in review papers, books, and nonreview papers already in hand (presumably many were from their own files) and to ask colleagues for recommendations. These are classic least-effort ways of minimizing work and maximizing payoff in retrievals. For those purposes, they are unexceptionable, which is why academics and other intelligent people use them so often. Nevertheless, they are almost guaranteed to miss relevant writings, particularly if literatures are large and heterogeneous. Moreover, because they are unsystematic, they cannot easily be replicated.

How many modes of searching are there? In an account from the LIS literature, Patrick Wilson categorizes five major modes, into which Cooper's more specific fifteen fit nicely (1992). Thomas Mann, a Library of Congress reference librarian, names eight modes, but they are simply variants on Cooper's fifteen, and they, too, fit well in Wilson's categories (1993). The convergence of these three independent accounts illustrates Wilson's claim, quoted earlier, that writings in combination may yield knowledge not found in any of them separately.

To show the convergence, table 4.2 places Cooper's and Mann's modes under Wilson's headings, all verbatim. Discussing computer searching, Mann combines citation databases with other kinds; here, they are put with his example of *citation searches in printed sources*; otherwise, his text is unaltered. As high-level generalizations, these modes do not date, even when the technologies within them have changed, and jointly they pretty well exhaust possible ways of searching. All of them ought to be considered by someone striving for high recall of documents.

The different modes can be used concurrently to save time. Research synthesists disinclined to go beyond their favorite few can extend their range by delegating searches to information specialists, including librarians, and by tasking team members to search independently. This delegation would particularly hold for less favored strategies, such as searching reference databases, browsing library collections, and discovering topical bibliographies. Nowadays, the Cochrane and Campbell databases would be important places to look.

The other obvious way to improve recall is to do forward searches in the citation indexes of the Web of Science (published by Clarivate Analytics), Scopus (published by Elsevier), and Google Scholar. The footnote chasing entries in table 4.2 reflect citation searches that go backward in time, that is, from a known publication to the earlier items it cites. The contrasting kind of citation search moves forward in time from a known publication to the later items that cite it. Although citation indexes will not infallibly yield new hits, it is fatuous to ignore them—roughly equivalent to not keeping abreast of appropriate statistical techniques.

The result of accepting higher retrieval standards is that, rather than veiling one's strategies, one can state them candidly as documentation in the final report (Rothstein et al. 2004). Whether such statements appear in the main text, endnotes, or appendices, they help later searchers replicate the original searcher's results and avoid unnecessary duplication of effort (Atkinson et al. 2015).

Because the thoroughness of the search depends on knowledge of the major modes for retrieving studies,

Table 4.2 Five Major Modes of Searching

Footnote Chasing
- *Cooper 1985*
  - References in review papers written by others
  - References in books by others
  - References in nonreview papers from journals you subscribe to
  - References in nonreview papers you browsed through at the library
  - Topical bibliographies compiled by others
- *Mann 1993*
  - Searches through published bibliographies (including sets of footnotes in relevant subject documents)
  - Related records searches

Consultation
- *Cooper 1985*
  - Communication with people who typically share information with you
  - Informal conversations at conferences or with students
  - Formal requests of scholars you knew were active in the field (for example, solicitation letters)
  - Comments from readers or reviewers of past work
  - General requests to government agencies
- *Mann 1993*
  - Searches through people sources (whether by verbal contact, email, electronic bulletin board, letters, and so on)

Searches in Subject Indexes
- *Cooper 1985*
  - Computer search of abstract data bases (for example, *ERIC, Psychological Abstracts)*
  - Manual search of abstract data bases
- *Mann 1993*
  - Controlled-vocabulary searches in manual or printed sources
  - Keyword searches in manual or printed sources
  - Computer searches—which can be done by subject heading, classification number, keyword . . .

Browsing
- *Cooper 1985*
  - Browsing through library shelves
- *Mann 1993*
  - Systematic browsing

Citation Searches
- *Cooper 1985*
  - Manual search of a citation index
  - Computer search of a citation index (for example, *SSCI)*
- *Mann 1993*
  - Citation searches in printed sources
  - Computer searches by citation

SOURCE: Adapted from Cooper 1985, Wilson 1992, and Mann 1993. The boldface headings are from Wilson.

a more detailed look at each of Wilson's five categories follows (1992). Cooper's and Mann's examples appear in italics as they are woven into the discussion. The first two modes, footnote chasing and consultation, are understandably attractive to most scholars, but may be affected by personal biases more than the other three, which involve searching impersonal databases or collections. Hence the discussion is somewhat weighted in favor of the latter. It is best to assume that all five are needed, although different topics may require them in different proportions.

4.7 FOOTNOTE CHASING

This is the adroit use of other authors' footnotes, or, more broadly, their references to the prior literature on a topic. Because *footnotes* may seem too much a humanities term, *references* will replace it in what follows. The reason research synthesists like to chase references is that doing so may immediately lead to usable primary studies. Moreover, the references of a substantive work do not come as unevaluated listings (like those in an anonymous bibliography), but as choices by an author whose judgment one can assess in the work itself. They are thus more like scholarly intelligence than raw data, especially if one values the author who provides them.

Reference chasing is obviously a two-stage process: to follow up on those made by someone else, the work in which they occur must be in hand. Some reference-bearing works will be already known; others must be discovered. Generally, the correlation between the familiarity of works and their physical distance from the researcher will be strong; known items will tend to be near by (for example, in the researcher's office), and unknown items, farther away (for example, in a local or nonlocal library collection or undiscovered online).

The first chasing many researchers do is simply to assemble the publications they already know or can readily discover. Indeed, from the standpoint of efficiency, the best way to begin a search is to first pick studies from one's own shelves and files and then to follow up leads to earlier work from their reference sections. Cooper's example of this was *references in nonreview papers from journals you subscribe to*. The term *snowballing* is used for multiple iterations of such searches—that is, using references to gather documents, then using the latter's references to gather more documents, and so on increasingly (Greenhalgh and Peacock 2005).

According to Green and Hall, meta-analysts doing a literature search should page through the volumes of the

best journals for a topic year by year—presumably not only those to which they personally subscribe, but also those in the library (1984, 46). An alerting service such as *Current Contents Connect* from Clarivate Analytics can be used as well. Despite its apparent cost in time, so-called hand searching may actually prove effective, given the small and focused set of titles in any particular table of contents. Items found in this fashion may lead as a bonus to earlier studies, through what Cooper calls *references in nonreview papers you browsed through at the library* (or nowadays online).

Existing syntheses and reviews are probably the type of works most likely to be known to a researcher, but a good starting assumption is that at least one is unknown. The first goal in a literature search, therefore, should be to discover previous *review papers or books by others* on a topic because they are likely to be both substantively important in their own right and a rich source of references to earlier studies. It is also important to ascertain the *nonexistence* of such reviews (if that is the case), because the researcher's strategy of presentation will depend on whether the literature of a topic has already been surveyed. Articles on the retrieval of systematic reviews are numerous (see, for example, Harrison 1997; Boynton et al. 1998; White et al. 2001; Montori et al. 2005; Papaioannou et al. 2009; Lee et al. 2012; Lunny, McKenzie, and McDonald 2016). It has been claimed that Google Scholar by itself is sufficient for such retrieval (Gehanno, Rollin, and Darmoni 2013); the ensuing controversy may be traced online.

Another goal in some cases should be to discover topical bibliographies compiled by others. The entries in freestanding bibliographies may be harder to judge for relevance than references in a book or article, especially if abstracts or summaries are lacking. Nevertheless, if a bibliography exists, tracking down its entries may diminish the need for further retrievals (Mann 2015, 169–77). Such bibliographies might be especially useful in discovering items not at the research front but of historical interest.

Mann's *related records searches* are a sophisticated form of reference chasing that became practical only with computerization. They make it possible to find articles that cite identical works. For example, the full record of a single article in the Web of Science (WoS) includes a Find Related Records hyperlink. Following that link will retrieve all articles that refer to at least one of the same earlier writings as the seed article. They are ranked high to low by the number of references they share with the seed. In some cases, the shared cited works can themselves be retrieved through the WoS interface.

Reference chasing is an inviting method—so inviting that the danger lies in stopping with it, on the pretext that any other method will quickly produce diminishing returns. A person serious about finding primary studies to consider will not assume without checking that they do not exist or that diminishing returns begin outside the office door. The substantive reason for concern is the possibility of failing to capture the full range of studies (and reported effects) that exist. Just as researchers' references may reflect personal biases, so may their collection of books and journal articles. The authors of these books and articles will tend to cite works compatible with biases of their own. In journals, this tendency produces citation networks that link some journals tightly and others loosely or not at all. The result is that specialties and disciplines fail to communicate despite interests in common. Thus, reference chasing may simply reinforce the homogeneity of findings, and other methods are needed to learn whether unknown but relevant writings exist. The next strategy is only a partial solution, as we shall see.

4.8 CONSULTATION

Many researchers trust people over bibliographies for answers on what is worth reading. Wilson's "consultation"—the finding of usable studies by talking and writing to others rather than by online searching—is illustrated in table 4.2 (1992). Cooper gives examples such as *informal conversations at conferences or with students*; Mann mentions ways of getting in touch. Actually, one is still searching bibliographies; they are simply inside people's heads. Everyone, including the most learned, does this; there is no more practical or fruitful way of proceeding. The only risk lies in relying on a personal network to the exclusion of other sources—a risk similar to overreliance on a personal library. Bibliographies, however dull, will help a person search a literature more thoroughly than recommendations from colleagues, however responsive.

It is not uncommon now for researchers to state formally that they consulted experts in seeking studies to review. A brief article in support of consultation reported that, for a review of studies of primary medical care, fifty of the items selected were found in databases, thirty-one through hand searching, and forty through expert advice (McManus et al. 1998). Some of the items were named by more than one source, but the McManus team added

that twenty-four of the references would have been missed entirely without the experts. A librarian objected in a letter that this merely knocks down a straw man, because no one in librarianship would claim that MEDLINE searches invariably generate comprehensive bibliographies (Due 1999). In a letter below Due's, the McManus team responded that what is self-evident to librarians is not necessarily so to clinicians. This may be true; one does hear of innocents who think that database or Web searches retrieve everything that exists on a topic.

The quality of advice that researchers can get depends on how well connected they are, and also perhaps on their energy in seeking new ties. Regarding the importance of personal connections, a noted author on scientific communication, the late Belver Griffith, told his seminar students, "If you have to search the literature before undertaking research, you are not the person to do the research." He was being only slightly ironic. In his view, you may read to get to a research front, but you cannot stay there by waiting for new publications to appear; you should be in personal communication with the creators of the literature and other key informants before your attempt at new research or synthesis begins. Some of these *people who typically share information with you* may be local—that is, colleagues in your workplace with whom face-to-face conversations are possible. Others may be geographically dispersed but linked through telephone calls, email, conferences, workshops, and so on in your "invisible college" (Cronin 1982; Price 1986).

Invisible colleges are social circles whose members communicate with each other because they share an intense interest in a set of research problems (Crane 1972). Although the exact membership of an invisible college may be hard to define, a nucleus of productive and communicative insiders can usually be identified. A nonmember of this core group can still try to elicit its advice on the literature by making *formal requests of scholars you knew were active in the field* or by seeking *comments from readers or reviewers of past work.*

The *National Faculty Directory,* the *Research Centers Directory,* and handbooks of professional organizations are useful when trying to find addresses or telephone numbers of persons to contact. Most research universities have easy-to-find faculty and departmental webpages containing CVs, interviews, and videos, as well as links to publications available to colleagues across disciplines. Moreover, important developments in science and social science are reported in major newspapers, posted on websites, tweeted, and blogged. Web searches for possible consultants are therefore often surprisingly productive, although beset by the problem of homonymic names (such as multiple Howard D. Whites).

Although consultation with people may well bring published writings to light, its unique strength lies in revealing unpublished works (Royle and Milne 2003). Clarity on the difference between published and unpublished is essential to appreciating this point. Writings such as doctoral dissertations and Education Resources Information Center (ERIC) reports are often called unpublished. Actually, a document has been published in the legal sense if anyone may copy it or obtain a copy of it (Strong 1990). This would include PDF files on the Web and machine-readable data files in archives. The confusion occurs because people associate publication with editorial quality control; they write *unpublished* when *unrefereed* or *not independently edited* would be more accurate. Conference papers are published by professional organizations, whose standards of refereeing run from very high to very low. Documents such as e-prints, reports, dissertations, and data files are published by their authors (or other relatively uncritical groups). Preprints—manuscripts sent to requestors after being accepted by an editor—are simply not yet published.

Most bibliographies and bibliographic databases cover only published documents. Yet research synthesists place unusual stress on including effects from unpublished studies in their syntheses to counteract a potential bias among editors and referees for publishing only effects that are statistically significant (Chalmers, Frank, and Reitman 1990; McAuley et al. 2000; Rothstein, Sutton, and Borenstein 2005). Documents of limited distribution—reports, dissertations, and other grey literature—may be available through standard online bibliographic services, such as *Dissertation Abstracts International* and the [U.S.] National Technical Information Service. Truly unpublished documents or data are only circulated among members of an invisible college (for example, papers in draft from distant colleagues) or revealed through extensive solicitation campaigns (for example, *general requests to government agencies).* There is no bibliographic control over unreleased writings in file drawers except the human memories one hopes to tap.

Consultation may lead to studies not cited because they have never been published, and, like reference chasing, it produces bibliographic advice that is selective rather than uncritical. But also like reference chasing, it can introduce bias into the search. The problem again is too much homogeneity in what people recommend.

Members of an invisible college may recommend only studies that support a dominant set of beliefs. This may be even more likely within groups in daily contact, such as departmental colleagues or teachers and their students.

The countermeasure is to seek heterogeneity. The national conferences of research associations are trade fairs for people with new ideas. Their bibliographic counterparts—big, open, and diversified—are the national bibliographies, the disciplinary abstracting and indexing services, the online library catalogs, the classified library stacks. What many researchers would regard as the weakness of these instruments—that editorially they are so inclusive—may be for research synthesists their greatest strength. We turn now to the ways in which one can use large-scale, critically neutral bibliographies to explore existing subject literatures, as a complement to the more selective approaches of reference chasing and consultation.

4.9 SEARCHES WITH SUBJECT INDEXING

Wilson defines subject searching as "the strategy of approaching the materials we want indirectly by using catalogs, bibliographies, indexes: works that are primarily collections of bibliographical descriptions with more or less complete representations of content" (1992, 156). Although it seems straightforward, retrieval of unknown publications by subject has in fact provided design problems to LIS for more than a century. The main problem, of course, is to effect a match between the searcher's expressed interest and the documentary description.

Only a few practical pointers can be given here. The synthesist wanting to improve recall of publications through a *manual search* or *computer search of abstract data bases* needs to understand the different ways in which topical literatures may be broken out from larger documentary stocks. Searchers get what they ask for, and so it helps to know that different writings will be retrieved by different ways of asking. What follows is intended to produce greater fluency in the technical vocabulary of document retrieval.

Various kinds of indexing bind groups of publications into literatures. These are authors' natural-language terms, indexers' controlled vocabulary terms, names of journals (or monographic series of books), and authors' citations. Although all are usable separately in retrievals from printed bibliographic tools, they can also be combined, to some degree, in online searching to improve recall or precision or both. If the synthesist is working with a professional searcher, all varieties of terms should be discussed in the strategy-planning interview. Moreover, the synthesist would do well to try out search statements personally, so as to learn how different expressions perform.

4.9.1 Natural Language and Keywords

When authors write, they manifest their topics with terms such as "aphasia," "teacher burnout," or "criterion-referenced education" in their titles, abstracts, and full texts. Because these terms are not assigned by indexers from controlled vocabularies but emerge naturally from authors' vocabularies, librarians call them *natural language* or *keywords*. Thus, insofar as all documents with "teacher burnout" in their titles or abstracts constitute a literature, that entire literature can be retrieved. (Many systems can now find natural-language terms of interest in multiple fields of a bibliographic record.) Moreover, Google searches are now carried out across the full texts of documents, and those full texts themselves may be downloadable.

Keywords is a slippery designation. By it, librarians usually mean all substantive words in an author's text. Authors may mean only the five or six phrases with which they index their own writings at the request of editors. Retrieval systems designers have used it to mean all nonstoplisted terms in a database, including the controlled vocabulary that indexers add. It can also mean the search terms that spontaneously occur to online searchers.

Google has so accustomed people to searching with natural language off the top of their heads that they may not realize there is any other way. (Other ways include controlled vocabulary and forward citation retrievals.) The Google search algorithm, though powerful, cannot yet distinguish among the different senses of a natural-language query such as "Wall Street." This explains why Google retrievals are often both huge and noisy—a failure of precision. Nor can Google expand a natural-language query to pick up different ways of expressing the same or related concepts; that is, it does not automatically link an input term such as "urban renewal" to "revitalization" and "gentrification"; or an input term such as "retirees" to "elderly," "older people," "senior citizens," and "old-age pensioners." This explains why Google retrievals miss potentially valuable material—a failure of recall. Regarding the latter, a famous 1985 study by David Blair and M. E. Maron showed that lawyers who trusted computerized natural-language retrieval to bring them all documents relevant to a major litigation got only

about one-fifth of those that were relevant. The lawyers thought they were getting three-quarters or more. Google taps by far the largest collection of documents in history—a wonderful thing—but it has not solved the recall problem that Blair and Maron addressed.

Research synthesists have a special concern in that they generally want to find empirical writings with measured effects and acceptable research designs (Cooper and Ribble 1989; Dieste and Padua 2007). This is a precision problem within the larger problem of achieving high recall (White et al. 2001). That is, given a particular topic, a researcher does not want to retrieve literally all writings on it—only those with a certain empirical content. Therefore, the strategy should be to specify the topic as broadly as possible (with both natural language and controlled vocabulary terms), but then to qualify the search by adding terms designed to match specific natural language in abstracts, such as "ANOVA," "random-" "control-" "t test," "F test" and "correlat-." Professional searchers can build up large, reusable groups of such terms, called hedges, so that if any one of them appears in an abstract, the document is retrieved (Bates 1992). A database of existing hedges is maintained by the Health Information Research Unit at McMaster University (2016).

The search strategy beginning with "ANOVA" presumes that abstracts of empirical studies state methods and results. It would thus be most useful in partitioning a large subject literature into empirical and non-empirical components. Of course, it would also exclude empirical studies that lacked the chosen signal words in their abstracts, and so would have to be used carefully. Nevertheless, it confers a valuable power.

Some professional searchers distinguish between hedges and filters (Campbell 2016). For them, hedges are reusable lists of *subject* terms, whereas filters are reusable *nonsubject* terms added to search strings to limit retrievals. Searchers for systematic reviews in medicine filter documents on grounds such as document type, research methods, characteristics of experimental subjects, and relation to clinical concepts (diagnosis, prognosis, treatment). The InterTASC Information Specialists' Sub-Group has published on the Web an extensive guide to search filters (Glanville, Lefebvre, and Wright 2008).

4.9.2 Controlled Vocabulary

Controlled vocabulary pertains to the terms added to the bibliographic record by the employees of abstracting and indexing services or large research libraries—broadly speaking, indexers. The major reason for controlled vocabulary is that authors' natural language in titles, abstracts, and full texts may scatter related writings rather than bringing them together (Mann 2015, 76). Controlled vocabulary counteracts scattering by restating the content of documents in standardized headings (such as "senior citizens" to unify documents that use various terms for that concept). It thereby creates literatures for searchers who would otherwise have to guess how authors might express a subject. Blair and Maron's lawyers failed on just this count, and they are not alone.

Controlled vocabulary consists of such things as hierarchical classification codes and subject headings for books, as well as descriptors for articles and reports (Mann 2015). For example, codes from the *Library of Congress Classification* scheme are assigned one to a book, so that each book will have a single position in collections arranged for browsing. On the other hand, catalogers may assign more than one heading per book from the *Library of Congress Subject Headings.* In the past, they usually assigned no more than three, because tables of contents and back-of-the-book indexes were presumed as complements, but now terms may be assigned much more generously.

Subject headings are as a rule the most specific terms (or compounds of terms, such as "reference services—automation—bibliographies") that match the scope of an entire work. In contrast, descriptors, taken from a list called a thesaurus, name salient concepts in writings rather than characterizing the work as a whole. Unlike compound subject headings, which are combined by indexers before any search occurs, descriptors are combined by the searcher at the time of a computerized search. Because the articles and reports they describe often lack the internal indexes seen in books, they are applied more liberally than subject headings, eight or ten being common.

Descriptors are important because they are used to index the journal and report literatures that synthesists typically want, and a nodding familiarity with tools such as the *Thesaurus of ERIC Descriptors,* the *Medical Subject Headings,* or the *Thesaurus of Psychological Index Terms* will be helpful in talking with librarians and information specialists when searches are to be delegated. Thesauri, created by committees of subject experts, enable one to define research interests in standardized language. They contain definitions of terms (called scope notes), pointers from nonpreferred terms to preferred equivalents ("for criterion-referenced education use competency-based education"), and displays of term hierarchies ("aversion therapy" is a kind of "behavior modification" and is placed under it). Their drawback is

that they are always a bit behind authors' natural language, because committee-based standardizations of vocabulary take time. For example, the term "evidence-based medicine" was not approved in *Medical Subject Headings* until 1997, although it had appeared in journals at least five years earlier (Harrison 1997). Therefore, a searcher should combine descriptors with natural language to achieve the desired fullness of recall.

In keeping with the earlier distinction between C2 reviews and Cochrane reviews, the social sciences present searchers with graver retrieval problems than medicine. According to Lesley Grayson and Alan Gomersall, the problems include "a more diverse literature; the greater variety and variability of secondary bibliographical tools; the increasing availability of material on the internet; and a less precise terminology" (2003, 2). Analyzable social-scientific studies are thus frequently harder to find than medical studies. Materials from certain nations or regions, especially those not in English, may not be covered by any tool. Abstracts may be missing or inadequate. Publications may lack vocabulary control altogether. If thesauri exist, they may be inconsistent in the descriptors and other types of indexing they offer, and indexers may apply them inconsistently. Some indexers, for example, might apply terms naming research methods, while other indexers omit them. Alison Wallace and her colleagues discuss at length their problems in retrieving documents relevant to housing policy (2006).

Supplements to descriptors include *identifiers* (specialized topical terms not included in the thesaurus) and *document types* (which partition literatures by publication format, such as article or book review, rather than subject). Being able to qualify an online search by document type allows synthesists to break out one of their favorite forms, past reviews of research, from a subject literature. For example, Marcia Bates combined descriptors and document types in a search statement still usable today in the ERIC database: "Mainstreaming AND (Deafness OR Hearing Impairments OR Partial Hearing) AND (Literature Reviews OR State of the Art Reviews)" (1992, 211).

Recalling the earlier distinction, the related subject descriptors "Deafness, Hearing Impairments, Partial Hearing" might go into a hedge, while the document-type descriptors would be a filter. In the Web of Science and Scopus, document types such as research reviews can be retrieved by checkbox.

Raya Fidel states that some online searchers routinely favor natural language over subject descriptors and document-type descriptors that require thesaurus lookups (1991). Her observation, which antedates the Web and Google, would be even more true today (Mann 2015, 114–15). Relatively few nonprofessional searchers ever learn to distinguish between natural language and controlled vocabulary; everything is keywords. Contemporary retrieval systems reflect the principle of least effort for the greatest number; they cater to searchers who do not know about controlled vocabulary and would probably not use thesauri even if they did. But that does not mean that most searchers are good searchers where both high recall and high precision are concerned (Mann 2015, 316–19). To retrieve optimally for systematic reviews, searchers should be able to exploit all resources available.

4.9.3 Journal Names

When editors accept contributions to journals, or, in the case of books, to monographic series, they make the journal or series name a part of the writing's bibliographic record. Writings so tagged form literatures of a kind. Abstracting and indexing services usually publish lists of journals they cover, and many, though not all, of the journal titles may be read as if they were broad subject headings.

To insiders, names such as *American Sociological Review* or *Psychological Bulletin* connote not only a subject matter but also a level of authority. The latter is a function of editorial quality control, and can be used to rank journals in prestige, which implicitly extends to the articles appearing in their pages. Earlier, a manual search through "the best" journals was mentioned—probably for items thought to be suitably refereed before publication. Actually, if the goal is to confine a subject search to certain journals, the computer offers an alternative. Journal and series names may be entered into search statements just like other index terms. Thus, one may perform a standard subject or citation search but restrict the final set to items appearing in specific journals or series. For example, one could combine "aphasia AND Journal of Verbal Learning and Verbal Behavior;" there might still be a fair number of abstracts to browse through, but all except those from the desired journal would be winnowed out.

4.10 BROWSING

Browsing, too, is a form of bibliographic searching. However, *browsing through library shelves* is not a majority strategy; only about one in four claimed to do it when Cooper wrote, and that was in the 1980s. Even delegating it to a librarian or information specialist is problematical,

because success depends on recognizing useful books or other writings that were not foreknown, and people differ in what they recognize.

The referees of a draft of this chapter claimed that no one browses in libraries any more. Today researchers set up journal table of contents feeds, visit websites, or follow conference hashtags on Twitter. They use blog feeds, saved searches, journal apps, and other automated alerts to monitor the latest prepublication drafts, conference proceedings, grey literatures, or open source publications. But wherever one browses, the principle is the same:

Library shelves: Face book spines classed by librarians and look for items of interest.

Journals: Subscribe to articles bundled by editors and look for items of interest.

Web: Use search terms to form sets of documents and look for items of interest.

Library classification codes, journal names, and terms entered into search engines are all ways of assembling documents on grounds of their presumed similarity. People who browse examine these collections in hopes they can recognize valuable items, as opposed to seeking known items by name. In effect, they let collections search them, rather than the other way around.

Wilson calls browsing book stacks "a good strategy when one can expect to find a relatively high concentration of things one is looking for in a particular section of a collection; it is not so good if the items being sought are likely to be spread out thinly in a large collection" (1992, 156). The Library of Congress (or Dewey) classification codes assigned to books are supposed to bring subject-related titles together so that knowledge of one book will lead to knowledge of others like it, and serendipitous finds do take place. It is also true that synthesists do more than gather articles to review. They need background knowledge of many kinds, and even old-fashioned library browsing may further its pursuit in some topical areas. For example, several textbooks on meta-analysis can be found if one traces the class number of, say, Kenneth Wachter and Myron Straf (1990) to the H62 section in large libraries using the Library of Congress classification scheme.

4.11 CITATION SEARCHES

The last type of strategy identified in table 4.2 is, in Cooper's language, the *manual* or *computer search of a citation index.* Authors' citations make topical or methodological links between studies explicit, and networks of cited and citing publications thus constitute literatures. To repeat, in forward citation searching, earlier writings become terms by which to retrieve the later items that cite them. This kind of retrieval has much to recommend it because it tends to produce hits different from those produced by retrievals with natural language or controlled vocabulary. In other words, it yields writings related to the topic that have relatively little overlap with those found by other methods (Pao and Worthen 1989).

The reasons for the lack of overlap are, first, that authors inadvertently hide the relevance of their work to other studies by using different natural language, which has a scattering effect; and, second, that indexers often fail to remedy this when they apply controlled vocabulary. Luckily, the problems of both kinds of terminology are partly corrected by citation linkages, which are vocabulary independent. They are also authors' linkages rather than indexers', and so presumably reflect greater subject expertise. Last, citation databases are multidisciplinary, and citations frequently cut across disciplinary lines. The items that cite a given work may be quite numerous, which improves the chances for disciplinary diversity. Hence, anyone who wants to achieve high recall should use forward citation searches with other retrieval techniques.

In the Web of Science, the cited items that one enters as search terms may be writings of any sort—books, articles, reports, government documents, conference papers, dissertations, films, and so on. However, the later citing items retrieved from the best-known WoS databases (Science Citation Index, Social Sciences Citation Index, and Arts and Humanities Citation Index) are taken only from journals. Thus, the articles (or other journal pieces) that cite the entry document can be retrieved, but not the books or conference papers that cite it. This fact was bemoaned for many years, and now, as a solution, WoS offers subscribers a separate Book Citation Index (coverage from 2005) and a Conference Proceedings Citation Index (coverage from 1990).

The WoS databases are available for separate or joint searching. The searcher can begin with modest bibliographic data: a known document, author, or organization. If need be, such data can be discovered by searching initially on natural language from the titles or abstracts of citing articles. Title terms may be intersected with names of cited documents or cited authors to improve precision.

The situation in Scopus is roughly parallel. It too originally allowed cited documents of any type as input but could retrieve only citing journal articles. Its claim to

fame was that its coverage of citing journals was even broader than that of WoS (if not as deep chronologically). Now it makes document types other than articles retrievable through its main interface—selected book series, conference proceedings, trade publications, and patents. By making books retrievable, both it and WoS seek to attract users from book-oriented fields in the social sciences and humanities.

Unlike WoS and Scopus, Google Scholar (GS) is free of charge and its ready availability has won it wide use. Completely automated, it covers citations to any sort of document from any sort of document, as long as they can be found by its Web crawlers. Hence, GS frequently records far more citations to an item than WoS or Scopus. Given an entry phrase such as words from a title, an author's name, or a subject term, it will retrieve documents on the Web bearing that phrase, plus the documents that cite them. That is, the retrieved documents are ranked high to low by the number of documents that cite them, and the latter can themselves be retrieved through clickable links. This makes GS good for extending one's reach in literature retrieval. Precisely because its use is so seductive, however, one should know that its lack of vocabulary control (and editorial oversight in general) can lead to numerous false drops.

The WoS databases also have problems of vocabulary control. For example, where authors are concerned, different persons can have the same name (homonyms), and the same person can have different names (allonyms, a coinage of White 2001). Homonyms degrade precision, because works by authors or citees other than the one the searcher intended will be lumped together with it in the retrieval and must be disambiguated. The WoS practice of using only initials for first and middle names worsens the problem. (For instance, "Lee, AJ" unites documents that "Lee, Alan J." and "Lee, Amos John" would separate. Authors' affiliations, if available, often disambiguate names, but not always.) Allonyms by contrast degrade recall, because the searcher may be unable to guess all the different ways an author's or citee's name has been copied into the database. The same author usually appears in at least two ways (for example, "Small HG" and "Small H"). Derek J. de Solla Price's name is cited in more than a dozen ways. Authors also change their names for various reasons, such as marriage. A nonproprietary service called ORCID addresses homonym and allonym problems by giving authors unique ID numbers on request.

One other matter relates to Clarivate Analytics, which obtained its citation databases from the former Institute for Scientific Information (ISI) and continues ISI's policy of indexing only journals that are cited above a certain threshold (Testa 2016). This policy has led to accusations of bias in favor of U.S. or North American or Anglo-American sources. The WoS indexes do have an English-language bias, because of the international prominence of English in science and scholarship; even so, leading journals in other languages are indexed. Yet researchers routinely lament the absence of journals they deem important. Unfortunately, the threshold is dictated by economic constraints on Clarivate Analytics and its subscribers, and many thousands of journals in both English and other languages will never make the cut. Tools such as Scopus, Google Scholar, Google Books, and the Chinese Social Sciences Citation Index are needed to extend researchers' capabilities for retrieval across relatively less-cited journals, languages, and formats.

4.12 FINAL OPERATIONS

The synthesist's ideal in gathering primary studies is to have the best possible pool from which to select those finally analyzed. Principled stopping rules for searchers appear in articles by Monika Kastner and her colleagues (2009) and by Andrew Booth (2010). In practice, the selection of studies has a discretionary element, attributable to individual tastes and expectations about readerships. Different synthesists interested in the same topic may differ on the primary studies to be included. If a set of studies is challenged as incomplete, it might be claimed that gathering a different set was not cost effective. The strategies and criteria actually used in searches will determine whether that defense has merit.

4.12.1 Judging Relevance

The comprehensiveness of a search may be in doubt, but not the fact that meta-analysis requires documents with comparable, quantified findings. The need for studies with appropriate data sharply divides what is relevant from what is not; an explicit test is the presence or absence of usable statistical tables. Of course, if nonquantitative studies are included, relevance judgments may be more difficult. Presumably, the judges will be scientists or advanced students with domain knowledge, high interest in the literature to be synthesized, and adequate time in which to make decisions. Further, they will be able to decide on the basis of abstracts or, better, full texts of documents. That limits the sources of variation in rele-

vance judgments to factors such as the uses to which documents will be put and the order in which documents are considered.

A remaining variable is openness to information. Although research synthesists as a group seem highly open to information, some are more open than others. This probably influences not only their relevance judgments, but also their appetites for multimodal literature searches and their degree of tolerance for studies of less than exemplary quality. Some want to err on the side of inclusiveness, and others want stricter editorial standards.

Where ancillary services are concerned, health professionals have doubted librarians' ability to judge the relevance of meta-analyses for evidence-based practice (Lewis, Urquhart, and Rolinson 1998). However, librarians and information specialists have long prepared to work on this front and by now may bring considerable sophistication to it (Cumming and Conway 1998; Tsafrir and Grinberg 1998; Wade et al. 2006; Rankin, Grefsheim, and Canto 2008; Grant and Booth 2009; Sheble 2014, 2016).

4.12.2 Document Delivery

The final step in literature retrieval is to obtain copies of items judged relevant. The online vendors cooperate with document suppliers so that hard copies of items may be ordered as part of an online search, often through one's academic library. Frequently the grey literature, such as dissertations, technical reports, and government documents, can also be acquired on the Web or through specialized documentation centers. When local collections fail, interlibrary loan services are generally reliable for books, grey publications, and photocopies of articles. The foundation of international interlibrary loan in North America is the Online Computer Library Center (OCLC) system, whose more than 435 million bibliographic records include those for a growing number of numeric data files in machine-readable form. In the United Kingdom, the British Library at Boston Spa serves an international clientele for interlibrary transactions. The transfer of texts, software, and data files is now routinely done by computer. Nevertheless, bibliographic organizations such as OCLC and the British Library are needed to make resources discoverable in the digital wilderness.

The key to availing oneself of these resources is to work closely with librarians and information specialists, whose role in scientific communication has received more emphasis here than is usual, so as to raise their visibility. If they themselves do not belong to synthesis teams, they may be able to advise or train team members in matters such as bibliographic databases and software, search strategies, and ways of obtaining documents. Locating an expert in high recall may take effort; persons with the requisite motivation and knowledge are not at every reference desk. But such experts exist, and it is worth the effort to find them.

4.13 REFERENCES

Ananiadou, Sophia, Brian Rea, Naoiaki Okazaki, Rob Procter, and James Thomas. 2009. "Supporting Systematic Reviews Using Text Mining." *Social Science Computer Review* 27(4): 509–23.

Atkinson, Kayla M., Alison C. Koenka, Carmen E. Sanchez, Hannah Moshontz, and Harris Cooper. 2015. "Reporting Standards for Literature Searches and Report Inclusion Criteria: Making Research Syntheses More Transparent and Easy to Replicate." *Research Synthesis Methods* 6(1): 87–95.

Bastian, Hilda, Paul Glasziou, and Iain Chalmers. 2010. "Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up?" *PLoS Medicine* 7(9): e1000326.

Bates, Marcia J. 1976. "Rigorous Systematic Bibliography." *RQ* 16(1): 7–26

———. 1992. "Tactics and Vocabularies in Online Searching." In *For Information Specialists; Interpretations of Reference and Bibliographic Work,* by Howard D. White, Marcia J. Bates, and Patrick Wilson. Norwood, N.J.: Ablex.

Bayliss, Susan E., Clare F. Davenport, and Mary E. Pennant. 2014. "Where and How to Search for Information on the Effectiveness of Public Health Interventions—A Case Study for Prevention of Cardiovascular Disease." *Health Information & Library Journal* 31(4): 303–13.

Beahler, Chris C., Jennifer J. Sundheim, and Naomi I. Trapp. 2000. "Information Retrieval in Systematic Reviews: Challenges in the Public Health Arena." *American Journal of Preventive Medicine* 18(4): 6–10.

Bernier, Charles L., and A. Neil Yerkey. 1979. *Cogent Communication; Overcoming Reading Overload.* Westport, Conn.: Greenwood Press.

Blair, David C., and M. E. Maron. 1985. "An Evaluation of Retrieval Effectiveness for a Full-Text Document Retrieval System." *Communications of the ACM* 28(3): 289–99.

Bland, Carole J., Linda N. Meurer, and George Maldonado. 1995. "A Systematic Approach to Conducting a Nonstatistical Metaanalysis of Research Literature." *Academic Medicine* 70(7): 642–53.

Boaz, Annette, Deborah Ashby, and Ken Young. 2002. "Systematic Reviews: What Have They Got to Offer Evidence Based Policy and Practice?" *ESRC UK Centre for Evidence Based Policy and Practice* working paper no. 2. London: King's College London. Accessed November 24, 2018. http://www.kcl.ac.uk/sspp/departments/politicaleconomy/research/cep/pubs/papers/paper-02.aspx.

Booth, Andrew. 2010. "How Much Searching Is Enough? Comprehensive versus Optimal Retrieval for Technology Assessments." *International Journal of Technology Assessment in Health Care* 26(4): 431–35.

Booth, Andrew, Anthea Sutton, and Diana Papaioannou. 2016. *Systematic Approaches to a Successful Literature Review*, 2nd ed. London: Sage Publications.

Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. 2009. *Introduction to Meta-Analysis*. Chichester, UK: John Wiley & Sons.

Boynton, Janette, Julie M. Glanville, David McDaid, and Carol Lefebvre. 1998. "Identifying Systematic Reviews in MEDLINE: Developing an Objective Approach to Search Strategy Design." *Journal of Information Science* 24(3): 137–54.

Bramwell, Vivien H. C., and Christopher J. Williams. 1997. "Do Authors of Review Articles Use Systematic Methods to Identify, Assess and Synthesize Information?" *Annals of Oncology* 8(12): 1185–95.

Bronson, Denise E., and Tamara S. Davis. 2012. *Finding and Evaluating Evidence: Systematic Reviews and Evidence-Based Practice*. New York: Oxford University Press.

Brown, Cecilia. 2010. "Communication in the Sciences." *Annual Review of Information Science and Technology* 44:287–316.

Campbell, Sandy. 2016. "What Is the Difference Between a Filter and a Hedge?" *Journal of the European Association for Health Information and Libraries* 12(1): 4–5.

Campbell Collaboration. 2016. "What Is a Systematic Review?" Accessed December 20, 2018. https://www.campbellcollaboration.org/research-resources/writing-a-campbell-systematic-review/systemic-review.html.

Centre for Reviews and Dissemination. 2009. *Systematic Reviews: CRD's Guidance for Undertaking Reviews in Health Care*, 3d ed. York: CRD, University of York. Accessed December 20, 2018. http://www.york.ac.uk/media/crd/Systematic_Reviews.pdf.

Chalmers, Thomas C., Cynthia S. Frank, and Dinah Reitman. 1990. "Minimizing the Three Stages of Publication Bias." *Journal of the American Medical Association* 263(10): 1392–95.

Cheung, Mike W.-L. 2015. *Meta-Analysis: A Structural Equation Modeling Approach.* Chichester, UK: John Wiley & Sons.

Committee on Scientific and Technical Communication. 1969. *Scientific and Technical Communication: A Pressing National Problem and Recommendations for Its Solution.* Washington, D.C.: National Academy of Sciences.

Conn, Vicki S., Sang-arun Isaramalai, Sabyasachi Rath, Peeranuch Jantarakupt, Rohini Wadhawan, and Yashodhara Dash. 2003. "Beyond MEDLINE for Literature Searches." *Journal of Nursing Scholarship* 35(2): 177–82.

Cooper, Harris M. 1985. "Literature Searching Strategies of Integrative Research Reviewers." *American Psychologist* 40(11): 1267–69.

———. 1987. "Literature-Searching Strategies of Integrative Research Reviewers: A First Survey." *Knowledge: Creation, Diffusion, Utilization* 8(2): 372–83.

———. 2000. "Strengthening the Role of Research in Policy Decisions: The Campbell Collaboration and the Promise of Systematic Research Reviews." Merrill Advanced Studies Center Conference on Making Research a Part of the Public Agenda. June. Accessed December 20, 2018. http://merrill.ku.edu/sites/masc.drupal.ku.edu/files/docs/2000whitepaper.pdf.

———. 2017. *Research Synthesis and Meta-Analysis: A Step-by-Step Approach*, 5th ed. Los Angeles, Calif.: Sage Publications.

Cooper, Harris M., and Ronald G. Ribble. 1989. "Influences on the Outcome of Literature Searches for Integrative Research Reviews." *Knowledge: Creation, Diffusion, Utilization* 10(3): 179–201.

Crane, Diana. 1972. *Invisible Colleges; Diffusion of Knowledge in Scientific Communities.* Chicago: University of Chicago Press.

Cronin, Blaise. 1982. "Invisible Colleges and Information Transfer: A Review and Commentary with Particular Reference to the Social Sciences." *Journal of Documentation* 38(3): 212–36.

Cumming, Lorna, and Lynn Conway. 1998. "Providing Comprehensive Information and Searches with Limited Resources." *Journal of Information Science* 24(3): 183–85.

Davies, Philip. 2000. "The Relevance of Systematic Reviews to Educational Policy and Practice." *Oxford Review of Education* 26(3/4): 365–78.

Dieste, Oscar, and Anna Griman Padua. 2007. "Developing Search Strategies for Detecting Relevant Experiments for Systematic Reviews." First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007). Madrid (September 20–21, 2007).

Dixon-Woods, Mary, Shona Agarwal, Bridget Young, David Jones, and Alex J. Sutton. 2004. "Integrative Approaches to Qualitative and Quantitative Evidence." London: Health Development Agency. Accessed November 24, 2018. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.96.8783&rep=rep1&type=pdf.

Doig, Gordon Stuart, and Fiona Simpson. 2003. "Efficient Literature Searching: A Core Skill for the Practice of Evidence-Based Medicine." *Intensive Care Medicine* 29(12): 2119–27.

Due, Stephen. 1999. "Study Only 'Proves' What Librarians Knew Anyway." *British Medical Journal* 319(7204): 260.

Egger, Matthias, George Davey-Smith, and Douglas G. Altman, eds. 2008. *Systematic Reviews in Health Care: Meta-Analysis in Context,* 2nd ed. New York: John Wiley & Sons.

Fink, Arlene. 2014. *Conducting Research Literature Reviews: From the Internet to Paper.* 4th ed. Thousand Oaks, Calif.: Sage Publications.

Fidel, Raya. 1991. "Searchers' Selection of Search Keys: III. Searching Styles." *Journal of the American Society for Information Science* 42(7): 515–27.

Furlan, Andrea D., Antii Malmivaara, Roger Chou, Chris G. Maher, Rick A. Deyo, Mark Schoene, Gert Bronfort, and Maurits W. van Tulder. 2015. "2015 Updated Method Guideline for Systematic Reviews in the Cochrane Back and Neck Group." *Spine* 40(21): 1660–73.

Ganann, Rebecca, Donna Ciliska, and Helen Thomas. 2010. "Expediting Systematic Reviews: Methods and Implications of Rapid Reviews." *Implementation Science* 5: 56–65.

Garfield, Eugene. 1989. "Reviewing Review Literature." In *Essays of an Information Scientist,* vol. 10. Philadelphia, Pa.: ISI Press.

Gehanno, Jean-François, Laetitia Rollin, and Stefan Darmoni. 2013. "Is the Coverage of Google Scholar Enough to Be Used Alone for Systematic Reviews." *BMC Medical Informatics and Decision Making* 13(7). Accessed November 24, 2018. http://www.biomedcentral.com/1472-6947/13/7.

Glanville, Julie M., Carol Lefebvre, and Kath Wright, eds. 2008. "ISSG Search Filter Resource." York, UK: The InterTASC Information Specialists' Sub-Group. Updated October 2, 2018. Accessed November 24, 2018. https://sites.google.com/a/york.ac.uk/issg-search-filters-resource/home.

Glass, Gene V. 2000. "Meta-Analysis at 25." Unpublished paper, Arizona State University. Accessed November 24, 2018. http://www.gvglass.info/papers/meta25.html.

Grant, Maria J., and Andrew Booth. 2009. "A Typology of Reviews: An Analysis of 14 Review Types and Associated Methodologies." *Health Information and Libraries Journal* 26(2): 92–108.

Grayson, Lesley, and Alan Gomersall. 2003. "A Difficult Business: Finding the Evidence for Social Science Reviews." *ESRC UK Centre for Evidence Based Policy and Practice* working paper 19. London: King's College London. Accessed November 24, 2018. https://www.researchgate.net/publication/260386917.

Green, Bert F., and Judith A. Hall. 1984. "Quantitative Methods for Literature Reviews." *Annual Review of Psychology* 35: 37–53.

Greenhalgh, Trisha, and Richard Peacock. 2005. "Effectiveness and Efficiency of Search Methods in Systematic Reviews of Complex Evidence: Audit of Primary Sources." *British Medical Journal* 331(7524): 1064–65.

Greenhouse, Joel B., Kenneth W. Wachter, and Miron L. Straf. 1990. "The Making of a Meta-Analysis: A Quantitative Review of the Aphasia Treatment Literature." In *The Future of Meta-Analysis,* edited by Kenneth W. Wachter and Miron L. Straf. New York: Russell Sage Foundation.

Grupp, Gunter, and Mary Heider. 1975. "Non-Overlapping Disciplinary Vocabularies. Communication from the Receiver's Point of View." In *Communication of Scientific Information*, edited by Stacey B. Day. Basel: S. Karger.

Harrison, Jane. 1997. "Designing a Search Strategy to Identify and Retrieve Articles on Evidence-Based Health Care Using MEDLINE." *Health Libraries Review* 14(1): 33–42.

Hawker, Sheila, Sheila Payne, Christine Kerr, Michael Hardey, and Jackie Powell. 2002. "Appraising the Evidence: Reviewing Disparate Data Systematically." *Qualitative Health Research* 12(9): 1284–99.

Higgins, Julian P. T., and Sally Green, eds. 2011. *Cochrane Handbook for Systematic Reviews of Interventions*, version 5.1.0. London: The Cochrane Collaboration.

Hunt, Morton. 1997. *How Science Takes Stock: The Story of Meta-Analysis.* New York: Russell Sage Foundation.

Hunter, John E., and Frank L. Schmidt. 2004. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*, 2nd ed. Thousand Oaks, Calif.: Sage Publications.

Jackson, Gregg B. 1978. *Methods for Reviewing and Integrating Research in the Social Sciences.* PB-283 747. Springfield, Va.: National Technical Information Service.

———. 1980. "Methods for Integrative Reviews." *Review of Educational Research* 50(3): 438–60.

Jesson, Jill, Lydia Matheson, and Fiona M. Lacey. 2011. *Doing Your Literature Review: Traditional and Systematic Techniques*. London: Sage Publications.

Jones, Myfanwy Lloyd. 2004. "Application of Systematic Methods to Qualitative Research: Practical Issues." *Journal of Advanced Nursing* 48(3): 271–78.

Kastner, Monika, Sharon E. Straus, K. Ann McKibbon, and Charlie Goldsmith. 2009. "The Capture-Mark-Recapture Technique Can Be Used as a Stopping Rule When Searching in Systematic Reviews." *Journal of Clinical Epidemiology* 62(2): 149–57.

Khangura, Sara, Kristin Konnyu, Rob Cushman, Jeremy Grimshaw, and David Moher. 2012. "Evidence Summaries: The Evolution of a Rapid Review Approach." *Systematic Reviews* 1(10). Accessed November 24, 2018. http://www.systematicreviewsjournal.com/content/1/1/10.

Kugley, Shannon, Anne Wade, James Thomas, Quenby Mahood, Anne-Marie Klint Jørgensen, Karianne Thune Hammerstrøm, and Nila Sathe. 2016. *Searching for Studies: A Guide to Information Retrieval for Campbell Systematic Reviews*. Oslo: The Campbell Collaboration. Accessed November 24, 2018. http://www.campbellcollaboration.org/artman2/uploads/1/Campbell_Methods_Guides_Information_Retrieval_1.pdf

Laird, Nan M. 1990. "A Discussion of the Aphasia Study." In *The Future of Meta-Analysis,* edited by Kenneth W. Wachter and Miron L. Straf. New York: Russell Sage Foundation.

Lee, Edwin, Maureen Dobbins, Kara DeCorby, Lyndsey McRae, Daiva Tirilis, and Heather Husson. 2012. "An Optimal Search Filter for Retrieving Systematic Reviews and Meta-Analyses." *BMC Medical Research Methodology* 12: 51. Accessed November 24, 2018. http://www.biomedcentral.com/1471-2288/12/51.

Lewis, Ruth A., Christine J. Urquhart, and Janet Rolinson. 1998. "Health Professionals' Attitudes Towards Evidence-Based Medicine and the Role of the Information Professional in Exploitation of the Research Evidence." *Journal of Information Science* 24(5): 281–90.

Lipsey, Mark W., and David B. Wilson. 2001. *Practical Meta-Analysis.* Thousand Oaks, Calif.: Sage Publications.

Littell, Julia H., Jacqueline Corcoran, and Vijayan Pillai. 2008. *Systematic Reviews and Meta-Analysis*. Oxford: Oxford University Press.

Lunny, Carole, Joanne E. McKenzie, and Steve McDonald. 2016. "Retrieval of Overviews of Systematic Reviews in MEDLINE Was Improved by the Development of an Objectively Derived and Validated Search Strategy." *Journal of Clinical Epidemiology* 74 (June): 107–18.

Mann, Thomas. 1993. *Library Research Models: A Guide to Classification, Cataloging, and Computers.* New York: Oxford University Press.

———. 2015. *The Oxford Guide to Library Research*, 4th ed. New York: Oxford University Press.

McAuley, Laura, Ba' Pham, Peter Tugwell, and David Moher. 2000. "Does the Inclusion of Grey Literature Influence Estimates of Intervention Effectiveness Reported in Meta-Analyses?" *Lancet* 356(9237): 1228–231.

McManus, Richard J., Sue Wilson, Brendan C. Delaney, David A. Fitzmaurice, Chris J. Hyde, Ros Tobias, Sue Jowett, and F. D. Richard Hobbs. 1998. "Review of the Usefulness of Contacting Other Experts When Conducting a Literature Search for Systematic Reviews." *British Medical Journal* 317(7172): 1562–63.

McMaster University. Health Information Research Unit. 2016. "Hedges." Accessed November 24, 2018. http://hiru.mcmaster.ca/hiru/HIRU_Hedges_home.aspx.

Moher, David, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G. Altman. 2009. "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement." *PLoS Medicine* 6(7): 1–6.

Moher, David, Jennifer Tetzlaff, Andrea C. Tricco, Margaret Sampson, and Douglas G. Altman. 2007. "Epidemiology and Reporting Characteristics of Systematic Reviews." *PLoS Medicine* 4(3): 447–55.

Montori, Victor M., Nancy L Wilczynski, Douglas Morgan, and R. Brian Haynes. 2005. "Optimal Search Strategies for Retrieving Systematic Reviews from Medline: Analytical Survey." *British Medical Journal* 330(7482): 68. Accessed November 24, 2018. http://bmj.bmjjournals.com/cgi/content/short/bmj.38336.804167.47v1.

Mulrow, Cynthia D., and Deborah Cook. 1998. *Systematic Reviews: Synthesis of Best Evidence for Health Care Decisions.* Philadelphia, Pa.: American College of Physicians.

Noonan, Eamonn, and Arild Bjørndal. 2011. "The Campbell Collaboration: Bringing an Evidence Perspective to Welfare, Justice, and Education." Editorial. *Cochrane Database of Systematic Reviews* 2011:ED000011.

Pao, Miranda Lee, and Dennis B. Worthen. 1989. "Retrieval Effectiveness by Semantic and Citation Searching." *Journal of the American Society for Information Science* 40(4): 226–35.

Papaioannou, Diana, Anthea Sutton, Christopher Carroll, Andrew Booth, and Ruth Wong. 2009. "Literature Searching for Social Science Systematic Reviews: Consideration of a Range of Search Techniques." *Health Information and Libraries Journal* 27(2): 114–22.

Pawson, Ray. 2006. *Evidence-Based Policy: A Realist Perspective*. London: Sage Publications.

Petticrew, Mark. 2001. "Systematic Reviews from Astronomy to Zoology: Myths and Misconceptions." *British Medical Journal* 322(7278): 98–101.

———. 2011. "When Are Complex Interventions 'Complex'? When Are Simple Interventions 'Simple'?" *European Journal of Public Health* 21(4): 397–99.

Petticrew, Mark, and Helen Roberts. 2006. *Systematic Reviews in the Social Sciences: A Practical Guide*. Oxford: Blackwell.

Petitti, Diana B. 2000. *Meta-Analysis, Decision Analysis and Cost-Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine*. New York: Oxford University Press.

Price, Derek J. de Solla. 1986. "Invisible Colleges and the Affluent Scientific Commuter." In his *Little Science, Big Science . . . and Beyond.* New York: Columbia University Press.

Rankin, Jocelyn A., Suzanne F. Grefsheim, and Candace C. Canto. 2008. "The Emerging Informationist Specialty: A Systematic Review of the Literature." *Journal of the Medical Library Association* 96(3): 194–206.

Rosenthal, Robert. 1984. *Meta-Analytic Procedures for Social Research.* Beverly Hills, Calif.: Sage Publications.

Rothstein, Hannah R., Alexander J. Sutton, and Michael Borenstein, eds. 2005. *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments.* New York: Wiley.

Rothstein, Hannah R., Herbert M. Turner, and Julia G. Lavenberg. 2004. *The Campbell Collaboration Information Retrieval Policy Brief.* Philadelphia, Pa.: Campbell Collaboration. Accessed November 24, 2018. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.131.8935&rep=rep1&type=pdf.

Royle, Pamela, and Ruairidh Milne. 2003. "Literature Searching for Randomized Controlled Trials Used in Cochrane Reviews: Rapid versus Exhaustive Searches." *International Journal of Technology Assessment in Health Care* 19(4): 591–603.

Saini, Michael, and Aron Shlonsky. 2012. *Systematic Synthesis of Qualitative Research*. New York: Oxford University Press.

Schlosser, Ralf W., Oliver Wendt, Katie L. Angermeier, and Manisha Shetty. 2005. "Searching for Evidence in Augmentative and Alternative Communication: Navigating a Scattered Literature." *Augmentative and Alternative Communication* 21(4): 233–55.

Sheble, Laura. 2014. "Diffusion of Meta-Analysis, Systematic Review, and Related Research Synthesis Methods: Patterns, Context, and Impact." Ph.D. diss., University of North Carolina at Chapel Hill.

———. 2016. "Research Synthesis Methods and Library and Information Science: Shared Problems, Limited Diffusion." *Journal of the Association for Information Science and Technology* 67(8): 1990–2008.

Shlonsky, Aron, Eamonn Noonan, Julia H. Littell, and Paul Montgomery. 2011. "The Role of Systematic Reviews and the Campbell Collaboration in the Realization of Evidence-Informed Practice." *Clinical Social Work Journal* 39(4): 362–68.

Smeets, Rob Johannes, Derick R. Wade, Alita Hidding, and J. Andre Knottnerus. 2006. "The Association of Physical Deconditioning and Chronic Low Back Pain: A Hypothesis-Oriented Systematic Review." *Disability and Rehabilitation* 28(11): 673–93.

Smith, Valerie, Declan Devane, Cecily M. Begley, and Mike Clarke. 2011. "Methodology in Conducting a Systematic Review of Systematic Reviews of Healthcare Interventions." *BMC Medical Research Methodology* 11(1): 15. Accessed November 24, 2018. http://www.biomedcentral.com/1471-2288/11/15.

Steering Group of the Campbell Collaboration. 2015. *Campbell Systematic Reviews: Policies and Guidelines. Supplement 1*. Oslo: Campbell Collaboration. Accessed November 24, 2018. https://campbellcollaboration.org/library/campbell-collaboration-systematic-reviews-policies-and-guidelines.html.

Strong, William S. 1990. *The Copyright Book; A Practical Guide*, 3rd ed. Cambridge, Mass.: MIT Press.

Stroup, Donna F., Jesse A. Berlin, Sally C. Morton, Ingram Olkin, G. David Williamson, Drummond Rennie, David Moher, Betsy J. Becker, Theresa Ann Sipe, and Stephen B. Thacker. 2000. "Meta-Analysis of Observational Studies in Epidemiology; A Proposal for Reporting." *Journal of the American Medical Association* 283(15): 2008–12.

Testa, James. 2016. "Journal Selection Process." Clarivate Analytics, June 26, 2018. Accessed November 24, 2018. http://wokinfo.com/essays/journal-selection-process.

Thomas, James, Angela Harden, Ann Oakley, Sandy Oliver, Katy Sutcliffe, Rebecca Rees, Ginny Brunton, and Josephine Kavanagh. 2004. "Integrating Qualitative Research with Trials in Systematic Reviews." *British Medical Journal* 328(7446): 1010–12.

Tsafrir, Jenni, and Miriam Grinberg. 1998. "Who Needs Evidence-Based Health Care?" *Bulletin of the Medical Library Association* 86(1): 40–45.

Wachter, Kenneth W., and Miron L. Straf, eds. 1990. *The Future of Meta-Analysis.* New York: Russell Sage Foundation.

Wade, Anne C., Herbert M. Turner, Hannah R. Rothstein, and Julia G. Lavenberg. 2006. "Information Retrieval and the Role of the Information Specialist in Producing High-Quality Systematic Reviews in the Social, Behavioural and Education Sciences." *Evidence & Policy: A Journal of Research, Debate and Practice* 2(1): 89–108.

Wallace, Alison, Karen Croucher, Mark Bean, Karen Jackson, Lisa O'Malley, and Deborah Quilgars. 2006. "Evidence for

Policy Making: Some Reflections on the Application of Systematic Reviews to Housing Research." *Housing Studies* 21(2): 297–314.

Wang, Morgan C., and Brad J. Bushman. 1999. *Integrating Results thorough Meta-Analytic Review Using SAS Software.* Raleigh, N.C.: SAS Institute.

White, Howard D. 1994. "Scientific Communication and Literature Retrieval." In *Handbook of Research Synthesis,* edited by Harris Cooper and Larry V. Hedges. New York: Russell Sage Foundation.

———. 2001. "Authors as Citers over Time." *Journal of the American Society for Information Science* 52(2): 87–108.

———. 2009. "Scientific Communication and Literature Retrieval." In *Handbook of Research Synthesis and Meta-analysis,* 2nd ed., edited by Harris Cooper, Larry V. Hedges, and Jeffrey C. Valentine. New York: Russell Sage Foundation.

White, V. J., Julie M. Glanville, Carol Lefebvre, and T. A. Sheldon. 2001. "A Statistical Approach to Designing Search Filters to Find Systematic Reviews: Objectivity Enhances Accuracy." *Journal of Information Science* 27(6): 357–70.

Whitlock, Evelyn P., Jennifer S. Lin, Roger Chou, Paul W. Shekelle, and Karen A. Robinson. 2008. "Using Existing Systematic Reviews in Complex Systematic Reviews." *Annals of Internal Medicine* 148(10): 776–82.

Wilson, Patrick. 1968. "Bibliographical Instruments and Their Specifications." In *Two Kinds of Power: An Essay on Bibliographical Control*. Berkeley: University of California Press.

———. 1977. *Public Knowledge, Private Ignorance; Toward a Library and Information Policy.* Westport, Conn.: Greenwood Press.

———. 1992. "Searching: Strategies and Evaluation." In *For Information Specialists; Interpretations of Reference and Bibliographic Work,* by Howard D. White, Marcia J. Bates, and Patrick Wilson. Norwood, N.J.: Ablex.

Wong, Sharon S.-L., Nancy L. Wilczynski, and R. Brian Haynes. 2006. "Developing Optimal Search Strategies for Detecting Clinically Sound Treatment Studies in EMBASE." *Bulletin of the Medical Library Association* 94(1): 41–47.

5

SEARCHING BIBLIOGRAPHIC DATABASES

JULIE GLANVILLE
York Health Economics Consortium, York, UK

CONTENTS

5.1 INTRODUCTION

This chapter provides an overview of how to select and search bibliographic databases, which are the backbone of most literature searches undertaken for research synthesis, and discusses the following aspects of searching databases:

- what bibliographic databases are designed to do, what they capture, and how they differ;
- why a range of databases might need to be searched when conducting a research synthesis;
- how to plan and develop a search strategy to search a database, giving consideration to identifying the concepts to be captured by the search strategy, the search terms to be used to capture the concepts and the combinations of terms and concepts;
- how to focus or limit the results of database searches;
- how to peer review a search strategy;
- how to manage search results using reference management software and how to select relevant records from search results; and
- how to record the search as it is being conducted and how to report the search strategy in a final report or other publication.

The chapter focuses on databases that record and index research relevant to the social, behavioral, and medical sciences. It assumes that the reader

- is acquainted with the key journals, handbooks, and reference tools in their discipline and the most important sources on the topic of synthesis;
- intends to conduct an extensive search of the literature, if appropriate to the synthesis question, to identify as many relevant studies as possible to contribute to the synthesis being conducted;
- is familiar with the structure and organization of research libraries and their access via catalogs and inquiry services;
- has read other chapters in this volume, particularly those on scientific communication (chapter 4) and grey literature (chapter 6); and
- is collaborating with an information specialist or librarian with experience of searching bibliographic databases in the topic of the synthesis being undertaken.

Table 5.1 presents a glossary of the key terminology used in the chapter.

Database searching is a skill informed by a knowledge of the design and content of individual databases and the facilities for searching offered by the interfaces to individual databases. Access to databases is provided by search interfaces that can offer a range of facilities to retrieve records. For example, the PsycINFO database (American Psychological Association 2016) is accessible via many interfaces including Ovid, DIMDI, EBSCO, and ProQuest. Each interface is unique in terms of look and feel as well as in the range and format of options it offers for searching and downloading records. Search strategies often need to be adapted to take account of these options.

Searching also draws on an awareness of how researchers might unintentionally fail to communicate their research fully in title and abstracts. In the limited number of words permitted in an abstract, authors may not capture every detail of their methods or list every outcome in their study.

Searching to inform research synthesis is often, but not always, undertaken from the perspective of trying to be exhaustive or extensive. Often, research synthesis seeks to find as much relevant research as possible (to minimize the impact of various biases, including publication bias) while trying to minimize the number of irrelevant research records that need to be processed as a result of the search. This means that search strategies tend to emphasize sensitivity (see table 5.1), and typically researchers will be

Table 5.1 Key Concepts in Database Searching to Inform Research Synthesis

Concept or Term	Definition
Bibliographic database	Collection of records describing publications. Typically each record provides the title, author, source information, and date. Many databases also include an abstract. Some databases may add subject index terms and other codes to facilitate retrieval.
Bibliographic reference management software	Software to manage records that have been downloaded from bibliographic databases. Examples include EndNote, Reference Manager, RefWorks, and Mendeley.
Concept	Topic or theme that forms part of a research question, for example, a population of interest such as children with attention deficit disorder.
Free text terms	Words in an information source record other than the indexing terms. Free-text terms are usually those in the title and abstract of a database record.
Hand searching	Searching the contents of a journal by looking at each article in sequence and making an assessment of the relevance of the article to the synthesis question. Hand searching may also be undertaken for sections of databases or websites.
Indexing	Addition of indexing terms to database records, to provide a single search term for records that might be described by authors using different terms.
Indexing language	Controlled vocabulary used to index records in a database to enhance consistent retrieval of records. For example, the ERIC Thesaurus in the ERIC database of educational research (Institute of Education Sciences n.d.).
Indexing term	Word or phrase from an indexing language.
Information source	Database, website, or library that provides access to research evidence and other documents.
Information specialist	An information scientist or librarian who has extensive experience of searching for research evidence from a variety of information sources.
Interface	Set of options or facilities that are available for searching a database. Options may include ways to combine sets of search results, including Boolean operators and proximity operators, ways to search for word variants (truncation, stemming, and wildcards) and the ability to restrict searches to specific fields such as the title.
Precision	Proportion of relevant records among all the records retrieved by a search strategy (relevant records retrieved divided by all records retrieved). In research synthesis, a precise search strategy is often traded off in favor of high sensitivity.
Proximity operators	Search operators that specify that a search term can be retrieved when it occurs within a certain distance from another search term. The distance can often be varied.
Search filters	Collection of search terms that identifies records about a specific population, study design, or other issue; ideally derived by research.
Search strategy	Collection of search terms used to interrogate a database to identify records relevant to a research synthesis question.
Sensitivity	Proportion of relevant records retrieved by a search strategy from a database (number of relevant records retrieved divided by total number of relevant records). Searches for research synthesis purposes usually aim to be sensitive. Also known as recall.
Specificity	Proportion of irrelevant records not retrieved by a search strategy from a database.
Text mining	Use of software to analyze unstructured text and identify patterns and derive information about a body of literature. In the searching context, can be used to identify terms, phrases, and collocated terms that might be used in search strategies. Also has applications in record selection.

SOURCE: Author's compilation.

willing to trade off precision or specificity (looking at as few irrelevant records as possible) to achieve sensitivity (Wood and Arber 2016b). It also means that searchers may need to search a range of databases because one database is unlikely to record all of the available research. Even where the same research is recorded in two databases, the search options available may result in the record being found in one database but not the other.

Information specialists, such as librarians, are important sources of expertise for teams undertaking research synthesis. Research has shown that trained information specialists tend to find more relevant studies than untrained researchers (Kuller et al. 1993; Erickson and Warner 1998). Information specialists who support research synthesis bring expertise in terms of being able to develop search strategies, knowing the technical issues involved in database searching, and knowing which resources are most relevant to specific questions. Consulting, or better still, involving an information specialist in the research synthesis is highly recommended (Higgins and Green 2011; Eden et al. 2011; Petticrew and Roberts 2006).

This chapter focuses on searching bibliographic (or reference) databases such as PsycINFO or Criminal Justice Abstracts (EBSCO Information Services n.d.a). These databases are collections of records containing the citation details and usually the abstracts of research publications. The records are representations of larger documents and have only a limited number of words to capture the full content of the larger document. To provide additional access points to the documents and so increase opportunities that searches will retrieve them, some databases add subject indexing (or thesaurus) terms to records. Other coding may also be added to database records, such as population codes, age codes or publication types.

Bibliographic databases such as PsycINFO are expensive to produce and access is usually by subscriptions. Academic libraries, professional bodies, and other organizations purchase subscriptions and provide access to databases to their members as a group. The larger the organization, the more databases may be available to an employee. Some large bibliographic databases, such as PubMed (U.S. National Library of Medicine n.d.b) and other government-funded resources, may be accessible free of charge to searchers. One-off access to individual subscription databases for the purposes of contributing to research synthesis can sometimes be purchased direct from the database publisher or via a commercial information intermediary. Professional organizations may provide access to some databases as a membership benefit.

Databases and database searching are constantly evolving. Searching databases as part of the research synthesis process is becoming increasingly evidence based, as can be seen in guidance from international initiatives, national organizations, and published handbooks (Higgins and Green 2011; Kugley et al. 2016; European Food Safety Authority 2010; Centre for Reviews and Dissemination 2009; Petticrew and Roberts 2006; Eden et al. 2011; Joanna Briggs Institute 2014). Guidance on best searching practice is easy to identify in some disciplines such as health care and social sciences, but not in all subjects. Searching toolkits are also available.

Searching for different types of evidence may also require different approaches. Searching for quantitative evidence is the main focus of this chapter. Syntheses of qualitative evidence may be better suited to a more exploratory, organic, and iterative search approach, in which it is not essential to find all relevant studies, but instead to find representative studies that cover an adequate range of relevant issues (Stansfield, Brunton, and Rees 2014). Further information about the evidence base for search approaches to undertaking syntheses of qualitative evidence is provided in a review by the information specialist Andrew Booth (2016). Points where practice may differ between searches for quantitative and qualitative evidence are highlighted in this chapter.

This chapter provides a general introduction, with the expectation that the reader will find relevant specific subject support for their research synthesis question from their local information specialist and from relevant research synthesis guidance in their particular discipline.

5.2 WHAT ARE BIBLIOGRAPHIC DATABASES?

Bibliographic databases tend to be focused on recording either a type of publication, such as journal articles, or all types of publication in a particular subject area. For example, the Cumulative Index to Nursing and Allied Health Literature (CINAHL) (EBSCO Health n.d.) database captures the nursing and allied health literature published in journals, books, reports, and dissertations (EBSCO Health n.d.). Brief details of a publication are captured in a database record. The volume of information within a database record varies by database and by the type of publication included. When encountering a database for the first time, it is important to read the descriptive information, FAQs, or Help section about the database as well as the searching options available. This information is important to ensure that searches are designed to elicit the most appropriate information from the database and to run efficiently. This

section describes the broad differences across databases collecting different publication types.

5.2.1 Synthesized Evidence

Before beginning a new research synthesis, it is wise to check that the question has not been published elsewhere recently. An increasing number of databases provide access to reviews and other evidence syntheses. Reviews can also be helpful for scoping the review synthesis question and providing examples of searches and other methods used by other researchers. Identifying whether a database of systematic reviews in a discipline exists should be a primary task in considering a new research synthesis.

Examples of collections of evidence syntheses include the following:

- The Cochrane Library contains reviews of health-care interventions (Cochrane Library n.d.).
- The Campbell Library has reviews of social work, criminology, and education (Campbell Collaboration n.d.).
- VetsRev has reviews of veterinary research (Centre for Evidence-Based Veterinary Medicine n.d.).
- The Environmental Evidence library offers access to completed and ongoing systematic reviews in environmental research.
- Epistemonikos collects published reviews and provides links to the eligible records in the reviews (Epistemonikos n.d.).

It is also important to try to identify ongoing systematic reviews. These may be recorded in the collections of evidence syntheses but may also be found in research registers. For example, in health care, the PROSPERO database is a registration route for systematic review protocols (National Institute for Health Research n.d.). Increasingly, journals such as *Systematic Reviews* publish review protocols, so searching bibliographic databases to identify protocols may also be productive.

5.2.2 Journal Literature

Bibliographic databases are the most efficient way to identify a set of potentially relevant studies that have been published as journal articles. Databases such as PsycINFO, Criminal Justice Abstracts, and ERIC (Institute of Education Sciences n.d.) have been designed to facilitate effective information retrieval through searches of the citation information, abstracts, and indexing terms of journal articles. The words in the title and abstract are known as free-text terms and are presented as provided by the authors of the articles. The indexing terms, which are derived from a controlled vocabulary or indexing language, are applied by the publisher of the database and are designed to provide a consistent term to describe a subject that might be described in different ways by different authors. Many electronic bibliographic databases include links to the full text of the article where available. Some database interfaces will also provide access from individual records to "related" articles, based on an analysis of similarities in the free-text terms or indexing terms.

Bibliographic databases of the journal literature can be broad ranging in their coverage, or more focused. For example, ERIC contains more than 1.5 million records of research publications in education; Criminal Justice Abstracts has more than five hundred thousand records indexing criminal justice research; and PsycINFO has more than four million records relating to psychology and the behavioral and social sciences. Each of these databases is indexed with its own indexing language: the ERIC Thesaurus, the Criminal Justice Abstracts Thesaurus, and the Thesaurus of Psychological Index Terms®. Scopus, in contrast, is a major cross-disciplinary database including records from journals in science, technology, medicine, social sciences, and the arts and humanities (SCOPUS n.d.). It does not apply its own indexing language to its records, so searching within such a database does not benefit from the assistance of additional terminology. Databases may also have a regional focus. For example, the LILACS database indexes medical journals from the Latin American and Caribbean region (Virtual Health Library n.d.).

Deciding which databases need to be searched for any topic under consideration for research synthesis needs to be informed by a knowledge of the unique content of the databases. Typically, in research synthesis we seek advice from research evidence, where possible, to guide our practice. CENTRAL (a database in the Cochrane Library), PubMed, and Embase (Elsevier Life Sciences. n.d.) are generally considered the most important sources for reports of clinical trials in health care (Lefebvre, Manheimer, and Glanville 2011). In veterinary medicine, a 2012 study identified CAB Abstracts as an essential source because it indexed more than 90 percent of the key veterinary journals identified by the authors (CABI 2016; Grindlay, Brennan, and Dean 2012). Research synthesis guidance documents can be valuable in determining the number of, and potentially most fruitful, databases to be searched.

That a database currently indexes a particular journal does not necessarily mean that it has always indexed that journal. Searchers should check the database coverage of key journals relevant to the research synthesis and search additional databases if coverage is partial. A database may not index the full content of a journal. The searcher should check the extent of indexing (cover to cover or selective) of key journals relevant to the research synthesis question and, if indexing is selective, should consider hand searching those journals (see table 5.1).

5.2.3 Books

Books are the best recorded publication format, but their importance to research synthesis varies by discipline. For example, books rarely feature in systematic reviews of healthcare since the main publication route for new research is journals. Researchers can identify books by searching library catalogs. Large national libraries, such as the U.S. Library of Congress, provide excellent access to books in their international multidisciplinary collections; books can also be retrieved through internet search engines. WorldCat provides access to the catalogs of more than ten thousand libraries worldwide (OCLC WorldCat n.d.). Some subject databases, such as PsycINFO and CINAHL, include books and book chapters, but coverage should be checked on a database by database basis.

Library catalog records do not typically contain abstracts, so the records offer limited information to search and do not typically provide detailed information on the individual chapters within books. The search interfaces to catalogs may be relatively simple relative to those for databases of journal articles. Therefore, as with most database searches, searches in library catalogs should be sensitive and use a range of synonyms to find records of potentially relevant books. These issues mean that searching catalogs can be time consuming and that searchers should allocate adequate resources for these searches.

5.2.4 Dissertations and Theses

Dissertations and theses are recorded and indexed in specialized information sources, such as ProQuest's Dissertations & Theses Global database (ProQuest n.d.). These databases are focused on recording this publication type and are designed to promote access to these potentially difficult to find documents by providing searchable records and options to order copies. In addition to large multinational databases, there are many national dissertation databases, such as the EThOS:UK E-Theses Online Service (British Library n.d.).

Some subject databases, such as PsycINFO and CINAHL, also index dissertations within their subject fields. WorldCat also contains records for dissertations and theses.

5.2.5 Conference Papers

Research findings are published in papers and posters delivered at congresses and conferences. Systematic reviews of the human health literature have concluded that 50 percent of trials reported in conference abstracts never reach full publication (Hopewell, Clarke, Stewart, et al. 2007). The picture in other disciplines may not yet have been reviewed, but primary studies indicate the possibility of similar issues (Snedeker, Totton, and Sargeant 2010). Therefore, to minimize publication bias, searchers should explore conference abstracts from which they may be able to trace subsequent publications (such as posters, PowerPoints, journal articles or reports), or contact the study authors for additional information.

The inclusion of conference abstracts in subject bibliographic databases is not universal. For example, Embase now includes large numbers of conference abstracts from medical conferences published in the journals that it indexes, whereas the National Library of Medicine does not for the journals it includes in PubMed.

Searchers can find conference abstracts via specific indexes such as the BIOSIS Citation Index and the Conference Proceedings Citation Index (Thomson Reuters 2014 n.d.a). However, even these databases are unlikely to retrieve all relevant studies presented at conferences. Searchers, therefore, should consider additional approaches such as hand searching or electronically searching conference proceedings that are made available online or in print. The growing trend for conference organizers to provide their abstracts on a conference website will improve current access to this type of publication. However, conference websites are often transitory and it cannot always be guaranteed that a conference website will exist a year hence. In addition, many conference websites can be accessed only by paying the conference attendance fee. Searchers should keep personal copies of any potentially interesting conference papers to ensure continued access.

Identifying conference papers is also valuable for research synthesis because it can identify recent research not yet indexed by bibliographic databases, as well as provide signals to journal articles that may have been included but were missed by the search strategy. A Cochrane review

concluded that a combination of hand searching conference proceedings and database searching could help fully identify relevant studies published in health journals (Hopewell, Clarke, Lefebvre, et al. 2007).

Choosing which conferences to search is likely to be informed by subject knowledge (or the advice of an information specialist) about the key conferences in the topic of the research synthesis. Once identified, the searcher should check whether the conferences are indexed in the journal databases to be searched: if they are, the number of conference websites that need to be individually searched can be minimized.

5.2.6 Ongoing Research

Research registers—databases that record ongoing research—are particularly useful for identifying ongoing or as yet unpublished studies, and thus are one way to reduce publication bias. Many research registers exist and most are publicly available online. They may be discipline specific, such as ClinicalTrials.gov, which records trials in health care. They may be country or region specific, such as the European Commission's CORDIS (European Commission n.d.). They may collect trials funded by particular organizations, such as the UK Research and Innovation Gateway to Research (2018). They may also be produced and maintained by manufacturers, pressure groups, and international organizations. No single database of research registers exists, however. Searchers should therefore allocate adequate resources to identify candidate registers and to search key international, national, subject, and funder registers relating to the topic of the research synthesis.

Many research register databases have very basic search interfaces and offer few of the sophisticated search options typical of large bibliographic databases, such as set combination and result downloading (Glanville et al. 2014). This means that searchers may have to use simple and repetitive searches. Also, searchers may find that the results are not formatted or downloadable and they may therefore need to cut and paste search results into nonbibliographic software such as Word or OneNote.

5.2.7 Citation Databases

Citation indexes, such as those published by Thomson Reuters (Science Citation Index Expanded, Social Sciences Citation Index, and Arts & Humanities Citation Index), or citation search services such as those provided by Google Scholar (Google n.d.), provide access to journal articles that cite other journal articles. Using selected key papers as the seeds, searchers can use these information sources to identify newer papers that may have cited the seed papers and hence may be reporting on a similar topic. Research in the human health literature for research synthesis has shown that citation searching can be a useful adjunct to database searching and hand searching (Greenhalgh and Peacock 2005; Linder et al. 2015); it may therefore be relevant in other disciplines as well. Citation analysis may also have potential to reduce the burden of searching bibliographic databases (Belter 2016). Different citation resources may yield different results for the same search question (Levay et al. 2015).

5.2.8 Grey Literature

The extensive range of grey literature, including research and technical reports, some of which is captured in databases such as the National Technical Information Service (NTIS), is described in chapter 6.

5.2.9 Other Searching Activities

As well as database searches, and depending on the synthesis topic and available resources, other research identification methods are options:

- Searchers should consider whether hand searching of key journals, databases, or conferences would be helpful. For health-care systematic reviews, research shows that searching the contents of a journal by looking at each article in sequence and making an assessment of the relevance of the article to the synthesis question may yield additional studies (Hopewell, Clarke, Lefebvre, et al. 2007). However, because this is a resource intensive activity, searchers should conduct an exploratory study when developing the research synthesis proposal, to assess whether this effort is merited.
- Searchers should check the references within relevant reviews and studies to identify any additional studies (Horsley, Dingwall, and Sampson 2011). This can provide reassurance that database searches have been adequate, but may also identify additional studies (Doree et al. 2018).
- Searchers should consider contacting experts when they have a list of known relevant studies to request recommendations of any additional relevant publications.

Searchers should ideally conduct citation searches on known relevant papers to identify later citations referring back to those key papers.

Searchers should consider searching databases for additional publications by named key authors;

Searchers should consider following any "related references" links offered by databases (Doree et al. 2018).

5.3 WHY SEARCH A RANGE OF BIBLIOGRAPHIC DATABASES?

Searchers need to be aware of two key challenges to identifying relevant studies for research synthesis that the literature search is seeking to minimize. First, searches are seeking to minimize reporting biases such as publication bias and, second, the search is trying to overcome the limitations inherent in database records and search interfaces. This section describes these challenges and the impact they have on literature searches.

5.3.1 The Challenge of Publication Bias

A number of reporting biases affect access to research findings (Song et al. 2010). A systematic review reports that "statistically significant, 'positive' results that indicate that an intervention works are more likely to be published, more likely to be published rapidly, more likely to be published in English, more likely to be published more than once, more likely to be published in high impact journals and, related to the last point, more likely to be cited by others" (Sterne, Egger, and Moher 2011, 298). These biases pose a significant challenge for searches seeking to retrieve all relevant studies for a research synthesis. As a result of publication bias, a proportion of research will not be published in peer-reviewed journals. Of the research that is published, a significant proportion will not be indexed in the major bibliographic databases (Sterne, Egger, and Moher 2011). This evidence relates to the human health field, but is likely to apply in other fields given that the impetus for publication is likely to be similar for most disciplines. Some evidence, for example, suggests that publication bias is also an issue in the field of food and feed safety (Glanville et al. 2014; O'Brien et al. 2006; Snedeker, Totton, and Sargeant 2010) and in veterinary and agricultural research (Nielen, Kruitwagen, and Beynen 2006; Berteaux et al. 2007).

Search strategies for research synthesis involving quantitative study designs should seek to minimize the effects of this bias by including searches beyond peer-reviewed journal literature, the grey literature (chapter 6), and other evidence identification approaches. Research synthesis of qualitative data may be conducted with less exhaustive approaches, but searchers should still review the range of resources to minimize bias by ensuring that enough types of information have been sampled (Booth 2016).

5.3.2 Limitations of Bibliographic Databases

The word limits for titles and abstracts in bibliographic databases and the variability in indexing terms mean that even when a study is recorded in a database, a specific search strategy may not find it (Lefebvre, Manheimer, and Glanville 2011; Kassai 2006; Snedeker, Totton, and Sargeant 2010; Rathbone et al. 2016; Kugley et al. 2016). Just because a record exists in a bibliographic database does not mean that it is easy to retrieve, even when using "appropriate" search strategies that search across indexing and free-text fields. Several explanations address failures to retrieve records. First, people (rather than machines) are still largely responsible for the selection and application of indexing terms to bibliographic database records, and as a result indexing is subjective and open to error. This means that research may not be indexed in the way that seems most obvious to us or in the way we anticipate. Second, database abstracts are brief (and in some cases nonexistent) and may not capture all aspects of a full document because of the limits of space. Third, authors may describe the same topic using different terminology in different records. Fourth, authors may not describe their methods or other elements of their research fully in the abstracts (Whiting et al. 2008; Booth 2016). For example, the abstract may not reflect all of the outcomes that the authors report in the full document, and thus searches that contain the outcome concept run the risk of missing relevant records. In health-care searches, outcomes rarely form part of the search precisely for this reason. Searchers rely on the record selection process rather than the search to check the desired outcomes are addressed in the publication. Similarly, authors may not report their methods clearly. These issues pose a significant challenge for a research synthesis search aiming to retrieve all relevant quantitative studies. Searchers need to be able to anticipate such issues and use methods to minimize their potential impact. These issues are also the reason that search strategy development is often an iterative process.

5.3.3 Importance of Extensive Searching

Searching multiple information sources, including a range of databases, increases the likelihood of retrieving relevant records (Avenell, Handoll, and Grant 2001; Grindlay, Brennan, and Dean 2012). It also reduces the impact of database indexing variability because a study not retrieved by a particular search strategy in one database may be retrieved in another (Lefebvre, Manheimer, and Glanville 2011). Searchers should consider the topic of their research synthesis when identifying bibliographic databases to search and identify databases that cover not only the specific topic, but also different publication types. In contexts such as agri-food-related health research, for example, these may include both medical and agricultural databases (Sargeant et al. 2006). Where possible, searchers should select databases that will also retrieve research published in languages other than English, as well as ongoing and recently completed research.

For research synthesis questions involving qualitative research, exhaustive searching may be less of a focus because researchers are trying to find a large enough sample of representative cases. In this context, the range of databases required needs to be assessed from a different perspective (Brunton, Stansfield, and Thomas 2012; Booth 2016) and the searcher needs to determine whether enough databases have been searched to ensure that a range of relevant issues will have been identified.

5.3.4 Bibliographic Database Selection

The research synthesis plan or protocol should include a specification of the search process. This will include a list of the information sources (including bibliographic databases) that will be searched and the search terms that will be used to search them. This section focuses on how to select the bibliographic databases.

Searchers select databases based on a combination of many factors related to the research question. Their objective is to ensure that enough databases are selected to provide extensive coverage of the topic of the research synthesis and the types of publications in which research evidence might have been published. The factors affecting selection decisions include the following:

Topic or disciplinary scope. In which disciplines is research on the topic being conducted and which bibliographic databases capture the research in those disciplines? For example, accuracy of eyewitness testimony in a court of law is a topic of interest to both psychologists studying memory, and legal professionals wishing to win their cases. At least two disciplinary databases, PsycINFO, which captures the psychological research literature, and the Index to Legal Periodicals Full Text, which captures legal research, might contain relevant literature. However, multidisciplinary databases such as Scopus could also yield relevant records. Ensuring adequate discipline coverage in the chosen databases is particularly important in interdisciplinary topics.

Access. Which databases are available (free of charge) via local institutions or organizations? What other access routes are available?

Usability. Can the results from database searches be easily downloaded in terms of both volume and formats that can be loaded into bibliographic management software?

Date. What period does the topic cover? If the topic has been a subject of research for decades, such as water fluoridation, then databases that cover long periods will be required in addition to those initiated more recently. On rare occasions, when the relevant literature extends back into the early part of the twentieth century, searchers may need to identify and search paper indexes.

Language and country. Research synthesis often seeks to minimize publication bias by accessing research in a range of languages and in regional and national databases. Research in health care shows that bias can arise from excluding trials reported in languages other than English, as researchers from non-English-speaking countries are more likely to publish trials with positive results in English-language journals and trials with nonsignificant results in non-English-language journals (Egger et al. 1997; Hopewell, McDonald, et al. 2007). If resources permit, searchers should search beyond the English-language literature to mitigate the impact of publication biases.

Publication type coverage. Databases that cover the various publication types described earlier should be identified.

Advice on identifying candidate databases should be sought from an information specialist. Most large academic libraries offer lists of databases and other reference sources by topic (such as the subject guides and indexes at the University of Minnesota Libraries). It may be worth searching more than one such online guide

because different libraries may subscribe to different resources. The *Gale Directory of Databases* (n.d.) lists available databases, for example.

Once selected, searchers will need to plan the order in which the databases should be searched. Searchers should search the databases with the richest information first because results from subsequent databases can be de-duplicated against those results. Searchers should prioritize the most specific subject databases because they are more likely to be indexed from the perspective of the topic and yield a higher proportion of relevant results. In contrast, multidisciplinary databases are likely to be indexed from a more general perspective and yield less focused results. Databases that include abstracts should probably be searched before those without abstracts so that the richest information is encountered first. In the absence of other factors, searchers should search the databases most likely to contribute the largest number of records first. They should also search databases that do not offer downloading options (and from which results will need to be cut and pasted) last, so that results can be matched against the previous search results to minimize the need for cutting and pasting records. The challenges of deciding when enough searches have been undertaken are discussed later.

5.4 PLANNING THE SEARCH STRATEGY

5.4.1 Context of the Research Question

Research synthesis is not a single concept and the search requirements may vary depending on the scope, timelines, and resources of the specific synthesis (Grant and Booth 2009). It is important that the searches are fit for purpose and designed with an awareness of what the synthesis is trying to achieve within allocated resources. A brief typology of the searches that might be undertaken for scoping reviews, rapid reviews, and full evidence syntheses are presented in table 5.2. Other typologies are available (Grant and Booth 2009).

The first stage of a research synthesis may be a scoping search to assess the size and makeup of the literature available for the evidence synthesis. This search is likely to be exploratory, high level, and much less exhaustive than the searches undertaken for the research synthesis proper. It is usually undertaken to develop the

Table 5.2 Typology of Evidence Synthesis Searches

Type	Characteristics
Scoping search	Undertaken to understand the scale and scope of the literature. Can be used to identify which databases, to understand how many records will be retrieved by the searches, and to identify key search terms. Review may inform the development of a research proposal and its budget. Undertaken in only a few databases and with a relatively focused search. May involve the assessment of selected results only. One key aspect may be the identification of earlier reviews on the topic of interest to aid scoping.
Rapid review searches	Undertaken for evidence syntheses that have to be undertaken in a short time frame. Involve searches of a few selected databases. Strategies may not be exhaustive and therefore are less likely to minimize publication biases. Should be reported in detail even though they may be briefer and less extensive than those for more extensive evidence syntheses. May be limited in pragmatic ways, such as by limiting by date, language, or study design. Will usually not involve additional methods to identify relevant records.
Main or full searches	Undertaken for an evidence synthesis that is aiming to be extensive or exhaustive to minimize publication biases. A sensitive approach is usually undertaken, involving searches of a number of databases using strategies with a range of relevant synonyms and as few concepts as possible. Will usually involve additional methods to identify relevant records, such as reference checking, hand searching, and citation searching. Should have as few limits as feasible. Will be reported in detail.

SOURCE: Author's compilation.

research synthesis proposal or protocol. It is likely to include a search for previous reviews of the topic. Published reviews can provide a range of helpful information, not only to inform the searches, but also to highlight other challenges that the review topic may carry with it. Scoping searches are vital for budgeting purposes to inform calculations on the cost of both the search process and the size of the literature to be processed for the synthesis.

The evidence synthesis will require a protocol or project plan, which will in turn include a description of the methods used to conduct the search. The searcher will need to gain an understanding of the purpose and scope of the research synthesis by reading the proposal and related documentation, as well as any key known relevant studies and related documents that have already been identified. The searcher should clarify the review question and any search issues arising with other team members. The searcher will also ascertain whether the synthesis is a rapid review or an extensive evidence synthesis, because this will determine several search decisions (see table 5.2).

Once the search topic and the purpose of the review is known, the searcher can begin to develop the detailed searches, taking into account the concepts that will feature in the search, the search terms that will capture the concepts, and the bibliographic databases that will be searched.

5.4.2 Identifying Concepts

The research question the research synthesis is seeking to answer is often broken down into its key concepts (Lefebvre, Manheimer, and Glanville 2011). These concepts are also used to develop the search strategy (de Vet et al. 2008; Lefebvre, Manheimer, and Glanville 2011; Booth 2016). Many conceptual models are available; the choice among them depends on the question being asked (see table 5.3). The important thing is that usually the final search strategies will not try to capture all of the

Table 5.3 Examples of Published Conceptual Breakdowns

Acronym	Concepts	Usage
PICO/PECO (European Food Safety Authority 2010; Lefebvre et al. 2011)	Population, intervention or exposure, comparator, outcomes	Reviews evaluating the effects of an intervention or exposure
PIT (de Vet et al. 2008; European Food Safety Authority 2010)	Population, index test, target condition	Reviews of test accuracy
PO (European Food Safety Authority 2010)	Population, outcome	Reviews that aim to answer descriptive questions: questions about prevalence, occurrence, consumption, and incidence
PICOT-D (Elias et al. 2015)	Population, intervention/exposure, comparator, outcomes + time + data	Reviews with measures of outcomes of interest, for example, blood glucose tests or hba1c levels
PICOCs (Petticrew and Roberts 2006)	Population, intervention or exposure, comparator, outcomes + context + study design	Reviews in the social sciences
ECLIPSE (Wildridge and Bell 2002)	Expectation, client group, location, impact, professionals, service	Reviews of service change
SPIDER (Cooke et al. 2012)	Sample, phenomenon of interest, design, evaluation, research type	Reviews of qualitative and mixed method studies
SPICE (Booth 2006)	Setting, perspective, intervention, comparison, evaluation	Reviews in which perspectives of the intervention need to be captured and the impact of the evaluation is important

SOURCE: Author's compilation.

concepts because search strategies with many concepts may be overprecise and lack sensitivity: a typical database record with only a title, abstract, and subject indexing may not contain all four or five of the desired concepts in a particular research question. Concepts not captured in the search strategy, however, will still be required at the record selection stage of research synthesis, when they are used to select records most likely to meet the eligibility criteria.

The concepts that should be included in the search strategy are determined by exploration, but certain key issues typically inform the choice:

- The most specific concept is often included in the search strategy because it is crucial to the research question and may yield the smallest number of records. For example, in a review of the diagnostic test accuracy of a new imaging technique to detect breast cancer, the imaging technique is likely to be the most specific concept relative to the disease being detected and the outcome of disease detection.
- The outcomes concept is often not captured in the search because the outcomes may be various, difficult to capture, or similar to the population concept. For example, in a review of the effectiveness of smoking cessation interventions, the outcomes can be described in many ways and may also involve search terms that have been used to capture the population, such as "smoking" or "tobacco." As noted earlier, outcomes addressed in a document may not all be listed in an abstract.
- The comparators (if these feature in the conceptual breakdown) are often not included in the search. For example, a review question on probiotic feed supplementation for the prevention of Salmonella infection in poultry may not include a concept for the comparators because these may be too various to describe or include "doing nothing" or "no supplementation," which can be difficult to capture in search terms. Comparators may be difficult to search for if they have not been explicitly described in the abstract: the idea of a comparison may be indicated by the wording, but the specific comparators may not be stated.
- Sometimes a concept to capture the study designs of interest may be included in the search strategy. Decisions on whether to introduce such a concept will hinge on how many study designs might need to be captured and whether the terminology used to describe the study designs is consistent within and across disciplines and databases. Where the study designs are few and consistently described, adding a study design concept to the search may be helpful but should be considered on a database by database basis. Strategies to find specific study designs may be available as published search filters.

Searchers should develop the strategy in a single database that has been identified as important to the research synthesis question. Later, they can adapt the search strategy to run in other databases, but this is usually only after a series of iterations to improve the search strategy. To illustrate the search development process, let us look at the example topic of memory accuracy of adult eyewitness testimony. The conceptual breakdown of this question might initially be considered to be adults, eyewitness testimony, and memory accuracy.

Increasingly, searchers may wish to make use of text mining (text analysis) tools, which can help assess the concepts available within a literature. Text analysis packages analyze the frequency of words and concepts in sets of records. This can reveal both frequently occurring words and phrases, which could be tested in strategies, and the presence of concepts within sets of records. Many packages provide helpful visual representations of the features of the records, which can help with both concept and term identification. It is possible to carry out a broad scoping search—such as "eyewitnesses AND memory"—in a database such as PsycINFO, download the results and then load the results into text visualization software such as VOSviewer, which displays the concepts within those records.

The searcher's next tasks are intertwined: to explore which terms will capture concepts and to decide which of the concepts will feature in the final strategy, and in what combination.

5.4.3 Identifying Search Terms

Once the searcher has identified potentially useful concepts, efforts usually focus on identifying as many relevant terms as possible (Centre for Reviews and Dissemination 2009; European Food Safety Authority 2010; Lefebvre, Manheimer, and Glanville 2011; Petticrew and Roberts 2006). Searchers should exploit search strategy techniques and database functionality within the available interfaces to achieve this, for example, through building a list of relevant and related search terms and by using truncation and proximity operators (Lefebvre, Manheimer, and Glanville 2011) (table 5.4). Inevitably,

Table 5.4 Database Interface Search Options

Option	Description	Examples
Truncation	Used to specify different word endings to a word root. This reduces the number of free-text terms that must be typed. Character will differ depending on the database interface. Some interfaces permit the specification of a maximum number of additional characters. A short word stem of three letters or less should be avoided because it is likely to retrieve too many irrelevant records. Should be used carefully because it may also retrieve words that are not required. Some database interfaces will permit left-hand truncation as well as right-hand truncation.	In the Web of Science interface, "lactobacill*" will identify records containing the terms lactobacillaceae, lactobacillus or lactobacilli. In the Ovid interface, "random$3" will find all word variants (within 3 letters) ranging from "random," through "randomly" to "randomise" and "randomize," but not "randomised" or "randomized." However, "random$" will retrieve all terms with the stem "random" no matter how many characters follow the stem.
Wildcards	Account for internal spelling variation and can be particularly useful when trying to capture differences in UK and US spelling.	In the Ovid interface, "randomi?ed" will identify records containing the word "randomized" as well as "randomised." In the EBSCO interface, the wildcard is ? or #. "ne?t" retrieves records containing "neat," "nest," or "next."
Phrases	Ensure that terms appear next to each other.	In the Ovid interface, "criminal adj justice" finds the exact phrase "criminal justice." In the EBSCO interface, exact phrases have to be surrounded by double quotes, for example, "criminal justice."
Proximity operators	Ensure that terms appear near to each other. Usually the maximum distance that the terms can be apart is specified. Offer better precision than the use of AND alone, since terms linked by AND may be widely separated within a record and semantically unrelated.	In the Ovid interface Breast adj3 (cancer* or neoplasm* or tumor* or tumour*) will search for "breast" within three words of any of the terms in the brackets. This search string would therefore identify a range of word groupings including "breast cancer," as well as "breast and colon neoplasms." In the Web of Science interface, the NEAR/x operator is available: Salmon near/4 virus.
Restricting search terms to specific fields	Ensures that the search terms are searched only in nominated fields such as the title or the author keywords.	In the PubMed interface, searches can be limited to the title by using the field code within square brackets: "criminal justice [ti]." In the Ovid interface, searches can be limited to the title field by using a suffix, for example, "Criminal.ti."
Explosion of subject headings	Many interfaces to databases with subject indexing offer the facility to "explode" subject index terms to automatically include any more specific terms in the search.	In the Ovid interface to PsycINFO, exploding "Short Term Memory" (exp short term memory/) would also retrieve records for "Iconic Memory."

SOURCE: Author's compilation.

conducting sensitive searches involves a trade-off in terms of reduced precision. This trade-off should be acknowledged as part of the research synthesis search process, and an appropriate balance should be sought within the context of the resources available.

The default search terms in any search strategy are likely to be those used to search the title and abstract of a record because at some point the search strategy will probably be used in databases that do not have subject indexing schemes. However, for databases that do offer indexing terms, the searcher will need to identify relevant indexing terms and include them in the search strategy. Database subject indexing schemes are often hierarchical, broader (more general) indexing terms having one or more narrower (or more specific) term or terms below them. Subject indexing schemes may also offer related terms that cross-reference related subject headings. Usually, the database interface provides search tools to assist with identifying subject indexing terms and terms around them in the hierarchy. For example, the "Map term to subject heading" function in the Ovid interface automatically identifies relevant index terms in an indexed database for any free-text terms entered.

Databases that offer subject indexing, such as PsycINFO, provide a way to search information sources using a controlled vocabulary. Terms from a controlled vocabulary, or thesaurus, are assigned to records, usually by a human indexer, to describe the content of the item. Subject indexing terms are valuable in increasing the sensitivity of a search because they provide a way of retrieving records whose authors may have used different words to describe the same concept. They can also be used to provide information in addition to that contained in the title and abstract. Indexing schemes are not interchangeable across databases; the indexing terms used in PsycINFO are likely to be different from those in another database. This is one reason a search strategy must be adapted before it can be run in another database. For example, MEDLINE does not have a Medical Subject Heading (MeSH) for "eye witness," "witness," or "testimony," but does have a MeSH for "Expert testimony," whereas the Criminal Justice Abstracts subject indexing scheme has several terms around the concept of "witness," including "Witness Credibility" and "Witnesses."

The initial development stage for a strategy to find records about the accuracy of eye-witness memory is shown in figure 5.1. Many of the subject indexing terms specific to PsycINFO relating to memory are specific terms and are below "Memory" in the subject indexing hierarchy. If all the more specific subject indexing terms below "Memory" are relevant to the question, the searcher can explode the term "Memory" and retrieve records with all the more specific indexing terms around different types of memory in one action. However, the detailed entries in the PsycINFO thesaurus suggest that some of the more specific memory terms would not be relevant, so a searcher would select only those headings that seem potentially helpful to reduce the number of irrelevant records retrieved. Explosion can be a time-saving feature, but should not be used automatically; searchers should explore its impact by looking at the thesaurus.

Searchers should be aware that controlled vocabularies used by databases have evolved over time, with index terms being added and removed to reflect developments

Database: PsycINFO <1806 to June Week 2 2016>
Search Strategy:

1	witnesses/ (4464)
2	witness$.ti,ab. (15890)
3	legal evidence/ (987)
4	legal testimony/ (1667)
5	or/1-4 (18470)
6	(memory or memories).ti,ab. (173295)
7	memory/ (55471)
8	false memory/ (1946)
9	explicit memory/ (1015)
10	long term memory/ (4118)
11	recall$.ti,ab. (49821)
12	memory decay/ (991)
13	retrospective memory/ (164)
14	short term memory/ (20914)
15	spatial memory/ (5024)
16	verbal memory/ (2591)
17	visual memory/ (2523)
18	or/6-17 (206518)
19	5 and 18 (2687)
20	"300".ag. (1512281)
21	(19 and "300").ag. (1163)

Key: / indicates a subject heading search
$ - truncation symbol, finds all terms with the stem
.ti,ab. searches for words in the title and abstract
Or/1-4 combines results in set lines 1, 2, 3, or 4
.ag. searches records with PsycINFO specific age group codings. And Boolean operator to identify records containing results in both sets.

Figure 5.1 Preliminary PsycINFO Search
SOURCE: Author's tabulation.

in the discipline covered by the indexing scheme. This can affect the design of the search strategy. For example, few MeSH index terms related to study design were available pre-1990. Therefore, the index terms must be supplemented with relevant free-text terms (searching the title and abstract) to retrieve this older material (Lefebvre, Manheimer, and Glanville 2011).

Indexing terms alone are not enough when searching for studies for a research synthesis; in addition, the accuracy of indexing can be affected by an author who reports study methods and objectives poorly, or by an indexer who fails to notice the methods reported in the publication. Indeed, some databases include records without indexing as well as records with indexing, so that a search limited to subject indexed records will miss those that have no indexing. Subject indexing can also fail to capture the topic of interest at the granularity required: papers on male breast cancer, for example, may be indexed under a heading that includes both male and female breast cancer. Care should also be taken when searching in topic areas that do not use the precisely defined vocabularies found in disciplines like medicine and veterinary science. Where subject indexing is inadequate, for whatever reason, searchers will be aware that they are relying much more on the free-text terms in the strategy, and that using the subject indexing terms may impact on precision.

Subheadings are used by some subject indexing schemes such as MeSH for MEDLINE. Subheadings are added to a subject indexing term to focus it. For example, in Ovid MEDLINE, the search construction "Probiotics/ae" will restrict the probiotics subject indexing term to those records where the indexer has noted that the adverse effects of probiotics are addressed. Where subheadings are offered, they can also be searched on their own and unattached to specific subject headings; in this context they are called floating subheadings. In Ovid MEDLINE, the search construction "Probiotics/ and ae.fs." will find records indexed with the term "Probiotics" and have adverse effects applied as a subheading to any of the indexing terms in the record. This is less precise than "Probiotics/ae" but may increase the sensitivity of the search. As with indexing terms, subheadings will have been applied to bibliographic records by a human indexer and are therefore not infallible. For this reason, it is suggested that, when constructing a search strategy for research synthesis, subheadings should not be used as the sole approach to searching but instead as an additional search approach.

Searchers can identify search terms in a number of ways: from key relevant papers provided by the review team (a technique often labeled as pearl growing) or identified from records retrieved by simple scoping searches. These records would be assessed for relevance by eye or by using text-mining software (Hausner et al. 2012, 2015; Paynter et al. 2016). Text-mining techniques can process large volumes of records rapidly and provide lists of the terms and phrases that occur more or less frequently within the records. Search terms can also be identified from the strategies included in published research syntheses, from experts on the research team, and by consulting online thesauri, dictionaries, and webpages.

Given the example topic of memory accuracy of adult eyewitness testimony, we might try some scoping searches within a relevant database such as PsycINFO. Typing in a search phrase such as "eye witness" into PsycINFO (Ovid interface) and using the "Map term to subject heading" yields the following candidate subject headings (assigned by indexers): witnesses, memory, suggestibility, and recall (learning). Subject indexing terms can also be obtained by looking at the records of key known studies, identifying the subject index terms, and then looking them up in the PsycINFO thesaurus within the Ovid interface.

Exploring the indexing term "Witnesses" shows related subject indexing that might also be helpful to add to the term lists to be tested in exploratory searches: legal evidence and legal testimony.

Searching on "memory" in PsycINFO brings up many potentially relevant headings, including the following:

- false memory,
- explicit memory,
- long term memory,
- memory,
- memory decay,
- retrospective memory,
- short term memory,
- spatial memory,
- verbal memory, and
- visual memory.

Each subject indexing term will have a definition or scope note, which may provide context and information to help with deciding whether it is a relevant indexing term to use in the strategy. Information will also indicate when the term was introduced into the database; to identify older records, additional search terms should be considered. The listings show that "eye witness" is not

available as a subject index term but that other terms are available.

As well as terms that capture the topic exactly, such as "memory" and the synonyms for those terms, such as "recall," related terms and broader or narrower terms can be helpful to increase the sensitivity of the search and to take account of author variability and records with limited searchable text. This is why "legal" and "forensic" might be useful additions. The following factors are also important to consider and take into account:

- differences in US and UK English spelling and national terminologies, for example, "behavior" as well as "behaviour";
- abbreviated and unabbreviated terms, as well as acronyms;
- differences in word ending (singular and plural, but also past and present tense, active and passive verb forms) (for example, in figure 5.1, "witness" is truncated using the Ovid $ option to ensure that "witness" and "witnesses" are retrieved);
- both the generic and branded names of products, such as pharmaceuticals, pesticides, and chemicals;
- scientific and common names of organisms; and
- changes in vocabulary over time (for example, third world country, developing country, low-income country).

When exploring the terms within a strategy, the use of facilities such as truncation and proximity operators offered by the database interface will also come into play. Typical search interface options are described in table 5.4 and some are shown in figure 5.1.

Once the searcher has compiled lists of candidate terms to reflect the concepts, ways to combine the terms and concepts together are explored.

5.4.4 Combinations: Using Boolean Operators

In many database interfaces, search terms, and concepts are combined together using Boolean operators (AND, OR, NOT).

The OR operator will find records containing one or more of the search terms; using OR makes the search results larger. It should be used to accumulate search terms for the same concept. For example, in figure 5.1, the terms related to the concept of "witnesses" are gathered in set 5 by combining sets 1 to 4 using OR. The "memory" terms are gathered together in a large set by combining sets 6 to 17 together using OR.

The AND operator will find records containing all of the concepts in the combination; using AND makes the search narrower or more focused. It should be used to join two (or more) concepts together. In figure 5.1, the two concepts "witnesses" and "memory" are combined together in set 19 using AND. The result of the AND combination is a smaller number of records than either "witnesses" or "memory" because the result set has to contain records that mention both concepts.

The NOT operator is used to exclude records from the search (Lefebvre, Manheimer, and Glanville 2011). However, it should generally be avoided because it can have a significant impact on the sensitivity of the search by inadvertently removing relevant records. For example, searching for "adults NOT children" would remove not only records that are just about children, but also any record that was about both adults and children, simply because it mentioned children (figure 5.2).

Exploring the records retrieved by scoping searches about eyewitness memory shows that a range of issues are investigated and might be potentially relevant. The exploration suggests that a simple three-concept search risks missing relevant studies. In the light of this, a series of searches might be considered, which seek to compensate for the fact that the literature is not captured neatly by one conceptual breakdown, and it is, in fact, multifaceted. An example of different conceptual breakdowns and a series of combinations, whose results are all gathered into a final result set, is shown in figure 5.3.

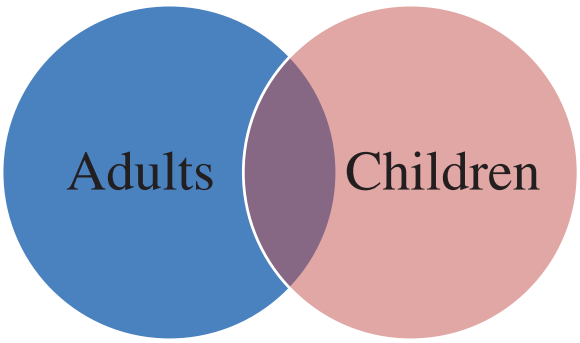

Figure 5.2. The Negative Impact of Searching for Adults NOT Children

SOURCE: Author's tabulation.
NOTE: Only the records in the darker shade will be retrieved.

1. Eyewitness concept AND testimony concept
2. Testimony concept AND memory concept
3. Eyewitness concept AND memory concept
4. Memory concept AND accuracy concept AND legal settings concept
5. 1 OR 2 OR 3 OR 4

Figure 5.3 Second Stage PsycINFO Search: Developing the Conceptual Approach

SOURCE: Author's tabulation.

The use of Boolean operators can be quite stringent and produces binary results: records are either retrieved or not retrieved because of the presence or absence of terms. The Boolean operators are not fuzzy and cannot reflect probabilities that a record may be relevant. As text mining becomes more widely available, we may see more occasions where, with complex multidimensional topics such as the eyewitness memory example, text mining is used to interrogate large sets of records (achieved, perhaps, through sensitive searches). Text-mining techniques can facilitate the design of queries to select records according to the probabilities that they are relevant and to present records in potential order of relevance. Researchers can then review records in order of possible relevance, decide at what point the relevant records have been exhausted, and possibly stop screening (O'Mara-Eves et al. 2015; Shemilt et al. 2014; Paynter et al. 2016).

Boolean logic is also unable to take account of the true focus of a document. For example, searches may retrieve records about "eye witnesses" when the only mention of the "eye witnesses" is in the final concluding sentence, which might be suggesting, perhaps, that "eye witness testimony should be sought." The reader understands that this is a recommendation and is not likely to reflect the main focus of the record. However, the simple presence of relevant terms means that Boolean searches will retrieve this record. Boolean logic is also unable to search using the meaning of words within records. Developments in text mining may again prove useful in improving precision in record retrieval because text-mining techniques can help with semantic analysis of records (O'Mara-Eves et al. 2015).

However, for most strategies the searcher will develop for bibliographic database searching, development still usually involves identifying concepts and combining them with Boolean operators. Typically, the most specific concept is developed and tested first. If search results are relatively small, or the synthesis team are willing to screen many records, a single concept may be adequate. However, the volume of results returned by the search may lead the research team to decide that additional concepts need to be added to the search to keep the number of retrieved records manageable. The searcher should explore the impact of adding the next most specific concept (using the AND operator) within the key database. The searcher will assess the impact of adding a second concept in terms of what proportion of the records that have been excluded were relevant. If a number of relevant records are lost by adding the second most specific concept, those records need to be explored to identify why they were missed. This might lead the searcher to the identification of additional terms to add to one or more concepts; it might also lead to the decision to abandon the second concept and to try another concept. Alternatively, it might lead to the conclusion that several search combinations will be necessary (as in figure 5.3). This type of multistranded or multifaceted approach is often seen in reviews of diagnostic test accuracy studies and complex topics (de Vet et al. 2008). This exploration continues until the strategy seems to be capturing as many relevant records as possible and not excluding large numbers of relevant records.

Searches for reviews of qualitative evidence, as noted earlier, may be developed differently from those for reviews of quantitative evidence. Rather than trying to develop a single search strategy, searching for qualitative studies may involve a series of searches looking at different aspects of the research question. New topics may be introduced to the search as the researchers read publications that suggest new avenues that should be explored. The series of searches is undertaken using a model of purposeful sampling (Booth 2016).

5.4.5 Focusing Searches

A range of methods to focus a search strategy may be considered; these are often expressed as limits. Some databases and database interfaces may offer built-in options to limit by one or more features. Database interface limits should be used with care because they may have unexpected effects; they also make the search reliant on the quality and consistency (and availability) of subject indexing within records. All limits should be justified within the research synthesis protocol (see chapter 21) to ensure that bias is minimized.

5.4.5.1 Population Limits Some research synthesis questions may contain concepts relating to an organism or population group, such as female adults, or domestic pets,

or bacteria. Databases may offer features to focus searches to those populations through indexing on issues such as age, sex, or species. The value and impact of using these options should be assessed carefully. The available searching and limiting options will vary from database to database. Developing searches to capture these issues can be challenging and the ISSG Search Filter Resource includes example filters that provide indications of how such searches could be developed (Glanville, Lefebvre, and Wright 2008).

In the example search, combining two concepts with AND to focus on memory and witnesses might then lead us to consider whether to add in a further concept for adults, since many of the papers retrieved would be about children as witnesses. Adding in a third concept, making use of PsycINFO's age group coding (ag field), shows that the number of records retrieved is reduced (figure 5.1). However, the records that are no longer retrieved by adding the age concept would need to be explored to ensure that focusing on adults was not eliminating relevant records.

5.4.5.2 Date and Language Limits If date or language limits have been agreed at the protocol development stage, these may be added to the search using a limiting option offered by the database interface or by restricting the results using a date or language field, where those are available. Date limits should only be considered if the context has changed since a specific date, for example, with the introduction of a new law, or standard, or technology (Lefebvre, Manheimer, and Glanville 2011). Language limits may have been considered when the protocol was developed, if resources do not permit assessment of studies in all languages.

5.4.5.3 Geographic Limits In cases in which the geographic context is crucial to the synthesis question, the search may include a concept to focus on geographical areas. Achieving such a focus is as challenging as devising any other concept. The concept will make use of information in the title, abstract, and subject indexing, as well as information in other fields such as the author affiliations and any other country coding that specific databases have applied to records. The ISSG Search Filter Resource contains some geographic filters that provide an indication of how geographic searches can be developed.

Sometimes it may be more efficient to remove geographic areas that are not relevant than to try to retrieve records by focusing on the relevant area. For example, if European countries are the focus, it might be safest to exclude studies about South America, Australasia, and the United States than to try to create a strategy sensitive enough to find all the European studies.

5.4.5.4 Publication Type or Format Limits Because studies are not always published in journal articles, it is often inappropriate to restrict the search in terms of excluding publication formats such as books or conference abstracts. The searcher will need to explore the impact of any such restrictions to ensure that relevant records would not be missed.

Sometimes, particular parts of journals—such as editorials, news stories, or comments—are viewed as unlikely to report relevant information for a synthesis, and are explicitly excluded from the search strategy. However, such exclusions should be made only after careful thought and investigation. For example, excluding letters may result in the loss of additional information where an author has reported an earlier study, a correction, or new information about a study that has been published elsewhere (Lefebvre, Manheimer, and Glanville 2011). Similarly, retractions of research are often first recorded in papers coded as "Comments" in MEDLINE, so excluding such papers (a common practice in searching) risks missing important relevant information (Wright and McDaid 2011).

5.4.5.5 Search Filters Search filters are tested, and ideally validated, search strategies designed to retrieve specific types of study or topic, such as a specific population, from a named database (Jenkins 2004; Wilczynski et al. 2016; Wood and Arber 2016a). They usually consist of indexing and free-text terms that describe the study design or topic of interest. The filter is added to the search strategy to provide an additional concept, for example, restricting the results of the search to a required study design or topic of interest. The ISSG Search Filter Resource is a free website that collates published and unpublished search filters grouped by study design and focus.

The value and availability of search filters for searches outside the health-care field is not well researched. In health care, published search filters have been developed primarily for use in large biomedical bibliographic databases such as MEDLINE. Little research has been undertaken into filters designed for use in the wider range of information sources required by reviewers searching, for example, the psychology, criminology, veterinary, or agricultural literature. Moreover, whereas research has suggested that filters to retrieve RCTs in MEDLINE and Embase are reliable, evidence for the sensitivity and recall of filters for the wide range of other study designs is limited in health care and even more scarce in other disciplines (Booth 2016; Petticrew and Roberts 2006).

Even when filters are available, they should be used with caution. Before incorporating any search filter into a search strategy, the searcher should assess the reliability of the filter's reported performance by exploring the methods used to create it. If a filter is several years old, its current effectiveness should be considered given the frequent changes in interfaces and indexing terms that affect databases (Lefebvre, Manheimer, and Glanville 2011). Guidance on critically appraising search filters is available (Bak et al. 2009; Glanville et al. 2008; Jenkins 2004).

We are likely to see the development of alternative approaches to applying search filters in bibliographic databases. Text-mining approaches allow us to conduct highly sensitive searches across a range of databases, to load results into the text mining applications, and then to interrogate the results either using machine-learning approaches or by developing relevance rules (O'Mara-Eves et al. 2015; Paynter et al. 2016). Machine learning involves "training" text-mining software to distinguish relevant records from irrelevant records. For complex synthesis questions that might benefit from text mining, it could be helpful to consult a text-mining expert early in the project to explore options for using text mining to aid the searching and record selection processes.

5.4.6 When to Stop Searching

Developing a search is an iterative and exploratory process. The searcher needs to explore trade-offs between search terms and assess their overall impact on the sensitivity and precision of the search. It is often difficult to decide in a scientific or objective way when a search is complete and search strategy development can stop. Searchers typically develop stopping decisions through the experience of developing many strategies. However, suggestions for stopping rules have been made around the retrieval of new records: we might stop the development process if adding in a series of new terms to a database yields no new relevant records or precision falls below a certain point (Chilcott et al. 2003). Stopping might also be appropriate when the removal of terms or concepts results in losing relevant records. Although many methods have been described to assist in deciding when to stop developing the search, few formal evaluations of the approaches have been undertaken (Booth 2010; Wood and Arber 2016b).

At a basic level, the searcher needs to investigate that a strategy is performing adequately. One simple test involves checking whether the search is finding the publications recommended as key publications or included in other similar reviews. However, it is not enough for the strategy to find only those records because it might be a sign that the strategy is biased to known studies and other relevant records might be being missed. The use of citation searches and reference checking are also useful checks of strategy performance. If those additional methods are finding documents that the searches have already retrieved, but that the team did not necessarily know about in advance, the results are one sign that the strategy is performing adequately. The searcher can also use the PRESS checklist to assess the quality of the strategy (McGowan et al. 2016). If some of the PRESS dimensions seem to be missing without adequate explanation, or arouse concerns, the searcher could conclude that the search may not yet be complete.

Other statistical techniques can be used to assess performance, such as capture recapture (Spoor et al. 1996) and the relative recall technique (Sampson and McGowan 2011; Sampson et al. 2006). These techniques may be most useful at the end of the search process because they rely on the achievement of several searches to make judgments about the overall performance of strategies. Capture recapture needs a set of hand searched or similar results to compare with a database search to estimate the number of missed studies. Relative recall requires a range of searches to have been conducted so that the relevant studies have been built up by a set of sensitive searches. The performance of the individual searches can then be assessed in each database by determining how many of the studies included in the research synthesis, indexed within a database, can be found by the database search used in the synthesis. If a search in a database did not perform well and missed many known relevant studies, that search strategy is likely to have been suboptimal given that it missed studies. If the search strategy found most of the studies that were available to be found in the database, then it was likely to have been a sensitive strategy. Assessments of precision could also be made, but these mostly inform future search approaches because they cannot affect the searches and record assessment already undertaken.

In research synthesis involving qualitative evidence, searching is often more organic and intertwined with the subsequent analysis of the identified research, such that the searching stops when new information ceases to be identified (Booth 2016). The reasons for stopping searching in this case need to be documented; methodologists have suggested that explanations or justifications for stopping may center on saturation (Booth 2016).

When developing search strategies, searchers will often find it helpful to make use of database interface facilities such as saved searches so that the search can be saved within the database interface and edited easily. This is more efficient than retyping the search regularly and minimizes retyping errors.

5.4.7 Search Strategy Peer Review

Research synthesis methods encourage the use of double independent reviewing for many tasks. Ideally, search strategies should be peer reviewed by an independent information specialist. However, this type of support may not always be available locally. To support opportunities for peer review, informal reciprocal networks are being developed. In the health-care domain, PRESSForum offers independent peer review of search strategies (PRESSForum n.d.). Information sharing discussion lists such as *expert-searching@pss.mlanet.org* can be helpful resources to ask questions about searching problems.

Published guidance in the form of a checklist, Peer Review of Electronic Search Strategies (PRESS), is also available to assist with detailed and consistent peer review and self-assessment of strategies (McGowan et al. 2016). The checklist is not weighted, so the relative importance of the dimensions of the checklist has to be determined on a case by case basis. The PRESS checklist can be used by the searcher, the peer reviewer, or by a journal referee to assess the quality of a search strategy. Other critical appraisal tools for search strategies are available (European Food Safety Authority 2015).

5.4.8 Main or Primary Searches

Once the searcher has finalized the search strategy in the key development database, and ideally had the strategy peer reviewed, the searches can be translated carefully to run in the other databases listed in the research synthesis protocol. In the example search, the PsycINFO strategy would be adapted to run in Criminal Justice Abstracts in the EBSCO interface by making the following changes:

- select the relevant subject indexing terms for Criminal Justice Abstracts;
- investigate whether it is possible to limit to adults in Criminal Justice Abstracts;
- convert the search syntax, such as truncation symbols, set combination, and field limits, from the Ovid interface to those required by the EBSCO interface.

If the strategy is to be run in a database without a subject indexing scheme, such as Science Citation Index (Thomson Reuters n.d.b), the searcher would remove the subject headings and would need to rethink or possibly remove the adult concept.

The searcher will also undertake any of the other research identification methods agreed to in the synthesis protocol. These might include grey literature searches (chapter 6), hand searching, reference checking, contacting experts, citation searches on key papers, named author searching, and following any "related references" links offered by some databases. It is probably most efficient for the searcher to do these searches after the results of the main database searches have been loaded into bibliographic software, so that only new records need to be added to the software. Some of these additional searches may yield results that can be easily uploaded. Others may result in the need to cut and paste references.

For syntheses of qualitative data, the searches may not be undertaken as a single block after which the search results are analyzed. Instead, searches may be more exploratory, organic, and sustained over a longer period (Booth 2016). A search might yield results that might suggest new topics, which then involve new searches. This process might continue until no further new topics are identified. It is likely to require a more extended series of dialogues between the searcher and the rest of the research team than a synthesis of quantitative evidence.

5.4.9 Managing the Search Results

The searcher will need to manage the records retrieved by the searches carefully so that none are misplaced or ignored during the rest of the research synthesis process. Bibliographic software can make the storage, de-duplication, and management of references retrieved from database searches more efficient. Bibliographic software includes packages such as EndNote, Reference Manager, RefWorks, Zotero, or Mendeley. A table listing a range of bibliographic software is provided in Wikipedia (Wikipedia 2018a). Many academic organizations provide bibliographic software to students and staff.

At the end of each search, the searcher should download the results of the search, if possible, as a structured (tagged) file that can be loaded into bibliographic software. A common structured format that permits loading into many bibliographic software packages is the RIS format (Wikipedia 2018b). Once loaded into a package, the research team can manage the records through the

various stages of the research synthesis process. Bibliographic software typically offers tools to rapidly de-duplicate records obtained from several databases. It can also index records on several fields, including fields that users can define and use to store process information, such as the information source from which records were downloaded, the date of the search, whether the document is eligible for the synthesis, notes on reasons for inclusion or exclusion, document ordering information (date, source, format, cost), and document storage information. Bibliographic software may also provide the option to create virtual groups of publications and to identify and link the full text of documents to the record. Bibliographic software can also link into word processing packages to generate the references for the final synthesis report if desired; references can be quickly formatted to meet different citation styles.

5.4.10 Record Selection

Record selection is described in detail in most research synthesis guidance with many practical tips (Higgins and Green 2011; Collaboration for Environmental Evidence 2013; Kugley et al. 2015; Centre for Reviews and Dissemination 2009; European Food Safety Authority 2010; Petticrew and Roberts 2006; Eden et al. 2011; Joanna Briggs Institute 2014). Record selection may be achieved within bibliographic software; records may also be exported to other software for selection, such as Excel, DistillerSR, or Covidence. Tools developed specifically for record selection can be identified from the SR Toolbox (Marshall n.d.). As text-mining techniques become more widely used, machine-learning tools may become a more common part of the record selection process (O'Mara-Eves et al. 2015). Best practice, advocated in many guidelines, is for record selection to be undertaken by two researchers independently and for a third researcher to be involved in cases of disagreement.

Record selection should be guided by the research synthesis protocol. Often researchers prepare a checklist, based on the protocol, to assist with deciding whether a record is to be included or excluded. Record selection may be a two- or three-stage process. Sometimes, stage one involves removing obviously irrelevant records rapidly. In our example, studies that are clearly only about children as eyewitnesses might be quickly removed as a stage 1 exercise. This permits the more relevant studies to be seen more clearly. The second stage is usually an assessment of how far a record meets the synthesis eligibility criteria based on information contained within the title and abstract of the record. If a record is clearly relevant, it is retained, but also if its relevance remains unclear, it is retained because the decision requires more information from the full document. If a record is clearly irrelevant, it is rejected. Investigators should assign a rejection reason if possible so that records can be revisited if necessary. Disagreement or a lack of clarity about the eligibility criteria may indicate a need for a team discussion. Disagreements may be adjudicated, perhaps by a senior investigator, and sometimes adjudication may involve a change to the eligibility criteria and an attendant change to the protocol.

After the research team has completed the selection process, it will seek to obtain full-text copies of the documents to assess whether the documents are truly relevant to the synthesis question. Access to full-text documents may be via the internet, library subscription services, or via interlibrary loan services. Sometimes the document's author will be contacted for a copy of a document. It is important to be vigilant for additional information relating to a publication. For example, a conference paper may have an abstract but also a PowerPoint presentation available. Journal articles may have supplementary information in files on the journal website that are separate from the paper itself. This information needs to be obtained as well to ensure that the research data are as complete as possible.

Research team members then read the full documents and consider whether the documents meet the eligibility criteria for the synthesis. The reasons for rejecting a document are recorded, and usually rejected documents are listed in an excluded studies table in the final report. Records relevant to the research question pass to the data extraction stage of the synthesis. Again, if research team members disagree about the eligibility of a document or are unclear about the eligibility criteria, they may need to discuss queries as a team. As in the abstract screening stage, the disagreement may be adjudicated, and sometimes the adjudication may involve a change to the eligibility criteria and an attendant change to the protocol.

As part of record selection, researchers also look for duplicate publications about the same study. Duplicate publications may be genuine duplicates, such as a single study reported in different journals, or may be publications reporting different aspects of the same study in different journals. Often, conference papers about different aspects of a study are published in addition to the full study report itself. Sometimes, different elements or analyses of a study may be published as separate papers.

Grouping publications by study is important to avoid double counting within the research synthesis.

5.5 RECORDING AND REPORTING THE SEARCH

Research synthesis methods encourage a scientific approach, and detailed reporting of the methods used to create the synthesis are an important element in demonstrating its scientific rigor (Kugley et al. 2016; Lefebvre, Manheimer, and Glanville 2011; 2013; European Food Safety Authority 2010; Centre for Reviews and Dissemination 2009; Petticrew and Roberts 2006; Eden et al. 2011; Joanna Briggs Institute 2014). The searcher should record the search process in adequate detail and report it in adequate detail in the final publication (Moher et al. 2009).

5.5.1 Recording the Search Process

The searcher should document search methods as the search progresses, to enable the search to be reported accurately. Elements of the PRESS checklist provide insights into some of the critical features of reporting the search:

> The search strategy should match the research question, the search concepts should be clear and there should not be too many or too few concepts. These elements of the checklist suggest that the process for arriving at the strategy should be recorded in case it needs to be explained at a later point. Specifically, the PRESS checklist expects an explanation of complex or unconventional strategies.
>
> The range of issues around the selection of subject headings in the PRESS checklist suggest that it is important for the searcher to record why subject headings have been chosen, and to defend the level of subject headings chosen and whether explosion has been employed.
>
> The PRESS checklist includes a number of questions around the range of text words identified and used in the strategy, which suggests that the searcher should explain the choice of text words.
>
> The PRESS checklist requires the searcher to provide appropriate limits and filters and to justify their choice. (McGowan et al. 2016)

Guidelines provide advice on best practice in recording search details (Kugley et al. 2016; Lefebvre, Manheimer, and Glanville 2011). The EFSA guidance for those carrying out systematic reviews in food and feed safety states that the following aspects should be recorded for each search:

> the name of the database;
>
> the date of the search for each database and the date range searched; and
>
> the full search strategy (all terms and set combinations) and the number of records retrieved (this information should be copied and pasted for all databases where possible; retyping searches should be avoided as this may introduce errors). (European Food Safety Authority 2010)

The searcher should keep notes of key decisions that may affect the synthesis' findings in a narrative format, such as the effects of selecting specific search headings, limiting the search in a particular way, or adding a search filter.

The searcher will also need to record the grey literature searches undertaken (chapter 6) and any other research identification which may have been undertaken, such as hand searching, reference checking, and contacting experts. Some of these searches may be less straightforward to record and searchers may wish to explore the use of notebook software such as OneNote or EverNote. OneNote and EverNote provide options to manage mixed-media records such as notes, screenshots, cuttings, and links to downloaded files.

5.5.2 Reporting the Search

Wherever the research synthesis is reported, the search is one indicator of the capacity of the synthesis to have captured as much relevant research as possible to answer the synthesis question or, in the case of reviews of qualitative evidence, as rich a set of information as possible. Searchers should report the search strategy in enough detail to facilitate an assessment of the search's quality in respect to the objectives of the research synthesis. Given evidence that the reporting of searches is less than optimal, reporting this element of the synthesis should be given appropriate attention. Many organizations provide guidance about the level of detail that should be included in search reports. For example, the PRISMA guidance on reporting systematic reviews and meta-analyses suggests "that all information sources should be described (databases with dates of coverage) and date last searched; [and] that the full electronic search strategy for at least one database should be presented, including any limits used, such that it could be repeated" (Moher et al. 2009).

This seems to be a minimum requirement, many guidelines suggesting that all strategies and study identification methods should be reported. Many syntheses include this level of detail, either in a report appendix or as a supplementary file to a peer-reviewed paper. The fullest possible reporting will make future updating much easier.

5.6 SUMMARY

Research synthesis involves identifying relevant information from one or more information sources, including bibliographic databases, using search strategies. The search and record selection process can be challenging, requiring a knowledge of query structuring, database content and structure, database interface variation, and software tools for record management and selection. Searchers will find that searches benefit from careful planning and adequate time to develop and complete strategies. Investing time in exploratory or scoping searches will assist in identifying concepts, terms, and search approaches used by previous researchers. Searchers should bear in mind the need for detailed recording and reporting of searches to inform research synthesis and to ensure that readers can assess the rigor and appropriateness of the search that underpins the synthesis.

5.7 REFERENCES

American Psychological Association. 2016. "PsycINFO." Last modified September 28, 2016. Accessed September 28, 2016. http://www.apa.org/pubs/databases/psycinfo.

Avenell, Alison, Helen Handoll, and Adrian Grant. 2001. "Lessons for Search Strategies from a Systematic Review, in the Cochrane Library, of Nutritional Supplementation Trials in Patients After Hip Fracture." *American Journal of Clinical Nutrition* 73(3): 505–10.

Bak, Greg, Monika Mierzwinski-Urban, Hayley Fitzsimmons, and Michelle Maden. 2009. "A Pragmatic Critical Appraisal Instrument for Search Filters: Introducing the CADTH CAI." *Health Information and Libraries Journal* 26(3): 211–19.

Belter, Christopher W. 2016. "Citation Analysis as a Literature Search Method for Systematic Reviews." *Journal of the Association for Information Science and Technology* 67(11): 2766–77. DOI: 10.1002/asi.23605.

Berteaux, Dominique, Brandee Diner, Cyril Dreyfus, Madrion Éblé, Isabelle Lessard, and Ilya Klvana. 2007. "Heavy Browsing by a Mammalian Herbivore Does Not Affect Fluctuating Asymmetry of Its Food Plants." *Écoscience* 14(2): 188–94.

Booth, Andrew. 2006. "Clear and Present Questions: Formulating Questions for Evidence Based Practice." *Library Hi Tech* 24(3): 355–68.

———. 2010. "How Much Searching Is Enough? Comprehensive Versus Optimal Retrieval for Technology Assessments." *International Journal of Technology Assessment in Health Care* 26(4): 431–35.

———. 2016. "Searching for Qualitative Research for Inclusion in Systematic Reviews: A Structured Methodological Review." *Systematic Reviews* 5(1): 74.

British Library. n.d. "EThOS: UK E-Theses Online Service." Accessed November 27, 2018. http://www.bl.uk/reshelp/findhelprestype/theses/ethos/index.html.

Brunton, Ginny, Claire Stansfield, and James Thomas. 2012. "Finding Relevant Studies." In *An Introduction to Systematic Reviews,* edited by David Gough, Sandy Oliver, and James Thomas. Farmington Hills, Mich.: Sage.

CABI. 2016. "CAB Abstracts." Last modified October 2018. Accessed November 27, 2018. http://www.cabi.org/publishing-products/online-information-resources/cab-abstracts.

Campbell Collaboration. n.d. "The Campbell Collaboration Online Library." Accessed November 27, 2018. http://www.campbellcollaboration.org/library.html.

Centre for Evidence-Based Veterinary Medicine. n.d. "VetSRev—Database of Veterinary Systematic Reviews." Last modified 2018. Accessed November 27, 2018. http://webapps.nottingham.ac.uk/refbase.

Centre for Reviews and Dissemination. 2009. *Systematic Reviews: CRD's Guidance for Undertaking Reviews in Health Care.* York: Centre for Reviews and Dissemination.

Chilcott, Jim, Alan Brennan, Andrew Booth, Jonathan Karnon, and Paul Tappenden. 2003. "The Role of Modelling in Prioritising and Planning Clinical Trials." *Health Technology Assessment* 7(23): 1–125.

Cochrane Library. n.d. Accessed November 27, 2018. http://www.cochranelibrary.com.

Collaboration for Environmental Evidence. 2013. *Guidelines for Systematic Reviews in Environmental Management.* Version 4.2. Bangor, UK: Bangor University, Collaboration for Environmental Evidence.

Cooke, Alison, Debbie Smith, and Andrew Booth. 2012. "Beyond PICO: The SPIDER Tool for Qualitative Evidence Synthesis." *Qualitative Health Research* 22(10): 1435–43.

de Vet, Henrica, Anne Eisinga, Ingrid Riphagen, B. Aertgeerts, and D. Pewsner. 2008. "Searching for Studies." In *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*, Version 0.4. London: The Cochrane Collaboration. Accessed November 27, 2018. https://methods.cochrane.org/sdt/handbook-dta-reviews.

Doree, Carolyn, Elke Hausner, Mariann Mathisen, and Siw Waffenschmidt. 2018. "Value of Using Different Search Approaches." HTA Vortal, SuRe Info. Last modified September 6, 2018. Accessed November 27, 2018. http://www.htai.org/vortal/?q=node/993.

EBSCO Health. n.d. "CINAHL Complete." Accessed November 27, 2018. https://www.ebscohost.com/nursing/products/cinahl-databases/cinahl-complete.

EBSCO Information Services. n.d.a. "Criminal Justice Abstracts." Accessed November 27, 2018. https://www.ebscohost.com/academic/criminal-justice-abstracts.

———. n.d.b. "Index to Legal Periodicals & Books Full Text." Accessed November 27, 2018. https://www.ebscohost.com/academic/index-to-legal-periodicals-and-books-full-text.

Eden, Jill, Laura Levit, Alfred Berg, and Sally Morton, eds. 2011. *Finding What Works in Health Care: Standards for Systematic Reviews.* Washington, D.C.: National Academies Press, Institute of Medicine.

Egger, Matthias, Tanja Zellweger-Zähner, Martin Schneider, Christoph Junker, Christian Lengeler, and Gerd Antes. 1997. "Language Bias in Randomised Controlled Trials Published in English and German." *Lancet* 350(9074): 326–29.

Elias, Beth L., Shea Polancich, Carolynn Jones, and Sean Convoy. 2015. "Evolving the PICOT Method for the Digital Age: The PICOT-D." *Journal of Nursing Education* 54(10): 594–99.

Elsevier Life Sciences. n.d. "Embase." Accessed November 27, 2018. https://www.embase.com/login.

Epistemonikos. n.d. Accessed November 27, 2018. http://www.epistemonikos.org.

Erickson, Sonya, and Evelyn R. Warner. 1998. "The Impact of an Individual Tutorial Session on MEDLINE Use Among Obstetrics and Gynaecology Residents in an Academic Training Programme: A Randomized Trial." *Medical Education* 32(3): 269–73.

European Commission. n.d. "CORDIS: Projects & Results Service." Accessed November 27, 2018. http://cordis.europa.eu/projects/home_en.html.

European Food Safety Authority. 2010. "Application of Systematic Review Methodology to Food and Feed Safety Assessments to Support Decision Making." *EFSA Journal* 8(6): 1637.

———. 2015. "Tools for Critically Appraising Different Study Designs, Systematic Review and Literature Searches." *EFSA Journal* 12(7): 836.

Gale. n.d. "Gale Directory Library." Accessed November 27, 2018. https://www.gale.com/databases/gale-directory-library.

Glanville, Julie, Sue Bayliss, Andrew Booth, Yenal Dundar, Hasina Fernandes, Nigel Fleeman, Louise Foster, Cynthia Fraser, Anne Fry-Smith, Su Golder, Carol Lefebvre, Caroline Miller, Suzy Paisley, Liz Payne, Alison Frice, and Karen Welch. 2008. "So Many Filters, So Little Time: The Development of a Search Filter Appraisal Checklist." *Journal of the Medical Library Association* 96(4): 356–61. DOI: 10.3163/1536–5050.96.4.011

Glanville, Julie M., Steven Duffy, Rachael McCool, and Danielle Varley. 2014. "Searching ClinicalTrials.gov and the International Clinical Trials Registry Platform to Inform Systematic Reviews: What Are the Optimal Search Approaches?" *Journal of the Medical Library Association* 102(3): 177–83. DOI: 10.3163/1536-5050.102.3.007.

Glanville, Julie M., Carol Lefebvre, and Kath Wright, eds. 2008. "ISSG Search Filter Resource." York, UK: The InterTASC Information Specialists' Sub-Group. Updated October 2, 2018. Accessed November 24, 2018. https://sites.google.com/a/york.ac.uk/issg-search-filters-resource/home..

Glanville, Julie, Hannah Wood, Mick Arber, Danielle Varbey, Geoff Frampton, and Hugh Brazier. 2014. "Technical Manual for Performing Electronic Literature Searches in Food and Feed Safety." York: York Health Economics Consortium.

Google. n.d. "Google Scholar." Accessed November 27, 2018. https://scholar.google.co.uk.

Grant, Maria J., and Andrew Booth. 2009. "A Typology of Reviews: An Analysis of 14 Review Types and Associated Methodologies." *Health Information and Libraries Journal* 26(2): 91–108.

Greenhalgh, Trisha, and Richard Peacock. 2005. "Effectiveness and Efficiency of Search Methods in Systematic Reviews of Complex Evidence: Audit of Primary Sources." *British Medical Journal* 331(7524): 1064–65.

Grindlay, Douglas J. C., Marnie L. Brennan, and Rachel S. Dean. 2012. "Searching the Veterinary Literature: A Comparison of the Coverage of Veterinary Journals by Nine Bibliographic Databases." *Journal of Veterinary Medicine Education* 39(4): 404–12.

Hausner, Elke, Charlotte Guddat, Tatjana Hermanns, Ulrike Lampert, and Siw Waffenschmidt. 2015. "Development of Search Strategies for Systematic Reviews: Validation Showed the Noninferiority of the Objective Approach." *Journal of Clinical Epidemiology* 68(2): 191–99.

Hausner, Elke, Siw Waffenschmidt, Thomas Kaiser, and Michael Simon. 2012. "Routine Development of Objectively Derived Search Strategies." *Systematic Reviews* 1: 19.

Higgins, Julian, and Sally Green, eds. 2011. *Cochrane Handbook for Systematic Reviews of Interventions*, Version 5.1.0. London: The Cochrane Collaboration.

Hopewell, Sally, Mike Clarke, Carol Lefebvre, and Roberta W. Scherer. 2007. "Handsearching Versus Electronic Searching to Identify Reports of Randomized Trials." In *Cochrane Database of Systematic Reviews* (2): MR000001. DOI: 10.1002/14651858.MR000001.pub2.

Hopewell, Sally, Mike Clarke, Lesley Stewart, and Jayne Tierney. 2007. "Time to Publication for Results of Clinical Trials." *Cochrane Database of Systematic Reviews* (2): MR000011. DOI: 10.1002/14651858.MR000011.pub2.

Hopewell, Sally, Steve McDonald, Mike Clarke, and Matthias Egger. 2007. "Grey Literature in Meta-Analyses of Randomized Trials of Health Care Interventions." *Cochrane Database of Systematic Reviews* (2): MR000010. DOI: 10.1002/14651858.MR000010.pub3.

Horsley, Tanya, Orvie Dingwall, and Margaret Sampson. 2011. "Checking Reference Lists to Find Additional Studies for Systematic Reviews." *Cochrane Database of Systematic Reviews* (8): MR000026. DOI: 10.1002/14651858.MR000026.pub2.

Institute of Education Sciences. n.d. "ERIC." Accessed November 27, 2018. http://eric.ed.gov.

Jenkins, Michelle. 2004. "Evaluation of Methodological Search Filters—A Review." *Health Information and Libraries Journal* 21(3): 148–63.

Joanna Briggs Institute. 2014. *2014 Reviewers' Manual*. Adelaide: University of Adelaide, Joanna Briggs Institute. Accessed November 27, 2018. http://joannabriggs.org/assets/docs/sumari/ReviewersManual-2014.pdf.

Kassai, Tibor. 2006. "The Impact on Database Searching Arising from Inconsistency in the Nomenclature of Parasitic Diseases." *Veterinary Parasitology* 138(3–4): 358–61.

Kugley, Shannon, Anne Wade, James Thomas, Quenby Mahood, Anne-Marie Klint Jørgensen, Karianne Thune Hammerstrøm, and Nila Sathe. 2016. *Searching for Studies: A Guide to Information Retrieval for Campbell Systematic Reviews*. Oslo: The Campbell Collaboration. Accessed November 24, 2018. http://www.campbellcollaboration.org/artman2/uploads/1/Campbell_Methods_Guides_Information_Retrieval_1.pdf.

Kuller, Alice B., Charles B. Wessel, David S. Ginn, and Thomas P. Martin. 1993. "Quality Filtering of the Clinical Literature by Librarians and Physicians." *Bulletin of the Medical Library Association* 81(1): 38–43.

Lefebvre, Carol, Eric Manheimer, and Julie Glanville. 2011. "Searching for Studies." In *Cochrane Handbook for Systematic Reviews of Interventions*, Version 5.1.0, edited by Julian Higgins and Sally Green. London: The Cochrane Collaboration.

Levay, Paul, Nicola Ainsworth, Rachel Kettle, and Antony Morgan. 2015. "Identifying Evidence for Public Health Guidance: A Comparison of Citation Searching with Web of Science and Google Scholar." *Research Synthesis Methods* 7(1): 34–45.

Linder, Suzanne, Geetanjali Kamath, Gregory Pratt, Smita Saraykar, and Robert Volk. 2015. "Citation Searches Are More Sensitive than Keyword Searches to Identify Studies Using Specific Measurement Instruments." *Journal of Clinical Epidemiology* 68(4): 412–17.

Marshall, Christopher. n.d. "SR Toolbox." Accessed November 27, 2018. http://systematicreviewtools.com/index.php.

McGowan, Jessie, Margaret Sampson, Douglas M. Salzwedel, Elise Cogo, Vicki Foerster, and Carol Lefebvre. 2016. "PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement." *Journal of Clinical Epidemiology* 75(1): 40–46.

Moher, David, Alessandro Liberati, Jennifer Tetzlaff, Douglas Altman, and The PRISMA Group. 2009. "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement." *PLoS Med* 6(7): e1000097.

National Institute for Health Research. n.d. "PROSPERO: International Prospective Register of Systematic Reviews." Accessed November 27, 2018. http://www.crd.york.ac.uk/PROSPERO/.

Nielen, M., Cas Kruitwagen, and Anton Beynen. 2006. "Effect of Vitamin E Supplementation on Udder Health: A Meta-Analysis." Exeter: Society for Veterinary Epidemiology and Preventive Medicine.

O'Brien, Sarah, Iain A. Gillespie, M. A. Sivanesan, Richard Elson, C. Hughes, and Goutam K. Adak. 2006. "Publication Bias in Foodborne Outbreaks of Infectious Intestinal Disease and Its Implications for Evidence-Based Food Policy. England and Wales 1992–2003." *Epidemiology and Infection* 134(4): 667–74.

OCLC WorldCat. n.d. Accessed November 27, 2018. https://www.worldcat.org.

O'Mara-Eves, Alison, James Thomas, John McNaught, Makoto Miwa, and Sophia Ananiadou. 2015. "Using Text Mining for Study Identification in Systematic Reviews: A Systematic Review of Current Approaches." *Systematic Reviews* 4(5).

Paynter, Robin, Lionel Bañez, Elise Berliner, Eileen Erinoff, Jennifer Lege-Matsuura, Shannon Potter, and Stacey Uhl. 2016. "EPC Methods: An Exploration of the Use of Text-Mining Software in Systematic Reviews. Research White Paper." Rockville, Md.: Agency for Healthcare Research and Quality.

Petticrew, Mark, and Helen Roberts, eds. 2006. *Systematic Reviews in the Social Sciences.* Oxford: Blackwell Publishing.

PRESSForum. n.d. Accessed November 27, 2018. http://pressforum.pbworks.com.

ProQuest. n.d. "ProQuest Dissertations & Theses Global." Accessed November 27, 2018. http://www.proquest.com/products-services/pqdtglobal.html.

Rathbone, John, Matt Carter, Tammy Hoffmann, and Paul Glasziou. 2016. "A Comparison of the Performance of Seven Key Bibliographic Databases in Identifying All Relevant Systematic Reviews of Interventions for Hypertension." *Systematic Reviews* 5(1): 27.

Sampson, Margaret, and Jessie McGowan. 2011. "Inquisitio Validus Index Medicus: A Simple Method of Validating MEDLINE Systematic Review Searches." *Research Synthesis Methods* 2(2): 103–109.

Sampson, Margaret, Li Zhang, Andra Morrison, Nicholas J. Barrowman, Tammy Clifford, Robert W. Platt, Terry P. Klassen, and David Moher. 2006. "An Alternative to the Hand Searching Gold Standard: Validating Methodological Search Filters Using Relative Recall." *BMC Medical Research Methodology* 6(1): 33.

Sargeant, Jan M., Mary E. Torrence, Andrijana Rajić, Annette M. O'Connor, and Jodi Williams. 2006. "Methodological Quality Assessment of Review Articles Evaluating Interventions to Improve Microbial Food Safety." *Foodborne Pathogens and Disease* 3(4): 447–56.

SCOPUS. n.d. Accessed November 27, 2018. https://www.scopus.com.

Shemilt, Ian, Antonia Simon, Gareth J. Hollands, Theresa Marteau, David Ogilvie, Alkison O'Mara-Eves, Michael P. Kelly, and James Thomas. 2014. "Pinpointing Needles in Giant Haystacks: Use of Text Mining to Reduce Impractical Screening Workload in Extremely Large Scoping Reviews." *Research Synthesis Methods* 5(1): 31–49.

Snedeker, Kate G., Sarah C. Totton, and Jan M. Sargeant. 2010. "Analysis of Trends in the Full Publication of Papers from Conference Abstracts Involving Pre-Harvest or Abattoir-Level Interventions Against Foodborne Pathogens." *Preventive Veterinary Medicine* 95(1–2): 1–9.

Song, Fujian, S. Parekh, Lee Hooper, Y. K. Loke, J. Ryder, A. J. Sutton, C. Hing, C. S. Kwok, C. Pang, and I. Harvey. 2010. "Dissemination and Publication of Research Findings: An Updated Review of Related Biases." *Health Technology Assessment* 14(8): 1–193.

Spoor, Pat, Mark Airey, Cathy Bennett, and Julie Greensill. 1996. "Use of the Capture-Recapture Technique to Evaluate the Completeness of Systematic Literature Searches." *British Medical Journal* 313(7053): 342–43.

Stansfield, Claire, Ginny Brunton, and Rebecca Rees. 2014. "Search Wide, Dig Deep: Literature Searching for Qualitative Research: An Analysis of the Publication Formats and Information Sources Used for Four Systematic Reviews in Public Health." *Research Synthesis Methods* 5(2): 142–51.

Sterne, Jonathon, Matthias Egger, and David Moher. 2011. "Addressing Reporting Biases." In *Cochrane Handbook for Systematic Reviews of Interventions*, Version 5.1.0., edited by Julian Higgins and Sally Green. London: The Cochrane Collaboration.

Thomson Reuters. 2014. "Biosis Citation Index." Accessed November 27, 2018. http://wokinfo.com/products_tools/specialized/bci.

Thomson Reuters. n.d.a. "Conference Proceedings Citation Index." Accessed November 27, 2018. http://thomsonreuters.com/en/products-services/scholarly-scientific-research/scholarly-search-and-discovery/conference-proceedings-citation-index.html.

Thomson Reuters. n.d.b. "Science Citation Index Expanded." Accessed November 27, 2016. http://thomsonreuters.com/en/products-services/scholarly-scientific-research/scholarly-search-and-discovery/science-citation-index-expanded.html.

UK Research and Innovation. 2018. "Gateway to Research." Last updated October 30, 2018. http://gtr.rcuk.ac.uk.

U.S. National Library of Medicine. n.d.a. "ClinicalTrials.gov." Accessed November 27, 2018. https://clinicaltrials.gov.

———. n.d.b. "PubMed." Accessed November 27, 2018. https://www.ncbi.nlm.nih.gov/pubmed/.

Virtual Health Library. n.d. "LILACS Database Search." Accessed November 27, 2018. http://bases.bireme.br/cgi-bin/wxislind.exe/iah/online/?IsisScript=iah/iah.xis&base=LILACS&lang=i&form=F.

Whiting, Penny, Marie Westwood, Margaret Burke, Jonathan Sterne, and Julie Glanville. 2008. "Systematic Reviews of Test Accuracy Should Search a Range of Databases to Identify Primary Studies." *Journal of Clinical Epidemiology* 61(4): 357–64.

Wikipedia. 2018a. "Comparison of Reference Management Software." Last modified November 3, 2018. https://en.wikipedia.org/wiki/Comparison_of_reference_management_software.

———. 2018b. "RIS (file format)." Last modified August 3, 2018. https://en.wikipedia.org/wiki/RIS_(file_format).

Wilczynski, Nancy L., Cynthia Lokker, Kathleen Ann McKibbon, Nicholas Hobson, and R. Brian Haynes. 2016. "Limits of Search Filter Development." *Journal of the Medical Library Association* 104(1): 42–46.

Wildridge, Valerie, and Lucy Bell. 2002. "How CLIP Became ECLIPSE: A Mnemonic to Assist in Searching for Health Policy/Management Information." *Health Information and Libraries Journal* 19(2): 113–15.

Wood, Hannah, and Mick Arber. 2016a. "Search Filters." HTA Vortal. Last modified April 11, 2016. Accessed September 28, 2016. http://www.htai.org/vortal/?q=node/573.

———. 2016b. "Search Strategy Development." HTA Vortal. Last modified March 30, 2016. Accessed September 28, 2016. http://www.htai.org/vortal/?q=node/790.

Wright, Kath, and Catriona McDaid. 2011. "Reporting of Article Retractions in Bibliographic Databases and Online Journals." *Journal of the Medical Library Association* 99(2): 164–67.

6

RETRIEVING GREY LITERATURE, INFORMATION, AND DATA IN THE DIGITAL AGE

DEAN GIUSTINI
University of British Columbia

CONTENTS

6.1 INTRODUCTION

The objective of this chapter is to discuss the retrieval of grey literature in support of the research synthesis and best practices that can be used to systematically conduct such a search. It is assumed that readers will be familiar with the basics of structured database searching in the social, behavioral, or medical sciences and will have read previous chapters in this volume, notably chapter 5, entitled "Searching Bibliographic Databases." To get the most from the present chapter, readers should read the descriptions of terms in table 5.1 of chapter 5, "Key Concepts in Database Searching to Inform Research Synthesis."

Grey literature is literature that has not been formally published, has limited distribution, or is not available via conventional channels (Auger 1998; Bonato 2018). Research synthesis is the practice of systematically retrieving, distilling, and integrating data from a variety of sources ("the evidence base") to draw more reliable conclusions from the literature (Cooper and Hedges 2009). Hannah Rothstein and Sally Hopewell observe "a critical relationship between the reliability and validity of a research synthesis, and the thoroughness of and lack of bias in the search for relevant studies" (2009, 104). In a high-quality review that aims to inform policy or practice, the goal is to identify all available evidence, including the grey literature, relevant to the question (Boland, Cherry, and Dickson 2013; Institute of Medicine 2011).

The chapter examines definitions and document types specified as grey literature, focusing on issues such as methodical search planning, mapping resources to a research question, and identifying as many potentially relevant resources for searching as possible. However, the strategies and techniques by which desired outcomes in grey literature searching are reached will be unique in each project. Key approaches and techniques are offered to assist searchers in building their search strategies and refining each one to suit an individual purpose and context.

Seeking advice from experts in systematic information retrieval is highly recommended before any significant grey literature searching (Lefebvre, Manheimer, and Glanville 2011; Boland, Cherry, and Dickson 2013). This chapter discusses the role of grey literature searching in amassing a representative body of literature for topics and the importance of peer review of search strategies or checking them with a librarian (McGowan et al. 2016a). It is also concerned with extending grey literature search techniques to locate a higher percentage of relevant documents.

6.2 WHAT IS GREY LITERATURE?

> "Grey literature stands for manifold document types produced on all levels of government, academics, business and industry in print and electronic formats that are protected by intellectual property rights, of sufficient quality to be collected and preserved by libraries and institutional repositories, but not controlled by commercial publishers; i.e. where publishing is not the primary activity of the producing body."
>
> Twelfth International Conference on Grey Literature, 2010 (Schöpfel 2010)

6.2.1 Terminologies, Types, and Definitions

Although subject to continuous revision, the best-known, classic definitions of grey literature are those formally accepted at the International Conferences on Grey Literature in 1997, 2004, and 2010 (Schöpfel 2010). The concept of a vast *fugitive* literature was once described by Charles Auger, who said that the research literature was characterized by "a vast body of documents" of "continuing increasing quantity" and was significant for "the difficulty it presented to the librarian" (1975). What these and other definitions underline are the challenges associated with finding grey literature today because of the speed and ease with which materials are published in the internet age. The question of whether these print era definitions are applicable in the digital age is also a challenge (Schöpfel and Rasuli 2018).

Amanda Lawrence refers to a body of grey literature that is produced informally and defined by its elusive nature (2012). Further, grey literature includes a range of documents such as government publications, policy documents, white reports, fugitive or nonconventional literature, unpublished literature, and nontraditional publications of various kinds in print and online (Lawrence 2012). Over the years, Vilma Alberani and Irwin Weintraub have each added a great deal to our evolving understanding of grey literature typologies and terminologies (Alberani, De Castro Pietrangeli, and Mazza 1990; Weintraub 2000). The best way to understand grey literature in the digital age is to compare it with the published literature across key characteristics (see table 6.1).

6.2.2 Characteristics

Grey literature documents increasingly refer to a broader and heterogeneous body of materials (Benzies et al. 2006; Bonato 2018). In the twenty-first century, many papers not formally published in academic journals can be classified as grey literature (Bellefontaine and Lee 2014; Paez 2017). Some of these papers are produced well beyond commercial outlets and are consistently elusive and hard to find

Table 6.1 Grey Literature Versus Published Literature

Issues	Grey literature: harder to find	Published literature: easier to find
Number of documents being published	Increasing rapidly, exponentially	Increasing also, but slower in pace
Speed of production	Instantaneous, due to the speed of desktop and self-publishing	Slow(er); time lag due to editing and peer-review processes
Costs	Low(er) (in most cases) or free; some market research firms and manufacturers charge for access	High(er) due to editorial, production, marketing costs. Publisher profit motives. Some predatory publishers.
Access	Free, open, immediate in many cases; some payment or association memberships required to access; some literature driven "underground"	Some free and open access; mostly paywalled (or restricted) due to digital rights management. Some literature is lost due to changing publishers
Quality	Highly variable but is often peer reviewed and written by experts; may be produced with a "hidden social or political agenda." Watch for bogus research firms.	Peer-reviewed quality can be high; but not immune from fraud, errors plagiarism, editorial interference, retracted publications.
Findability	Improving, but often dispersed; may be "hit and miss" and hidden in the deep web. Linked data holds some promise to bring relevant documents together.	Highly variable. May be elusive if subscription databases are required. Some literature is only accessible to native speakers of grey report languages (for example, Chinese, Russian)
Archiving and preservation	Weaker archival and preservation practices and policies; may be difficult due to volume and multiple formats; rise of datasets; lack of infrastructure	Stronger, due to better digital infrastructure(s) and libraries; some problems due to legal restrictions, and publisher amalgamations
Impact on libraries	Challenges and opportunities re: access, cataloguing, description, preservation and findability	Complex due to rising costs, back-up copies, copyright, licensing, digital rights management, storage
Role of publishers	Some content available via the creative commons or made open access for wider dissemination	Commercial interests of publishers based on economic models, not scholarly models of communication

SOURCE: Author's compilation.

(Coonin 2003). These are not the only factors in determining whether something is grey, however. According to Rose Relevo and Howard Balshem, "Grey literature is, by definition, not systematically identified, stored, or indexed and therefore it can be difficult to locate" (2011, 2).

Taryn Rucinski identifies some of the more than 150 possible types of grey literature such as discussion papers, newsletters, surveys, working papers, technical reports, trade association publications, institutional or association reports, conference proceedings, academic and government reports (2015). Other grey documents include unpublished manuscripts, surveys, product catalogs, presentations, preprints, practice guidelines, and lecture notes, to name a few (Giustini 2016). In a recent monograph, Sarah Bonato explores the diversity of grey literature in the digital age in the form of governmental reports, working papers, and other unpublished research (2018).

New scholarly publishing platforms and social media have introduced categories of literature that fall into a distinct "grey zone" (Banks 2009; Aloia and Naughton 2016; Giustini 2016). Some researchers refer to information produced informally such as emails, meeting minutes and personal memories as "grey information" (Adams et al. 2016). "Grey data" refers to user-generated content on the web in the form of tweets, blog posts, and Facebook status updates (Adams et al. 2016). Increasingly, unpublished research and the data they cite are referred to as grey literature (Godin et al. 2015).

Joachim Schöpfel and Dominic Farace note that better conceptual frameworks are needed to understand and monitor grey literature's evolution (2011). The world of information in the twenty-first century is often portrayed as black and white but new shades of grey are needed (Rucinski 2015). According to Julie Gelfand and Daniel Tsang, "the definition [of grey literature] has been challenged as still being too narrow considering new forms and practices of scholarship and research underway and the methods of publishing now widely available" (2015, 30). Grey literature will continue to change in response to the production of new scientific knowledge in the digital age.

6.3 VALUE, IMPACT, AND QUALITY

The importance of grey literature searching in the research synthesis varies from topic to topic and the type of review undertaken (Grant and Booth 2009; Booth, Sutton, and Papaioannou 2016). In some cases, the benefits of including grey literature will outweigh the time and resources required to search for it (Paez 2017). The decision about whether to perform extensive grey literature searching after the standard bibliographic databases may be the most complex issue to resolve (Benzies et al. 2006). Generally, the impact of grey literature is related to its role in broadening the set of papers found for the research synthesis. Despite its ephemeral nature, grey literature can enhance a research synthesis when published evidence is scant or topics are new, changing quickly or interdisciplinary (Relevo and Balshem 2011). In 2006, Karen Benzies and her colleagues published a checklist to determine whether a state-of-the-evidence review would benefit from grey literature and made recommendations to include a consensus in the research is lacking and when the availability of existing evidence is scant or low quality. In pediatrics, Lisa Hartling and her colleagues conclude that grey literature searching represented a small proportion of included studies and rarely changed the results or conclusions of a review (2017). Still, inclusion of grey literature may have a direct impact when relevant studies are relatively few (Enticott, Buck, and Sawyer 2018), or when vested interests in the published literature may be questionable (Hartling et al. 2017).

In health technology assessment, some types of grey literature may be vital to the research synthesis, such as clinical study reports, synopses, regulatory data, trial registry records, conference proceedings, and abstracts (Halfpenny et al. 2016; Farrah and Mierzwinski-Urban 2019). According to some, grey literature encompasses materials such as unpublished trial data, government documents, and manufacturers' information (Relevo and Balshem 2011). In the context of medical devices and drug research, and in health or public policy where the literature may be limited, locating and retrieving relevant grey literature is an essential part of the review. Grey literature can be important in dynamic and innovative fields where relatively little academic work has been done and when practice is seen to be ahead of research investigations (Adams, Smart, and Huff 2017).

Research synthesists should be aware of the problems associated with searching major bibliographic databases alone. Database bias (and other biases) may result from the inclusion or exclusion of research papers indexed in widely used databases such as MEDLINE and Embase (Egger et al. 2003). A bias toward positive findings may result when papers are found only by searching the major databases. This is because statistically significant positive results are more likely to be published in peer-reviewed, indexed papers. Grey literature may be important to the overall results of the review when studies with null results are located (Adams et al. 2016). Similarly, excluding

grey literature can lead to inflated estimates of treatment effects (Blackhall 2007). Without searching for the grey literature the research synthesis can result in a distorted or incomplete view of a topic (Schmucker et al. 2013, 2017).

Schöpfel asserts that grey literature's importance depends on each discipline (2010). In the biomedical and life sciences, the traditional preference is for peer-reviewed papers; in agriculture and aeronautics, grey literature plays a more prominent role (Schöpfel 2010). The Institute of Medicine states that grey literature searching should form part of all systematic reviews (2011). Jean Adams and her colleagues emphasize the value of grey literature in the investigation of public health interventions (2016). Grey literature provides useful perspectives in engineering and law (Rucinski 2015). In the social sciences, a search for grey literature lends credibility and value to review syntheses of various kinds (Adams, Smart, and Huff 2017). Grey literature can provide more details than the published literature because no length restrictions come into play (Adams, Smart, and Huff 2017). However, as more supplemental files and data are attached to the published journal literature, this distinction may become less obvious over time.

6.3.1 Challenges in Use

Debate is mounting about whether every research synthesis should include a search for grey literature (Bellefontaine and Lee 2014; Hartling et al. 2017). Time-consuming and complex literature searches, which cover the grey literature and all relevant languages and databases, are normally recommended to prevent reporting biases from being introduced into the research synthesis (Egger et al. 2003). All relevant specialist theses and dissertations on the topic, and their references, should at least be reviewed (Bellefontaine and Lee 2014). Andrew Booth mentions the depth of reporting possible in PhD dissertations but warns against their uncritical inclusion, especially when their results swamp the findings in other smaller studies (2016).

The research synthesis team can determine the degree of comprehensiveness by taking into account the specific requirements of the review and its resources (Egger et al. 2003). Some researchers argue in favor of finding all evidence regardless of the time it takes (Boland, Cherry and Dickson 2013). However, the time it takes to find only one additional obscure report or study that in all probability will not change the results of a review may not be worthwhile (Boland, Cherry, and Dickson 2013; Finfgeld-Connett and Johnson 2013). In some cases, grey literature may be helpful "to tip the balance" when evidence for an intervention is inconclusive (Hickner, Friese, and Irwin 2011, 32). Systematic exclusion of the grey literature would be ill advised in some cases given the scarcity of information in certain disciplines (Martinez, Williams, and Yu 2015).

In some disciplines, the belief is persistent that the quality of grey literature is uneven or low relative to traditionally published papers (Hopewell, MacDonald, et al. 2007). Some researchers have said that in light of transparency and quality concerns, papers available in certain grey literature outlets are of limited value (Martinez, Williams, and Yu 2015). Further, the perception is prevalent that grey literature is produced by authors with no academic credentials or an interest in publishing in outlets that adhere to scholarly publishing norms (Adams, Smart, and Huff 2017). Peer-reviewed articles may be sufficient for some topics, leaving little to be gained by adding the grey literature. Papers indexed outside MEDLINE or Embase are more difficult to find and may require translation into English, which will increase costs and delay the conclusion of a review (Egger et al. 2003). Further, "trials that are difficult to locate tend to be of lower methodological quality than trials that are easily accessible and published in English" (Egger et al. 2003, 3).

Some researchers believe that grey literature is not consistently peer-reviewed in the way academic journal articles are reviewed (Boland, Cherry, and Dickson 2013). In a 2004 survey, 44.2 percent of respondents felt that grey literature had benefited from some kind of peer review (Boekhorst, Farace, and Frantzen 2005). Others say that grey literature is often high quality and written by experts (Yasin and Hasnain 2012). Quality-control mechanisms, such as editing and peer review, are often but not always part of publishing grey literature (Conn et al. 2003).

Research synthesists can evaluate the quality of component papers in the research synthesis through robust critical appraisal. The Critical Appraisal Skills Programme (CASP) Tools developed by Oxford University in the United Kingdom and the Joanna Briggs Institute (JBI) in Australia can be used to evaluate grey literature. The AACODS checklist, a tool specifically designed for evaluating grey literature, which stands for authority, accuracy, coverage, objectivity, date, and significance can also be used (Tyndall 2010). Affan Yasain and Muhammad Hasnain created a similar list of issues to ask in assessing grey literature based on a DARE (Database of Abstracts of Reviews of Effects) checklist (2012). Because these questions included where a paper was discussed, Yasain

and Hasnain observe that grey reports are more likely to be discussed online and cited if they are valued as research.

6.3.2 Costs and Resources

One of the significant challenges with grey literature is that it is the least efficient type of literature to find (Cook et al. 2001). Researchers see its value but may not be fully aware of the required investment of time to find it (Booth, Sutton, and Papaioannou 2016). Performing exhaustive searches of grey literature presents a considerable work burden to the searcher (Balshem et al. 2013). Internet searches, where much of the grey literature is located, can be difficult to design and time consuming to execute (Benzies et al. 2006).

One study reported on how long it took to perform expert searches (developing, refining, and adapting searches to different databases), specifying two weeks of a librarian's time and a large investment of resources (Greenhalgh and Peacock 2005). Ahlam Saleh, Melissa Ratajeski, and Marnie Bertolet find that the average time spent searching for the systematic review was twenty-four hours, within a range of two to 113 hours, and half the searches taking eight or fewer hours (2014). They identified the time spent searching by examining eighteen systematic reviews that reported some form of grey literature searching. The average time spent searching online was approximately seven hours, within a range of twenty minutes to fifty-eight hours. Locating grey literature consumed about 27 percent of the total searching time.

6.3.3 Challenges in Finding and Preserving

Even when the quality of grey literature is acceptable, the question remains as to whether it will be found efficiently, even within the standard bibliographic databases. Grey literature was present in the majority (68 percent) of bibliographic databases and nearly all institutional repositories (95 percent) that Wanda Marsolek and her colleagues examined (2018). Internet search engines and open access have made grey literature more accessible (Banks 2004; Marsolek et al. 2018), but a range of acquisition and cataloging issues continue to affect its findability (Childress and Jul 2003; Okoroma 2011; Vaska and Vaska 2016). Poor findability is often related to weak preservation and data management practices (Gelfand and Tsang 2015). In fact, some grey literature producers do not consistently preserve publications for the future thereby creating "endangered documents" (Schöpfel and Farace 2011). Some scientists do not consider the long-term durability of their unpublished datasets thereby making them hidden or invisible to researchers; in one analysis, the authors sought to retrieve lost data and found that grey data substantially increased the size of their study sample (Augusto et al. 2010). Recent international requirements for data management for funded research at the National Institutes of Health and the Canadian Institute of Health Research (CIHR), for example, should further improve the archiving and preservation of scientific datasets into the future (NIH 2015; Canada 2018).

Julia Gelfand and Daniel Tsang suggest that not all grey literature (or data) should be preserved (2015). Janice Kung and Sandy Campbell point to a lack of policies for selecting valuable research data for preservation (2016). Making a determination of what to preserve is a challenge for librarians and researchers. This is why the definition of grey literature, updated at the International Conference on Grey Literature in 2010, emphasizes preserving grey literature of "sufficient quality." Preservation is also at the center of GreyNet International's Pisa Declaration (Giustini 2014). Other international projects are looking at document preservation such as LOCKSS (Lots of Copies Keep Stuff Safe), based at Stanford University and CLOCKSS (Controlled Lots of Copies Keep Stuff Safe), an independent nonprofit organization in the United States. Portico, operated by the organization that produces JSTOR, is another example (Mering 2015). Ultimately, researchers can take steps to ensure that their papers are preserved for the long term.

6.4 PREPARING FOR SEARCHING

Extensive grey literature searching begins with methodical planning and preparation; this includes determining the data and citation requirements of the research synthesis (Adams et al. 2016). Planning to search for the grey literature outside bibliographic databases is an important consideration in several types of research and research methods (Booth, Sutton, and Papaioannou 2016). Narrowing down which resources are appropriate for grey literature searching is key to planning (Booth 2016).

6.4.1 Sources

A number of subject guides and checklists are worth consulting and have been published by organizations such as the Canadian Agency for Drugs and Technologies in Health (CADTH), the Agency for Healthcare Research and Quality (AHRQ) and the Cochrane Collaboration (see table 6.2). Academic subject guides written by librarians

Table 6.2 Sources of Grey Literature and Data

This list of guides and starting points is not exhaustive but will help to locate online resources and websites that pertain to your research synthesis. The following list is organized into five main categories, and begins with where to start with early search planning by using various meta-lists, followed by meta-search tools and larger more comprehensive search tools, databases, archives, and repositories. Some examples include the directories of open access journals, data repositories, clinical trial registries and specialized bibliographic databases. Identify items of interest and visit their websites to locate publications that are relevant.

1. Meta-lists, structured checklists and other information starting points
   - Cochrane Handbook for Systematic Reviews of Interventions (https://training.cochrane.org/handbook)
   - Campbell Collaboration Information Retrieval Guide (https://campbellcollaboration.org/information-retrieval-guide.html)
   - Grey Matters: a practical tool for searching health-related grey literature (https://www.cadth.ca/resources/finding-evidence/grey-matters)
   - HTAi Vortal (http://vortal.htai.org, see also SuRe Info)
   - Joanna Briggs Institute Reviewers Manual (https://wiki.joannabriggs.org/display/MANUAL/Joanna+Briggs+Institute+Reviewer%27s+Manual)
   - National Library of Medicine. "HTA101: VII. Retrieve Evidence". National Information Center on Health Services Research and Health Care Technology (NICHSR, https://www.nlm.nih.gov/nichsr/hta101/ta10109.html)
   - Public Health Grey Literature Sources (Canada and Beyond) (http://www.ophla.ca/pdf/Public%20Health%20Grey%20Literature%20Sources.pdf)
   - Subject guides written by librarians via Google search
   - Summarized Research in Information Retrieval (SuRe Info, http://vortal.htai.org/?q=sure-info)
   - Wikipedia (https://en.wikipedia.org/wiki/Main_Page)

2. Meta-search databases, platforms, repositories
   - Academic Search (https://www.ebsco.com/products/research-databases/academic-search-complete)
   - BASE (Bielefield Academic Search Engine, https://www.base-search.net/)
   - Cochrane Library (https://www.cochranelibrary.com/)
   - Campbell Collaboration Library (https://campbellcollaboration.org/library.html)
   - DOAJ (Directory of Open Access Journals, https://doaj.org/)
   - ERIC (Education Resources Information Center, https://eric.ed.gov/)
   - Google Book Search (https://books.google.com/), Google Scholar (https://scholar.google.ca/), and Microsoft Academic Search (https://academic.microsoft.com/)
   - GreyGuide (http://greyguide.isti.cnr.it/)
   - Institute of Medicine (IOM, http://www.nationalacademies.org/hmd/)
   - McMaster Health Systems Evidence (https://www.healthsystemsevidence.org/)
   - New York Academy of Medicine, "The Grey Literature Report" (http://www.greylit.org/)
   - National Technical Information Service (NTIS, https://www.ntis.gov/)
   - OA subject repositories; arXiv (https://arxiv.org/), bioRxiv (https://www.biorxiv.org/), ChemRxiv (https://chemrxiv.org/), engrXiv (https://blog.engrxiv.org/), SocArXiv (https://osf.io/preprints/socarxiv/), Social Science Research Network (SSRN, https://www.ssrn.com/en/)
   - OA publishing platforms: BioMedCentral (https://www.biomedcentral.com/), PubMedCentral (https://www.ncbi.nlm.nih.gov/pmc/)
   - OpenGrey System for Information on Grey Literature in Europe (http://www.opengrey.eu/)
   - OAISter (http:/oaister.worldcat.org/)
   - OpenDOAR: Registry of Open Access Repositories (http://v2.sherpa.ac.uk/opendoar/)
   - PQDT Open (https://pqdtopen.proquest.com/search.html), Thesis Canada (http://www.bac-lac.gc.ca/eng/services/theses/Pages/theses-canada.aspx), NTDLD (http://www.ndltd.org/), British Library EThOS (https://ethos.bl.uk/) and Dart Europe (http://www.dart-europe.eu/basic-search.php)
   - Scopus (https://www.elsevier.com/solutions/scopus)
   - Web of Science (https://clarivate.com/products/web-of-science/)
   - WorldCat (https://www.worldcat.org/)

(*Continued*)

Table 6.2 (*Continued*)

3. Some producers of syntheses
   - Agency for Healthcare Research and Quality (AHRQ, https://www.ahrq.gov/)
   - Cochrane Collaboration (http://www.cochrane.org/)
   - Campbell Collaboration (http://www.campbellcollaboration.org)
   - Joanna Briggs Institute (http://joannabriggs.org/)
   - Canadian Agency for Drugs and Technologies in Health (CADTH, https://www.cadth.ca/)
   - EPPI-Centre database of educational research (http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=185)
4. Data repositories and trial registries
   - ClinialTrials.gov (https://clinicaltrials.gov/)
   - ISRCTN Registry–metaRegister of Controlled Trials (http://www.isrctn.com/page/mrct)
   - PROSPERO: International prospective register of systematic reviews (https://www.crd.york.ac.uk/PROSPERO/)
   - OpenTrials (https://www.crd.york.ac.uk/PROSPERO/)
   - Registry of Research Data Repositories (re3data, https://www.re3data.org/)
   - WHO International Clinical Trials Registry Platform (https://www.who.int/ictrp/en/)
5. Highly specialized bibliographic databases
   - Ageline (https://www.ebsco.com/products/research-databases/ageline)
   - Agricola (https://www.ebsco.com/products/research-databases/agricola)
   - PapersFirst (FirstSearch, https://help.oclc.org/Discovery_and_Reference/FirstSearch/FirstSearch_databases)
   - POPLINE (https://www.popline.org/poplinesubjects)

SOURCE: Author's compilation.

will also be useful structured tools in outlining the steps of grey literature searching and framing its principles for each discipline (Vaska and Vaska 2016).

Creating a document with detailed search steps should be viewed as adhering to reporting standards such as PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) (Godin et al. 2015). According to Dagmara Chojecki and Lisa Tjosvold, however, recording and reporting the search are two separate processes: "that of documenting the search strategy and that of reporting the search strategy." Further, "documenting the search can be seen as recording information about the search simultaneously as it is run. This information is often used for internal purposes and records. Reporting the search involves the final formal write up of the search methodology that will be included in the published version of the review" (2016).

Grey literature searching presents the researcher with various challenges in terms of search reproducibility. This is because grey literature searching is a more uncontrolled and iterative type of searching than that performed in standard bibliographic databases (Booth, Sutton, and Papaioannou 2016; Godin et al. 2015). The resulting uncontrollable mass of documents found by searching for grey literature on some websites is often due to their lack of organization. Their unique search facilities impose additional challenges and limitations that require repetitious searching and navigation in many places (Stansfield, Dickson, and Bangpan 2016). Further, many of these websites do not permit the narrowing of topics or provide features such as sorting, refining, or saving results. Some of the larger concerns expressed about the difficulty of using these websites are their inability to perform complex Boolean search queries and to export citations (Stansfield, Dickson, and Bangpan 2016).

6.4.2 Search Construction

I now turn to a priori planning, expecting that readers will develop a more structured process for their searching. Search planning should result in locating a higher percentage of the overall relevant literature. A search planning document or grid is the first step in this construction (see table 6.3).

There is no true "gold standard" for grey literature searching and insufficient detail about how to conduct such a search (Adams et al. 2016; Godin et al. 2015). The goal in systematic searching is to generate as exhaustive a list as possible of significant published and unpublished studies that are relevant to a research question (Lefebvre, Manheimer, and Glanville 2011). Systematic searchers should aim to strike a balance between sensitivity (as many

Table 6.3 Grey Literature Search Construction and Planning

1. Create a search document (planning grid) with tables from left to right using Word or an Excel spreadsheet (see also table 6.4).
2. State the review question using a framework (PICO, ECLIPSE) to break down the topic with all concepts listed (or use your own conceptual breakdown).
3. List key databases, web resources, conference proceedings, repositories, and search engines (platforms or vendors used) to be searched (and by whom).
4. Take note that online resources will cover different subject areas and formats; each will have its own unique organization.
   - Search strategies should be tailored to each website, accounting for characteristics of search interfaces and areas of focus of host organization.
   - Ensure websites include affiliated organizations and website addresses.
5. List all free-text terms, concepts, associated synonyms, subjects, search strings (strategies) and modify as needed for each site that will be searched.
   - Consider spellings, and differences in Canadian, American, or British English.
   - Statement of the search terms used and any search restrictions.
   - If possible, save longer search sets by using saved search features or store searches in a separate Word document.
6. Perform some test or pilot searches at key websites; record number of hits; compare with other resources searched.
7. Indicate when (dates) the searches will be conducted in a timeline, and any time, language, publication (or other) limits to be used.
8. Indicate total number of results found; number of results retrieved and saved.
9. Use comments column at far right to note unique search features of resources/websites.
10. Indicate how citations will be imported into Endnote, RefWorks, Mendeley or other bibliographic referencing software such as CSV or tabbed format or RIS.

SOURCE: Author's compilation.

potentially relevant records as possible) and specificity (the relevant records found as a proportion of all records examined) (Shaw et al. 2004). Ultimately, searching for grey literature will involve some compromise in sensitivity but should still aim to identify as much of the available relevant evidence as possible (Boland, Cherry, and Dickson 2013).

Planning systematic literature searches and managing the results are distinguishing characteristics of research syntheses (Shaw et al. 2004). Broadly speaking, the searcher will be asked to use reproducible scientific methods in the search, as increasingly required by grant-funding agencies (Hickner, Friese, and Irwin 2011). Rothstein and Hopewell have said that "the soundness of a research synthesis depends heavily on the degree to which the search for relevant studies is thorough, systematic, unbiased, transparent, and clearly documented" (2009, 105). How to manage hundreds, in some cases thousands, of citations and full-text documents will also need to be determined (Booth, Sutton, and Papaioannou 2016).

Two reference managers used to manage citations in research are Endnote, which offers a free basic package and a subscription-based version, and Mendeley, which is free to use, though a fee-based premium model is also available. Mendeley can be used to download batch records by dragging folders of pdfs directly into a library of references (Saleh, Ratajeski, and Bertolet 2014). A handy way to search across documents is programmed into Mendeley, as are ways to create research groups to share unpublished and published materials with others. Mendeley is able to create individual citation records for each file "dropped" into the system, and offers screen

scraping facilities and metadata extraction for grey literature (Price 2016).

Other reference management tools are RefWorks, a subscription-based tool, and Zotero, which is open source with some novel features such as website archiving. The RefGrab-It tool in RefWorks, a handy drag-and-drop bookmarklet, aids the searcher in capturing metadata from websites; Endnote has a capture references feature and bookmarklet to perform a similar function. Two obstacles in managing grey literature are the nonstandard search interfaces that are encountered and dealing with citation inaccuracies (Kratochvíl 2016). Websites that include grey literature may both be difficult to search and present a range of challenges in citation management such as exporting records and duplicate citations.

Duplicate (and triplicate or more) citations can be a challenge in the review synthesis and make overall data and citation management more difficult (Adams et al. 2016). The challenge is reinforced by the multiple versions of documents encountered in grey literature searching. Some search engines, such as Yahoo, allow the development of APIs to search and remove duplicates from web searches (Bellefontaine and Lee 2014). Endnote uses de-duplication algorithms that can be adjusted to assist in efficient search planning and capture of citations (Bramer et al. 2016). Neal Haddaway and his colleagues discuss some of the advantages and disadvantages of the many available methods to extract full citations (and citation metadata) from websites by using citation management and analysis software versus web-crawling software (2017). For example, they discuss the software tool Import.io (https://import.io) as an alternative web-based platform for extracting data from websites to document searches. Import.io allows the user to define precisely what information should be extracted from a website and can be used across a range of search engines, though it can be a difficult tool to implement without adequate software knowledge.

6.4.3 Driving the Process

The first stage in the research synthesis is to adequately refine the topic and convert it into an answerable question (Booth, Sutton, and Papaioannou 2016). Researchers developing ideas for a research synthesis come equipped with fuzzy or grey questions (McKimmie and Szurmak 2002). Framing a question is thus the driving force behind evidence-based practice (Eldredge 2000); further, it provides an opportunity to clarify the aims of research.

In several evidence-based practices, researchers use frameworks to structure their research questions. These frameworks or mnemonics help to define topics, break them down into searchable parts, and determine whether they are feasible for investigation. Two frameworks are widely used. One is PICO (patient, intervention, comparison, outcome), which has been used in medicine to break down clinical questions (Cooke, Smith, and Booth 2012; Methley et al. 2014). Another is ECLIPSE (expectation, client, location, impact, professionals, service), which was developed to address questions in health policy and management (Wildridge and Bell 2002). Some researchers find it helpful to add study design (S) to the PICO framework, but both frameworks are flexible enough to be adapted accordingly. A good third option, SPICE (setting, perspective, intervention, comparison, evaluation), was developed to answer questions in the social sciences (Booth, Sutton, and Papaioannou 2016). SPICE provides librarians with a useful framework given the S (setting) and perspective (P) parts in the model, which are useful in focusing research questions in the information professions (for examples of the use of PICO, ECLIPSE, and SPICE, see table 5.3).

After the review parameters are established, the searcher should take a closer look at how to plan and document the actual searches (see table 6.4). The planning document, or search grid, is a diary of the searches performed that consists of the names of databases, websites, and online resources as well as any key journals to be searched by hand. The inclusion and exclusion criteria for the research project should be explicit, and the search should be developed accordingly. The planning phase relies on an exhaustive listing of concepts from the review question and is concept centered; all key concepts and the ways they can be expressed in natural language are the foundation for searching (Booth, Sutton, and Papaioannou 2016). Key terms and concepts determine, to a large extent, the organizing framework for the review (Webster and Watson 2002).

It should be determined in early pre-searches ("orienting the searcher to the literature") that a good-quality review does not already exist on the topic. If it does, the question is whether it needs to be updated or a completely new review is needed (Booth, Sutton, and Papaioannou 2016). To determine whether any systematic reviews are under way, check PROSPERO and journals such as *Systematic Reviews*, which publishes systematic review protocols.

Examining the methods of already found, relevant papers is useful to see whether reviewers were explicit

Table 6.4 Checklist to Document Grey Literature Searches

- ☑ Documentation includes a clearly stated question and scope of the research project.
- ☐ Description of the rationale and methods used to develop the search strategy are included; why some grey literature sources were searched (and others were not).
- ☐ Topic is broken down into workable components; search terms, key concepts, keywords; use of truncation, stemming, wildcards, and variants are listed.
- ☐ Boolean operators, proximity operators, and search restrictions are stated (anything you wish to exclude from the search).
- ☐ Variant spellings and differences in Canadian, American, or British English are noted.
- ☐ Screenshots and print-outs of detailed searches are added and longer search sets are saved in a Word document.
- ☐ Authors and experts, and their affiliated organizations, are noted and searched specifically.
- ☐ Websites and online resources were searched and in priority order (some duplication of the standard bibliographic databases was required).
- ☐ Online resources, affiliated organizations, and web addresses are listed.
- ☐ When (date) searches were conducted in each resource, and number of hits, are indicated; end date of searches (for example, 2010 through 2017) is noted.
- ☐ Comments are included to note when resources and websites were last updated in case you need to revisit and update your searches.
- ☐ (Un)indexed journals that will be hand searched are indicated; a record of experts and organizations contacted; other nondatabase methods of searching are listed.
- ☐ Special search techniques such as reference harvesting and citation searching are indicated.

SOURCE: Author's compilation.

NOTE: For more detailed information on the documentation of searches, see Chojecki and Tjosvold (2016). http://vortal.htai.org/sites/default/files/Elements%20of%20the%20Search%20to%20be%20Reported%20Table%201.pdf (accessed January 19, 2019).

about their search strategies. Gold standard searches or "pearls" and a set of relevant papers can be used to create more specific searches as necessary (Bellefontaine and Lee 2014; Hinde and Spackman 2015). Bibliographic databases such as Cochrane, MEDLINE, PsycINFO, or Sociological Abstracts are useful in locating published syntheses and searching across a body of literature (Booth, Sutton, and Papaioannou 2016). Some of the larger bibliographic databases such as Academic Search Premier and ERIC are useful in finding well-cited and seminal papers. Resources such as Google Scholar and Wikipedia can be useful to refine topics as searches are conducted across an interdisciplinary pool of papers and grey literature (Bramer et al. 2016; Spencer, Krige, and Nair 2014).

6.5 IDENTIFYING KEY SOURCES

Before reading this section, readers should be acquainted with the key information sources, monographs, databases, and websites in their disciplines and relevant resources that pertain to their research questions. If readers are unaware of these resources, academic libraries can create subject guides and organized lists of reference sources and bibliographic databases by topic for this purpose (for example, see http://guides.library.harvard.edu/sb.php and http://guides.library.ubc.ca/). To appreciate the scope of grey literature searching, see the example in appendix 1 ("Acupuncture in the Management of Drug and Alcohol Dependence") where a hypothetical research question is posed and a range of sources of information to search are identified.

Initial searches in the major bibliographic databases and library catalogs will provide searchers with a general idea of the overall size and quality of monographs and journal literature for a topic. Narrowing the topic will invariably be necessary for some topics, but it must first be ascertained that the available body of evidence is substantial enough to address the research question fully. Scoping the topic is part of establishing the feasibility of the review and determining the quantity and quality of literature to answer a question (Booth 2016). This early searching should give the searcher some general idea of

grey literature suppliers and producers for the topic. The searcher's findings can be discussed with the review synthesis team before proceeding.

After the scoping is completed, key websites and sources of information can be investigated in more detail and the most relevant selected (Brien et al. 2010). The searcher selects resources based on a combination of factors related to the research question. The objective is to ensure that selected resources provide extensive coverage of a topic as well as access to the types of publications where relevant evidence is likely to be published. Asking the review team for suggestions will assist in identifying these resources. Locating subject-specific databases, websites, and online resources will also involve some exploration (Booth, Sutton, and Papaioannou 2016).

The term *website* in grey literature searching is used broadly to help identify search engines and websites at organizations and government agencies, institutional repositories, research registries, and academic libraries (Stansfield, Dickson, and Bangpan 2016). Certain entries in Wikipedia can be used as starting points to locate associations and experts in different disciplines (Spencer, Krige and Nair 2014). Deciding which grey literature websites should be listed in the planning document is informed by some knowledge of their content strengths. Many resources can be identified by using guides and structured checklists (see table 6.2), pinpoint Google search techniques (see appendix 2) and consultations with subject experts (Mahood, Van Eerd, and Irvin 2014).

Grey literature searching benefits from clear structure but may require some flexibility and accepting grey literature as it appears in the search, even if this appears to be unsystematic (Booth 2016). Serendipitous discovery methods such as cited reference searching and browsing should be viewed as part of the search strategy (Brien et al. 2010; Hartling et al. 2016). Many searchers use multiple approaches to grey literature searching to narrow down and identify appropriate sources of information and websites (Godin et al. 2015). Another way is to "pre-specify exactly what forms of literature are being sought and then [to] select sources and strategies for these specific forms" (Booth 2016, 7).

For example, searchers can create lists of publications that are important to the review, such as reviews, clinical trials, and policy reports, and which organizations are likely to produce, such as the Cochrane Collaboration, WHOICTRP, and the Canadian Centre for Policy Alternatives (Balshem et al. 2013; Booth 2016). Another way is to identify platforms or catalogs to search, such as specialist theses repositories, clinical trial registries, or library catalogs as well as relevant websites such as Theses Canada and Networked Digital Library of Theses and Dissertations where these documents are located (see table 6.5). To address local or regional issues in a research synthesis, Google searches within geographic regions can bolster other pinpoint strategies (Godin et al. 2015; see also appendix 2). Lists of grey literature producers, such as those curated by the New York Academy of Medicine, are useful in building lists of credible organizations and associations (see table 6.6).

Some documents considered partly grey are available in key open-access repositories such as PubMedCentral and BioMedCentral. Other important paper repositories such as arXiv and the Social Sciences Research Network aim to serve specific disciplines in the sciences and social sciences (Booth, Sutton, and Papaioannou 2016). When looking for grey literature, the searcher should both investigate well-respected universities or research centers known for conducting research pertaining to the discipline or topic, and conduct searches for materials in their library catalogs and within their institutional repositories. Librarians and searchers will want to check libraries locally for materials as well as those in larger unified catalogs such as WorldCat and AMICUS. Identifying resources may include a combination of electronic databases, national registries of research, websites, and conference proceedings (Booth, Sutton, and Papaioannou 2016).

Relevant publications are found by visiting key producers' websites (Booth, Sutton, and Papaioannou 2016). Nonprofit organizations as well as think tanks, businesses, and foundations produce a great deal of grey literature (Schöpfel and Farace 2011). Navigating the organization's website and seeking menu-based options labeled "publications," "reports," or "documents" will be important. Look for a website's site map to determine whether links to full text are provided. Advanced search screens in Google and Yahoo can be used to perform site searches for document types such as portable digital files, PowerPoints, and papers delivered at conferences (appendix 2). When websites do not provide a good search facility or access to full documents to nonmembers (for example, conference websites), the searcher should consider asking the site manager for assistance or speaking to a librarian about an interlibrary loan.

6.5.1 Finding and Mapping Resources

Finding and mapping a set of resources to a particular question involves browsing (or "surfing") for information;

Table 6.5 Identifying Resources by Method, Tools, and Types of Grey Literature

Method	Tools (Examples)	Used to Find
Contacting experts	Email, Google Scholar, Twitter, Facebook, Academia.edu, university websites	Unpublished or in-process grey literature, information, data
Clinical trial registries	WHO International Clinical Trials Registry Platform	Some unpublished research data, in-process studies and trials
Grey literature repositories	OpenGrey, local university and subject-based repositories	Some unpublished or in-process grey literature, information, data
Hand searching	The searcher's hands (manual methods), computer browsing screen by screen	Core journals, conference proceedings, open access journals not indexed in major bibliographic databases
Library catalogs	U.S. Library of Congress, U.S. National Library of Medicine, local university libraries, national libraries and union catalogs such as WorldCat, AMICUS	Some grey literature and book chapters; unpublished or in-process grey literature, information, data
Reference list scanning	The searcher's eye (harvest references from relevant documents and bibliographies)	Some unpublished or in-process grey literature, information, data
Theses and institutional repositories	Theses Canada; PQDT Open, NTDLD, university repositories	Dissertations, theses, other academic outputs such as conference abstracts, posters, presentations
Web (general) searching	Google, Google Scholar, Bing, Yahoo	A range of grey literature, information, facts, or data

SOURCE: Author's tabulation based on McArthur 2016.

once websites are determined relevant, they should be listed on the *planning document* (Booth, Sutton, and Papaioannou 2016). The resources listed in table 6.2 will assist in locating unpublished literature types such as dissertations, theses, and conference proceedings as well as key organizations, directories, and lists; they provide an idea about how to find journals, subject repositories in physics, chemistry, medicine, and psychology and institutional repositories at universities and research centers.

Refining and mapping topics to relevant resources is done in consultation with the review team (Brien et al. 2010). To aid in listing resources in some priority order, librarians can use tiered-rating systems and Likert scales to evaluate online resources from somewhat important to highly important. Performing basic or simple test or "pilot" searches at top-tier or highly valued websites will provide a way to scan for materials and select search terms (Aromataris and Pearson 2014). As websites are considered irrelevant, they should be noted as such on the planning grid.

To address the potential for database and or searcher bias, searchers should aim to learn more about relevant resources beyond those purchased by their local libraries (Finfgeld-Connett and Johnson 2013). Some vendors of fee-based databases provide free trials of resources to be searched for a short period of time. Librarians possess a wide range of knowledge and of information sources and can help build lists of possible resources (Booth, Sutton, and Papaioannou 2016). Many librarians will know about a range of resources and the extent of coverage within them (Rosenthal 1994). Some newer interdisciplinary topics will require extensive searches because the literature is widely dispersed across resources and databases (Bonato 2018).

Newly located grey literature should be evaluated the way other resources are (Tyndall 2010). Deciding which online sources and producers are credible is a challenge but an evaluation can be made based on authority, sponsorship, place of the host organization, and affiliated agencies such as the sponsoring government or nongovernmental agency. Tyndall's AACODS checklist is helpful in evaluating individual grey literature papers (2010); the National Library of Medicine in the United States offers a tutorial on how to evaluate websites in health and medicine, many

Table 6.6 Locating Credible Sources and Producers of Grey Literature

Searching for grey literature requires attention to the types of organizations likely to publish or investigate the same (or similar) research questions that you are trying to investigate. This list of grey literature producers illustrates the range of organizations producing grey literature around the world.

For more detail, consult GreySource - Index to Web based Resources in Grey Literature. This index is an international list organized by topics such as agriculture, environment, humanities and social science, biology and medicine, earth science, engineering, and more. The New York Academy of Medicine also keeps comprehensive lists of Grey Literature-Producing Organizations.

Types of grey literature producers

- Academic and research institutions
- Associations (annual conference proceedings)
- Charitable foundations, organizations, and non-profits
- Conferences and congresses
- Government agencies, departments and committees (for example, municipal, provincial, national)
- Intelligence agencies (for example, domestic, military, national security)
- International organizations and agencies
- Libraries, museums, archives
  - Institutional repositories
  - Theses and dissertations databases
- Market research firms
- Multinational companies
- Non-governmental organizations (NGOs)
- Private companies and agencies
- Professional organizations
- Public institutions
- Research centers and institutes
- Scientific laboratories
- Special interest groups / societies / foundations
- Statistical organizations
- Think tanks and policy institutes
- Universities and colleges

SOURCE: Author's tabulation.

of which principles will apply across disciplines (https://medlineplus.gov/webeval/webeval.html).

Some resources reported in languages other than English may be considered for certain topics and to avoid linguistic bias, though what documents in foreign languages add to the research synthesis is not always clear (Song et al. 2010). One group of researchers found that use of non-Western databases was essential because their topic—on faith-based interventions in mental health—encompassed many religions and languages (Wright, Cotrell, and Mir 2014). Some topics in public policy and health technology assessment require the retrieval of papers from countries where policy documents such as drug monographs, government white papers, and research reports are published in languages other than English. Google Translate (a free web-based translation tool) can provide rough translations but will not be satisfactory for many topics. Searching Google in countries such as Canada, the United Kingdom, or Australia can assist in locating grey literature among a set of relevant national internet domains. To assist in locating these sites, see the Wikipedia list of Google domains (https://en.wikipedia.org/wiki/List_of_Google_domains).

6.6. DEVELOPING SEARCH STRATEGIES

Specialized search techniques can help ensure that a search is highly sensitive. For a good review of a topic, searches should never be confined to any single approach, set of documents, or geographic area. Searching many sources, including a range of databases, increases the likelihood of retrieving relevant records (Avenell, Handoll, and Grant 2001; Grindlay, Brennan, and Dean 2012). Between thirty

and seventy websites may have to be searched, but the precise number of top sites may be lower (Pullin et al. 2013). In one study, an investment of effort was required across multiple web sources and yielded twenty-one grey reports—12.5 percent of the total publications in the review (Saleh, Ratajeski, and Bertolet 2014).

6.6.1 Competencies in Searching

Searching for grey literature should account for how documents are likely to be described; any free-text terms that appear in the title or abstracts of relevant items should be reflected in the search strings (Booth 2016). Given the limited number of words in a document or bibliographic record, authors will not capture every detail of their topic or methods in their descriptions (Whiting et al. 2008). Most websites provide a search field of some kind in which to enter search terms; in some cases, advanced search functions will also be an option. Many websites will allow limits by publication date, recently uploaded, popular documents on the site, and geographic region. Many website search engines cannot perform the complex queries performed in standard bibliographic databases. For example, some sites will not offer batch downloading, saved searches, or an alerting service as new materials are entered into the database.

Some search tools have prescribed limits on the number of search terms that can be entered. To search thoroughly, the searcher should aim to include the most relevant keywords, phrases, and synonyms on the planning grid. Google Scholar, for example, has a limit of 256 characters (Bramer, Giustini, and Kramer 2016). It may be important to conduct specific field searching (searching in the "title" or author fields) to improve precision, and keep detailed documentation on what is working or not. To increase retrieval, combine variant terms and synonyms using OR. If necessary, the searcher should separate synonyms and use them separately or in combination and find out whether adjacency searching is possible on a given website by using the NEAR or ADJ commands or by looking at the help pages or frequently-asked questions.

Search queries will need to be simplified at websites because the full strategies used in the bibliographic databases are not always possible to replicate. Here, Occam's razor (the law of parsimony) may be worth remembering: simple, short search strategies may be just as effective as more complex ones for identifying relevant papers (Finfgeld-Connett and Johnson 2013). For example, in a study of patients' perceptions of living with a chronic health problem, researchers determined that three broad terms were as effective as more complex strategies (Flemming and Briggs 2007). Search terms should be entered in priority order and in varying combinations to search exhaustively.

In 2010, Jessie McGowan and her colleagues published a tool to evaluate *pre-planned* searches in seven categories: translation of the search strategy; use of Boolean and proximity operators and subject headings; use of natural language or free-text; spelling syntax and line numbers; limits and filters and adapting the strategy to each new database. They produced a checklist of questions (and a 2016 update, in which one category was removed) to guide peer review of electronic search strategies (PRESS) (McGowan et al. 2010, 2016b). Both are recommended as search strategies are refined; further, reflect on the ten key competencies in grey literature searching and consider the important skills required from the searcher pertaining to each (table 6.7).

Table 6.7 Ten Key Competencies in Searching for Grey Literature

1. Define grey literature and explain its characteristics vis-à-vis "published literature"
2. Develop grey literature search strategies and approaches that are reproducible
3. Identify key resources and websites to locate the most relevant grey literature
4. Undertake comprehensive searches using diverse tools, engines, and resources
5. Use hand searching, snowballing, and citation searching to increase sensitivity
6. Document and report the search process to meet required scientific standards
7. Manage citations using Endnote, Mendeley, RefWorks, or related tools
8. Budget for and estimate the costs associated with major grey literature searching
9. Understand review synthesis methods and communicate the value of grey literature to the research synthesis team (that is, to help minimize publication and linguistic bias, to avoid distorted view of literature)
10. Develop strategies to communicate effectively with the research team throughout the entire research process

SOURCE: Author's tabulation.

6.6.2 Google Scholar in Searching

Google Scholar has an important place in grey literature searching (Haddaway et al. 2015; Bramer, Giustini, and Kramer 2016) especially in combination with bibliographic database searching (Bramer et al. 2017). Its initial role may be in searching broadly across the web to verify citations and inform the information-gathering (pre-search) phase (Bramer et al. 2013). Google Scholar can be used to locate highly cited and free documents rather than as a tool to conduct literature reviews (Martín-Martín, Orduna-Malea, Harzing et al. 2016; Bramer et al. 2013). Some problems with Google Scholar are due to its lack of search functionality and transparency of coverage (Sibbald et al. 2015; Bramer, Giustini, and Kramer 2016). Another problem is that Google Scholar searches cannot be rerun reliably, which presents a challenge for reporting and replicability (Bramer et al. 2013).

Google Scholar is thought to index "moderate amounts" of grey literature (Haddaway et al. 2015). Some researchers avoid scrolling through thousands of Google Scholar results by limiting themselves to the first two or three hundred results (Haddaway et al. 2015). To capture relevant papers, some researchers limit screening to the first hundred results (Godin et al. 2015). Others opt for as many as the first five hundred results in the hopes of finding relevant papers (Bellefontaine and Lee 2014). Only the first thousand results of any search can be viewed in Google Scholar even when the search engine indicates its search results are much higher (Bramer, Giustini, and Kramer 2016).

Google Scholar does not offer saved searching, search histories, or expert search commands. It does, however, offer an advanced search for guided use of Boolean commands and field searching (Bellefontaine and Lee 2014). Search engines that search the deep web show some potential in locating grey literature (Speirs 2013; Rudesill, Caverlee, and Sui 2015; Olson 2013). All major web search engines—such as Google, Bing, and Yahoo—cover less than 20 percent of the web in their results and little of the deep web, if at all. Each search engine therefore has limitations in grey literature searching (Bellefontaine and Lee 2014).

6.6.3 Hand Searching, Harvesting, and Altmetrics

Hand searching refers to the manual search of print journals to locate relevant articles and citations that have been missed in electronic searching (Booth, Sutton, and Papaioannou 2016). According to the *Cochrane Handbook*, "handsearching involves a manual page-by-page examination of the entire contents of a journal issue or conference proceedings to identify all eligible reports" (Lefebvre, Manheimer, and Glanville 2011, 6.2.2.1). The need to search issue by issue or page by page in key journals may need to be taken into account in search planning (Rothstein and Hopewell 2009).

Hand searching is recommended even when the journals identified for hand searching are indexed by the traditional biomedical databases (Helmer et al. 2001). Some studies comparing electronic searching with hand searching found that little benefit was to be gained from hand searching (Rothstein and Hopewell 2009). It has been shown to be effective in finding additional papers in some studies but is time consuming (Armstrong et al. 2005). In an era when journals are "born digital" and digitized back to issue number one, it is fair to question whether hand searching should be adapted to online workflows or kept as a manual process, though this should be tested empirically.

A search filter, also known as a search hedge, consists of both controlled and uncontrolled terms that can be used to increase overall yield for a set of possibly relevant documents that might not otherwise be found by hand. Similarly, a search filter is a combination of terms used in a single nested search query such as "teens OR teenagers OR adolescents" (Haynes et al. 1994). Search filters are thus pretested strategies designed to identify desired concepts from vast amounts of literature indexed in the traditional bibliographic databases (Lee et al. 2012). Customized filters can be used to locate papers by study design, by population, or within specific geographic areas. Search filters are useful when many synonymous terms, spelling variations, and different languages are part of the topic. Certain websites may not allow the use of search filters but each term or set of terms that make up the filter can be searched separately.

Snowballing (also known as reference harvesting, backward searching, treeing through references, the ancestry approach or pearling) is a search technique used to track down references in relevant documents (Greenhalgh and Peacock 2005; Booth, Sutton, and Papaioannou 2016). The best starting point for reference harvesting is to identify a small group of relevant documents for review because they are more likely to cite papers similar in nature (Hinde and Spackman 2015). Searching reference lists is almost universally recommended as an effective way to

identify supplementary relevant materials. In one study, snowballing found 29.2 percent of all items retrieved for a review (Helmer et al. 2001). Snowballing is sometimes found to be more effective than other forms of searching (Hinde and Spackman 2015).

A few studies provide evidence of the effectiveness of reference harvesting. Trisha Greenhalgh and Richard Peacock searched the reference lists of previously identified papers that yielded the most (12 percent) of useful additional studies: reference harvesting proved effective in finding one useful paper for every fifteen minutes of searching (2005). Scanning references identified a quarter of the potentially useful studies identified by means other than searching of electronic databases (Helmer et al. 2001).

Digital forms of reference searching and citation harvesting can increase the total number of papers. In PubMed, for example, the "Related Citations" feature leads to articles similar in subject content to the one viewed. In Google Scholar, the "Cited by" feature will point to potential grey literature among the references cited. Bibliographic coupling or co-citations in Scopus and Web of Science can be helpful in leading researchers to grey literature. One study finds that cited reference searching may actually yield more relevant studies than performing more database searching (Hartling et al. 2016). Citation searching is widely viewed as a useful technique and may reduce overall search burden (Linder et al. 2015; Belter 2016).

Some bibliographic databases such as ERIC and CINAHL point searchers to possibly relevant materials based on the materials already viewed through features such as "Related articles" and "Cited within this database." Innovative harvesting in some of the standard bibliographic databases now include altmetrics (nontraditional metrics), which refer to how many mentions a given paper has received, social exchanges about it on blogs, Twitter, Wikipedia, and Facebook, and the impact of these activities (Lindsay 2016). Altmetric.com, PlumAnalytics, and ImpactStory each currently offer this service within their own websites or as added features on various search engines and websites (Lindsay 2016). A related altmetrics trend measures the impact of individual articles by usage such as number of page views, citations, and downloads.

6.6.4 Identifying Experts to Increase Recall

Identifying experts in various disciplines is a proven method to increase search recall (Booth 2016). Experts are often able to identify specialist websites with relative ease and professional confidence (Adams, Smart, and Huff 2017). The number of content experts that should be contacted will vary depending on the scope of a project and its timelines (Godin et al. 2015). Asking colleagues and key experts about unpublished studies will lead to papers that even extensive database searching may fail to identify; requesting datasets directly from authors is also useful (Schroll, Bero, and Gøtzsche 2013). Author profiles can be created on Google Scholar ("Google Scholar Citations") and may be helpful in tracking down experts and following their publication metrics (Giustini 2016). Informal channels of communication can sometimes be the only way to obtain unpublished datasets. Formal emails and letters of request can also help in identifying unpublished studies.

Online social sites, discussion forums, and blogs authored by scholars may provide useful information about experts and the most viewed or downloaded papers (Thelwall and Kousha 2015). Mendeley, Academia.edu and ResearchGate are social networks catering to academics who use them to create profiles and list their publications. They also provide platforms for researchers to load all kinds of published and unpublished papers (Bornmann and Haunschild 2015; Citrome 2014; Thelwall and Kousha 2015; Martín-Martín, Orduna-Malea, Ayllón et al. 2016). Academic social network sites are potentially rich stores of grey literature and social discussion about unpublished research. Many experts load their preprints to get feedback from peers, to interact, and to build their social networks (Martín-Martín, Orduna-Malea, Ayllón et al. 2016). The citation metrics gathered at the sites (as well as via Twitter) provide information about author impact factors, readership patterns, and whose research is being cited and by whom (Thelwall and Kousha 2015). Invariably, these sites are useful in finding experts, contact information and grey papers and other materials.

6.6.5 When to Stop Searching

At some point in planning (and perhaps during the actual searches, or during an evaluation), it is wise to determine when to stop searching. Some researchers determine this stopping point based on an estimate of how close the searcher has come to finding everything relevant (Kastner et al. 2009). A priori stopping rules, and identifying a saturation point, are among the solutions proposed to help searchers identify a stopping point (Kastner, Straus, McKibbon et al. 2009; Booth 2010). Reasonable limits

should be placed on how many resources can be effectively searched but currently there seems to be some disagreement about when that point is reached (Hartling et al. 2016). Stopping rules decided in advance of searching can be used to rationalize the decision to stop. For example, there is no inherent value in continuing literature searches unless they improve sensitivity (Kastner et al. 2009; Booth 2010).

Regardless of the type of research undertaken, ending the search for grey literature can be based on the judgment of the searcher; the key is not to terminate a search too early. An attempt should be made to ensure that decisions to stop searching are justified given the needs of the review (Booth 2010). In some cases, it is impossible to know what has not been found; accepting that no single review will be informative forever may be justification enough to stop searching (Booth 2010). In the end, the decision to stop will need to be documented and explained within the context of the review itself (Finfgeld-Connett and Johnson 2013). In the case of knowledge-building and theory-generating reviews, the key is saturation of concepts and the full explication of interrelationships among them (Strauss and Corbin 1998).

6.7 RECORDING AND REPORTING

Some researchers devise a robust system of recording searches to fit the individual project, relying on checklists, planning grids, Excel spreadsheets, or a combination of these (Godin et al. 2015; Chojecki and Tjosvold 2016). Using Excel to record searches performed on websites is not seamless and may require manual editing (Godin et al. 2015). To aid in accurate reporting, some searchers create Word documents and take screenshots of live searches, indicating the date and time when searches were performed and by whom.

The validity of a review can be evaluated in part on how explicit the searches for grey literature were (Hopewell, Clark, and Mallett 2005). Transparent and detailed reporting ensures that searches are reproducible. Searchers are required to provide enough detail to enable searches to be repeated later, tested, and updated as necessary. In the internet era, several variables limit the reproducibility of web searching due to the unpredictability of tools and ongoing changes to web content and addresses (Stansfield, Dickson, and Bangpan 2016; Briscoe 2015).

The PRISMA guidelines are explicit and helpful about recording and reporting searches for systematic reviews (Liberati et al. 2009), and can be applied to a grey literature searching (Godin et al. 2015). PRISMA stipulates that the following should be included: a description of all information sources in the searches, the name or names of those conducting the searches, the date the searches were performed, and a full search strategy of at least one database including all search terms and combinations (Liberati et al. 2009).

Claire Stansfield, Kelly Dickson, and Mukdarut Bangpan explore some of the challenges of conducting website searches and emphasize systematic but practical record keeping (2016). Key information about search methods, sources, search queries, and results should be recorded (McArthur 2016). Web searching should be reported to an extent that search strategies are transparent and reproducible; the aim is to include complete, detailed search strategies (which in some instances can be copied and pasted) and the number of records retrieved. Other details such as hand searching, contact with experts, reference lists, and citation searching should also be included in planning documents.

6.8 CONCLUSION

This chapter has discussed the systematic retrieval of grey literature in support of the research synthesis and several current techniques to conduct it efficiently. The search for grey literature is a scientific undertaking and requires both adhering to a predetermined search plan and an awareness of a reliable set of strategies and practices to ensure that the search is thorough, systematic, unbiased, transparent, and clearly documented.

Whether a research synthesis requires an extensive and potentially costly search for grey literature is a question that depends on local resources and the goals of individual projects. The research synthesis team can determine the degree of comprehensiveness of grey literature searching taking into account the requirements of the review and its resources (Egger et al. 2003). To retrieve grey literature effectively when required—similar to what is conducted in standard bibliographic databases—structured checklists and research guides will help searchers retrieve grey literature and enhance the likelihood that the process will be efficient (McGowan 2016b; Vaska and Vaska 2016). Searches for the grey literature should be monitored and evaluated for efficiency throughout the process as important papers appear in the process and more targeted searches are performed. Above all, the searcher should be prepared to discover retrieval challenges that are encountered, especially those beyond the scope of topics covered in this chapter.

Expert searchers such as librarians and information specialists are critical contacts in systematic searching and in identifying sources of grey literature. The interpre-

tive services of a librarian is critical at other points. As well as offering help to improve the process and substance of individual search strategies, librarians can assist researchers in selecting software programs for citation management and data management planning (for the value of the information specialist in the systematic review, see Wade et al. 2006). Librarians provide supporting documentation for reviews, search tutorials, and subject guides (Vaska and Vaska 2016). They are indispensable when assistance is needed for document delivery or interlibrary loan requests.

Retrieving grey literature, information, and data are complex and time-consuming tasks but rewarding scientific activities in their own right. The search and retrieval of grey literature may be difficult to manage and document but achieving this starts with a plan guided by emerging protocols from the literature. With the rise of the web, open access to research and data reporting requirements at grant-funding agencies worldwide, the expectation is that grey literature, information and data will play key roles in the synthesis of research in the social, behavioral, and medical sciences well into the future.

Further research needs to identify better practices in grey literature searching. The requirements of systematic review searching can confound even the most well-informed searcher given the almost daily changes to search engines and how and where grey materials are produced. To ensure long-term access to and preservation of these materials, researchers should work with librarians on solving some of the problems associated with finding grey literature. Despite the rise of the internet (or because of it) and the unprecedented production of scientific information in the digital age, a lot of grey literature searching is overly inefficient and repetitive. Too much is left to serendipity even when the search is thoroughly planned and a librarian is consulted. The process of locating valuable, relevant grey papers among millions of disparate web documents would benefit enormously from better cataloging and preservation practices well before the searcher looks for them in the research synthesis.

6.9 APPENDIX

6.9.1 Appendix 1. Case Study: Identifying Key Resources

Acupuncture in the Management of Drug and Alcohol Dependence

"Is acupuncture effective in managing drug and alcohol dependence?" The goal is to find as many randomized controlled trials (RCTs) as possible, perform a systematic review of the literature and a subsequent quantitative, meta-analysis.

Keywords and Phrases It is advisable to organize keywords, wildcards, and combinations before searching. Most specialized databases will have different types of search interfaces and functionality, but be as systematic as possible. A range of possible keyword combinations for this topic include acupuncture, meridian, acupressure, electroacupuncture, shiatsu, drug*, polydrug*, substance, alcohol, beer, wine, spirits, tranquilize, tranquilizer, narcotic, opiate, solvent, inhalant, street drug*, prescri*, non-prescri*, nonprescri*, abuse*, use*, usin*, misus*, utliz*, utilis*, depend, addict, illegal, illicit, habit*, withdraw*, behavio*, abstinen*, abstain*, abstention, rehab, intox*, detox, dual, diagnosis, and disorder.

Other Considerations

Synonyms, such as adolescents, teens, youth

Acronyms, such as AUD (alcohol use disorder), DUI (driving under the influence)

Differences in terminology across national boundaries, such as liquor, spirits

Differences in spellings, such as tranquilizer and tranquiliser

Old and new terminology, such as addiction and dependence

Brand and generic names, such as hydromorphone and dilaudid

Lay and medical terminology such as drunk and alcohol-dependent

Major Bibliographic Databases and Search Engines This list is not exhaustive and is meant to provide a starting point for the published literature. In some cases, these resources will not be available at your institution and access to them will have to be worked out:

MEDLINE, PubMed, and Embase

Cochrane Database of Systematic Reviews

Cumulated index to nursing and allied health literature (CINAHL)

Google and Google Scholar

PsycINFO, Sociological Abstracts

PubMedCentral

ScienceDirect

Academic Search Complete

Agricola

AMED (Allied and Complementary Medicine Database)

Key Websites and Online Resources Many of these resources will help locate unpublished studies and papers. Contacting relevant organizations will help in discovering what websites, search engines, and online resources exist (special deep web databases, library catalogs not crawled by Google, and so on). Most websites now provide a jumping-off point for your searching and are increasingly sophisticated even if you cannot perform the kind of structured searching you do in the major bibliographic databases. If you are unfamiliar with the topic for the research synthesis and do not yet know which organizations exist in a given field, a number of subject guides and organizational directories are available to help focus search efforts and guide you along the way. Examples of relevant organizations include the following:

British Acupuncture Council http://www.acupuncture.org.uk/

Canadian Interdisciplinary Network for Complementary and Alternative Medicine Research (IN-CAM), http://www.incamresearch.ca

National Acupuncture Detoxification Association (NADA), http://www.acudetox.com

National Center for Complementary and Integrative Health (NCCIH), https://nccih.nih.gov/health/acupuncture

National Institute on Alcohol Abuse and Alcoholism (NIAAA), http://www.niaaa.nih.gov/

National Institute on Drug Abuse (NIDA), https://www.drugabuse.gov/

Specialist Website and Database Examples These websites and specialized resources will further increase the comprehensiveness of grey literature searches:

AcuTrials®, http://acutrials.ocom.edu/

Acubase, http://www.acubase.fr/

Acubriefs, http://acubriefs.blogspot.ca/

Ageline, https://www.ebscohost.com/academic/ageline

Canadian Centre on Substance Abuse (CCSA), http://www.ccsa.ca/Pages/default.aspx

Drug Database (Alcohol and other Drugs Council of Australia), https://www.informit.org/index-product-details/DRUG

Networked Digital Library of Theses and Dissertations (NDLTD), http://www.ndltd.org/

PEDro, https://www.pedro.org.au

Traditional Chinese Drug Database (TCDBASE), http://tcm.cz3.nus.edu.sg/group/tcm-id/tcmid_ns.asp

Databases in the area of acupuncture and Traditional Chinese Medicine (TCM) are numerous, such as the Chinese Technical Periodicals (VIP), Chinese Biomedical Literature Database (CBM) and China National Knowledge Infrastructure (CNKI). See, for example, http://caod.oriprobe.com/packages/TCM.htm.

Library and Union Catalogs Library catalogs and discovery layers in academic, special, and public libraries are excellent sources of grey literature. Catalogs both provide access to local and regional materials and inform researchers that they exist. Library catalogs are fertile sources for bibliographic verification and resource discovery in grey literature searching. Library catalogs are good sources of grey literature and often index books, dissertations, government, and technical reports, particularly if the authors are affiliated with the parent organization as scholars or researchers. The following are a few examples for the acupuncture topic:

AMICUS, http://amicus.collectionscanada.ca/aaweb/aalogine.htm

Drug Policy Alliance, Lindesmith Library, http://www.drugpolicy.org/resources-publications

Centre for Addiction and Mental Health Library, http://www.camh.ca/en/education/about/services/camh_library/Pages/camh_library.aspx

Your Local Library

National Research Council (NRC) library catalog, http://cat.cisti.nrc.ca/search

WorldCat, https://www.worldcat.org/

Repositories (pre-prints, protocols, registries, research "in progress") Here are a few examples for the acupuncture topic:

ClinicalTrials.gov, http://clinicaltrials.gov

Cog Prints, http://cogprints.org

Defense Technical Information Network, http://stinet.dtic.mil/dtic

Directory of Open Access Journals, http://www.doaj.org

Open Trials, http://opentrials.net

PROSPERO, University of York, National Institute of Health Research (UK), https://www.crd.york.ac.uk/prospero

PubMedCentral, http://pubmedcentral.gov

Social Science Research Network (SSRN), https://papers.ssrn.com/sol3/results.cfm

WHO International Clinical Trials Registry Platform, http://www.who.int/ictrp/en/

6.9.2 Appendix 2. Advanced Google Search Commands

Advanced Google search commands give you the ability to use more pinpoint search strategies to locate grey literature more efficiently. However, it may be necessary to repeat any Google searches because using the site: command may require double-checking at the website itself. To use the advanced search page at Google, see: https://www.google.ca/advanced_search.

Searching by	Description	Example
"Phrase"	Forces a specific word order	"alcohol dependency"
Specific "word"	Quotation marks around a word turns off synonyms and spell checking	"addictions"
Site or domain search:	Search within a particular website (gov.ca) or domain (.ca)	site:drugabuse.gov/ site:.gov inurl:.ca
Filetype:	Searches for a particular filetype	filetype:pdf,doc,pptx
Intitle:	Searches only in the title	intitle: "systematic review"

For more information on advanced searching on Google and Google Scholar, see the Google Advanced Power Searching page: http://www.powersearchingwithgoogle.com/course/aps/skills

6.10 REFERENCES

Adams, Jean, Frances C. Hillier-Brown, Helen J. Moore, Amelia A. Lake, Vera Araujo-Soares, Martin White, and Carolyn Summerbell. 2016. "Searching and Synthesising 'Grey Literature' and 'Grey Information' in Public Health: Critical Reflections on Three Case Studies." *Systematic Reviews* 5(1): 164.

Adams, Richard J., Palie Smart, and Anne Sigismund Huff. 2017. "Shades of Grey: Guidelines for Working with the Grey Literature in Systematic Reviews for Management and Organizational Studies." *International Journal of Management Reviews* 19(4): 432–54.

Alberani, Vilma, Paola De Castro Pietrangeli, and A. M. Mazza. 1990. "The Use of Grey Literature in Health Sciences: A Preliminary Survey." *Bulletin of the Medical Library Association* 78(4): 358–63.

Aloia, Danielle, and Robin Naughton. 2016. "Share# GreyLit: Using Social Media to Communicate Grey Literature." *The Grey Journal* 12(2).

Armstrong, Rebecca, Nicki Jackson, Jodie Doyle, Elizabeth Waters, and Faline Howes. 2005. "It's in Your Hands: The Value of Handsearching in Conducting Systematic Reviews of Public Health Interventions." *Journal of Public Health* 27(4): 388–91.

Aromataris, Edoardo, and Alan Pearson. 2014. "The Systematic Review: An Overview." *American Journal of Nursing* 114(3): 53–58.

Auger, Charles P. 1975. *Use of Reports Literature*. London: Butterworths.

———. 1998. *Information Sources in Grey Literature Guides to Information Sources,* 4th ed. London: Bowker Saur.

Augusto, Laurent, Mark Ronald Bakker, Christian Morel, Céline Meredieu, Pierre Trichet, Vincent Badeau, Dominque Arrouays, Claude Passard, David Achat, Anne Gallet-Budynek, Dominque Merzeau, Didier Cantaloupe, Mohamed Najar, Jacques Ranger. 2010. "Is 'Grey Literature' a Reliable Source of Data to Characterize Soils at the Scale of a Region? A Case Study in a Maritime Pine Forest in Southwestern France." *European Journal of Soil Science* 61(6): 807–22.

Avenell, Alison, Helen Handoll, and Adrian Grant. 2001. "Lessons for Search Strategies from a Systematic Review, in The Cochrane Library, of Nutritional Supplementation Trials in patients After Hip Fracture." *American Journal of Clinical Nutrition* 73(3): 505–10.

Balshem, Howard, Adrienne Stevens, Mohammed Ansari, Susan Norris, Devan Kansagara, Tatyana Shamliyan, Roger Chou, Mei Chung, David Moher, and Kay Dickersin. 2013. *Finding Grey Literature Evidence and Assessing for Outcome and Analysis Reporting Biases When Comparing Medical Interventions: AHRQ and the Effective Health Care Program*. Rockville, Md.: Agency for Healthcare Research and Quality.

Banks, Marcus. 2004. "Connections Between Open Access Publishing and Access to Gray Literature." *Journal of the Medical Library Association* 92(2): 164–66.

———. 2009. "Blog Posts and Tweets: The Next Frontier for Grey Literature." *The Grey Journal* 8(1): 54–59.

Bellefontaine, Sarah, and Catherine Lee. 2014. "Between Black and White: Examining Grey Literature in Meta-Analyses of Psychological Research." *Journal of Child & Family Studies* 23(8): 1378–88.

Belter, Christopher W. 2016. "Citation Analysis as a Literature Search Method for Systematic Reviews." *Journal of the Association for Information Science and Technology* 10.1002/asi.23605.

Benzies, Karen M., Shahirose Premji, K. Alix Hayden, and Karen Serrett. 2006. "State-of-the-Evidence Reviews: Advantages and Challenges of Including Grey Literature." *Worldviews on Evidence-Based Nursing* 3(2): 55–61.

Blackhall, Karen M. 2007. "Finding Studies for Inclusion in Systematic Reviews of Interventions for Injury Prevention the Importance of Grey and Unpublished Literature." *Injury Prevention* 13(5): 359.

Boekhorst, Albert K., Dominic J. Farace, and Jerry Frantzen. 2005. "Grey Literature Survey 2004: A Research Project Tracking Developments in the Field of Grey Literature." *The Grey Journal* 1(1): 41.

Boland, Angela, M. Gemma Cherry, and Rumona Dickson, eds. 2013. *Doing a Systematic Review: A Student's Guide*. Thousand Oaks, Calif.: Sage Publications.

Bonato, Sarah. 2018. *Searching the Grey Literature: A Handbook for Searching Reports, Working Papers, and Other Unpublished Research*. Lanham, Md.: Rowman & Littlefield.

Booth, Andrew. 2010. "How Much Searching Is Enough? Comprehensive Versus Optimal Retrieval for Technology Assessments." *International Journal of Technology Assessment in Health Care* 26(4): 431–35.

———. 2016. "Searching for Qualitative Research for Inclusion in Systematic Reviews: A Structured Methodological Review." *Systematic Reviews* 5(1): 1.

Booth, Andrew, Anthea Sutton, and Diana Papaioannou. 2016. *Systematic Approaches to a Successful Literature Review*. Thousand Oaks, Calif.: Sage Publications.

Bornmann, Lutz, and Robin Haunschild. 2015. "Which People Use Which Scientific Papers? An Evaluation of Data from F1000 and Mendeley." *Journal of Informetrics* 9(3): 477–87.

Bramer, Wichor M., Dean Giustini, Gerdien B. de Jonge, Leslie Holland, and Tanja Bekhuis. 2016. "De-Duplication of Database Search Results for Systematic Reviews in EndNote." *Journal of the Medical Library Association* 104(3): 240.

Bramer, Wichor M., Dean Giustini, and Bianca M. R. Kramer. 2016. "Comparing the Coverage, Recall, and Precision of Searches for 120 Systematic Reviews in Embase, MEDLINE, and Google Scholar: A Prospective Study." *Systematic Reviews* 5(1): 1.

Bramer, Wichor M., Dean Giustini, Bianca M. R. Kramer, and P. F. Anderson. 2013. "The Comparative Recall of Google Scholar versus PubMed in Identical Searches for Biomedical Systematic Reviews: A Review of Searches Used in Systematic Reviews." *Systematic Reviews* 2(1): 1.

Bramer, Wichor M., Melissa L. Rethlefsen, Jos Kleijnen, and Oscar H. Franco. 2017. "Optimal Database Combinations for Literature Searches in Systematic Reviews: A Prospective Exploratory Study." *Systematic Reviews* 6(1): 245.

Brien, Susan E., Diane L. Lorenzetti, Steven Lewis, James Kennedy, and William A. Ghali. 2010. "Overview of a Formal Scoping Review on Health System Report Cards." *Implementation Science* 5(1): 1.

Briscoe, Simon. 2015. "Web Searching for Systematic Reviews: A Case Study of Reporting Standards in the UK Health Technology Assessment Programme." *BMC Research Notes* 8 (April): 153.

Canada, Government of. 2018. "Draft Tri-Agency Statement of Principles on Digital Data Management for Consultation." Science.gc.ca. Last modified May 15, 2018. Accessed November 30, 2018. http://www.science.gc.ca/eic/site/063.nsf/eng/h_97610.html.

Childress, Eric, and Erik Jul. 2003. "Going Grey: Grey Literature and Metadata." *Journal of Internet Cataloging* 6(3): 3–6.

Chojecki, Dagmara, and Lisa Tjosvold. 2016. "Documenting and Reporting the Search Process." HTAi Vortal, December 20, 2016. Accessed November 30, 2018. http://vortal.htai.org/?q=node/1065.

Citrome, Leslie. 2014. "Beyond PubMed: Searching the Grey Literature for Clinical Trial Results." *Innovations in Clinical Neuroscience* 11(7–8): 42–46.

Conn, Vicki S., Jeffrey C. Valentine, Harris M. Cooper, and Marilyn J. Rantz. 2003. "Grey Literature in Meta-Analyses." *Nursing Research* 52(4): 256–61.

Cook, Alison M., Ilora G. Finlay, Adrian G. K. Edwards, Kerenza Hood, Irene J. Higginson, Danielle M. Goodwin, Charles E. Normand, and Hannah-Rose Douglas. 2001. "Efficiency of Searching the Grey Literature in Palliative Care." *Journal of Pain and Symptom Management* 22(3): 797–801.

Cooke, Alison, Debbie Smith, and Andrew Booth. 2012. "Beyond PICO: the SPIDER Tool for Qualitative Evidence Synthesis." *Qualitative Health Research* 22(10): 1435–43.

Coonin, B. 2003. "Grey Literature: An Annotated Bibliography." http://personal.ecu.edu/cooninb/Greyliterature.htm.

Cooper, Harris, and Larry V. Hedges. 2009. "Research Synthesis as a Scientific Process." In *The Handbook of Research Synthesis and Meta-Analysis*, edited by Harris Cooper, Larry V. Hedges, and Jeffrey C. Valentine. New York: Russell Sage Foundation.

Egger, Matthias, P. Juni, C. Bartlett, F. Holenstein, and J. Sterne. 2003. "How Important Are Comprehensive Literature Searches and the Assessment of Trial Quality in Systematic Reviews? Empirical Study." *Health Technology Assessment* 7(1):1–76.

Eldredge, Jonathan D. 2000. "Evidence-based Librarianship: Formulating EBL Questions." *Bibliotheca Medica Canadiana* 22(2): 74–77.

Enticott, Joanne, Kimberly Buck, and Frances Shawyer. 2018."Finding 'Hard to Find' Literature on Hard to Find Groups: A Novel Technique to Search Grey Literature on Refugees and Asylum Seekers." *International Journal of Methods in Psychiatric Research* 27(1): e1580.

Farrah, Kelly, and Monika Mierzwinski-Urban. 2019. "Almost Half of References in Reports on New and Emerging Non-drug Health Technologies Are Grey Literature." *Journal of the Medical Library Association: JMLA* 107(1): 43–48.

Finfgeld-Connett, Deborah, and E. Diane Johnson. 2013. "Literature Search Strategies for Conducting Knowledge-Building and Theory-Generating Qualitative Systematic Reviews." *Journal of Advanced Nursing* 69(1): 194–204.

Flemming, Kate, and Michelle Briggs. 2007. "Electronic Searching to Locate Qualitative Research: Evaluation of Three Strategies." *Journal of Advanced Nursing* 57(1): 95–100.

Gelfand, Julia M., and Daniel C. Tsang. 2015. "Data: Is It Grey, Maligned or Malignant?" *The Grey Journal* 11(1): 29–40.

Giustini, Dean. 2014. "Endorse the 'Pisa Declaration on Policy Development for Grey Literature.'" *Search Principle* (blog), August 12, 2014. Accessed November 30, 2018. https://blogs.ubc.ca/dean/2014/08/endorse-the-pisa-declaration-on-policy-development-for-grey-literature.

———. 2016. "Grey Literature." HLWIKI International, December 28, 2016. Accessed November 30, 2018, http://hlwiki.slais.ubc.ca/index.php/Grey_literature.

Godin, Katelyn, Jackie Stapleton, Sharon I. Kirkpatrick, Rhona M. Hanning, and Scott T. Leatherdale. 2015. "Applying Systematic Review Search Methods to the Grey Literature: A Case Study Examining Guidelines for School-Based Breakfast Programs in Canada." *Systematic Reviews* 4(1): 138.

Grant, Maria, and Andrew Booth. 2009. "A Typology of Reviews: An Analysis of 14 Review Types and Associated Methodologies." *Health Information and Libraries Journal* 26(2)*:* 91–108.

Greenhalgh, Trisha, and Richard Peacock. 2005. "Effectiveness and Efficiency of Search Methods in Systematic Reviews of Complex Evidence: Audit of Primary Sources." *British Medical Journal* 331(7524):1064–65.

Grindlay, Douglas J. C., Marnie L. Brennan, and Rachel S. Dean. 2012. "Searching the Veterinary Literature: A Comparison of the Coverage of Veterinary Journals by Nine Bibliographic Databases." *Journal of Veterinary Medicine Education* 39(4): 404–12.

Haddaway, Neal Robert, Alexandra Mary Collins, Deborah Coughlin, and Stuart Kirk. 2015. "The Role of Google Scholar in Evidence Reviews and Its Applicability to Grey Literature Searching." *PloS One* 10(9): e0138237.

———. 2017. "A Rapid Method to Increase Transparency and Efficiency in Web-Based Searches." *Environmental Evidence* 6(1): 1.

Halfpenny, Nicholas James Anthony, Joan Mary Quigley, Juliette Catherine Thompson, and David Alexander Scott. 2016. "Value and Usability of Unpublished Data Sources for Systematic Reviews and Network Meta-Analyses." *Evidence Based Medicine* 21(6): 208–13.

Hartling, Lisa, Robin Featherstone, Megan Nuspl, Kassi Shave, Donna M. Dryden, and Ben Vandermeer. 2017. "Grey Literature in Systematic Reviews: A Cross-Sectional Study of the Contribution of Non-English Reports, Unpublished Studies and Dissertations to the Results of Meta-Analyses in Child-Relevant Reviews." *BMC Medical Research Methodology* 17(1): 64.

Hartling, Lisa, Robin Featherstone, Megan Nuspl, Kassi Shave, Donna M. Dryden, and Ben Vandermeer. 2016. "The Contribution of Databases to the Results of Systematic Reviews: A Cross-Sectional Study." *BMC Medical Research Methodology* 16(1): 127.

Haynes, R. Brian, Nancy Wilczynski, K. Ann McKibbon, Cynthia J. Walker, and John C. Sinclair. 1994. "Developing Optimal Search Strategies for Detecting Clinically Sound

Studies in MEDLINE." *Journal of the American Medical Informatics Association* 1(6): 447–58.

Helmer, Diane, Isabelle Savoie, Carolyn Green, and Arminee Kazanjian. 2001. "Evidence-Based Practice: Extending the Search to Find Material for the Systematic Review." *Bulletin of the Medical Librarians Association* 89(4): 346–52.

Hickner, Andy, Christopher R. Friese, and Margaret Irwin. 2011. "Development and Testing of a Literature Search Protocol for Evidence Based Nursing: An Applied Student Learning Experience." *Evidence Based Library and Information Practice* 6(3): 28–39.

Hinde, Sebastian, and Eldon Spackman. 2015. "Bidirectional Citation Searching to Completion: An Exploration of Literature Searching Methods." *PharmacoEconomics* 33(1): 5–11.

Hopewell, Sally, Mike Clarke, and Sue Mallett. 2005. "Grey Literature and Systematic Reviews." *Publication Bias in Meta-analysis: Prevention, Assessment and Adjustments* (2005): 48–72.

Hopewell, Sally, Steve McDonald, Mike Clarke, and Matthias Egger. 2007. "Grey Literature in Meta-Analyses of Randomized Trials of Health Care Interventions." *Cochrane Database of Systematic Reviews* (2): MR000010. DOI: 10.1002/14651858.MR000010.pub3.

Institute of Medicine. 2011. "IOM Standard 3: Standards for Finding and Assessing Individual Studies." In *Finding What Works in Health Care: Standards for Systematic Reviews*, edited by J. Eden, L. Levit, A. Berg, and S. Morton. Washington, D.C.: National Academies Press.

Kastner, Monika, Sharon E. Straus, K. Ann McKibbon, and Charlie H. Goldsmith. 2009. "The Capture–Mark–Recapture Technique Can Be Used as a Stopping Rule When Searching in Systematic Reviews." *Journal of Clinical Epidemiology* 2(62): 149–57.

Kratochvíl, Jiří. 2016. "Comparison of the Accuracy of Bibliographical References Generated for Medical Citation Styles by EndNote, Mendeley, RefWorks and Zotero." *Journal of Academic Librarianship* 43(1): 57–66.

Kung, Janice Yu Chen, and Sandy Campbell. 2016. "What Not to Keep: Not All Data Has Future Research Value." *Journal of the Canadian Health Libraries Association* 37(2): 53–57.

Lawrence, Amanda. 2012. "Electronic Documents in a Print World: Grey Literature and the Internet." *Media International Australia* 143(1): 122–31.

Lee, Edwin, Maureen Dobbins, Kara DeCorby, Lyndsey McRae, Daiva Tirilis, and Heather Husson. 2012. "An Optimal Search Filter for Retrieving Systematic Reviews and Meta-Analyses." *BMC Medical Research Methodology* 12(1): 1.

Lefebvre, Carol, Eric Manheimer, and Julie Glanville. 2011. "Searching for Studies." In *Cochrane Handbook for Systematic Reviews of Interventions*, Version 5.1.0, edited by Julian Higgins and Sally Green. London: The Cochrane Collaboration.

Liberati, Alessandro, Douglas G. Altman, Jennifer Tetzlaff, Cynthia Mulrow, Peter C. Gøtzsche, John P. A. Ioannidis, Mike Clarke, P. J. Devereaux, Jos Kleijnen, and David Moher. 2009. "The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration." *Annals of Internal Medicine* 151(4): W65–94.

Linder, Suzanne, Geetanjali Kamath, Gregory Pratt, Smita S. Saraykar, and Robert J. Volk. 2015. "Citation Searches Are More Sensitive than Keyword Searches to Identify Studies Using Specific Measurement Instruments." *Journal of Clinical Epidemiology* 68(4): 412–17.

Lindsay, J. Michael. 2016. "PlumX from Plum Analytics: Not Just Altmetrics." *Journal of Electronic Resources in Medical Libraries* 13(1): 8–17.

Mahood, Quenby, Dwayne Van Eerd, and Emma Irvin. 2014. "Searching for Grey Literature for Systematic Reviews: Challenges and Benefits." *Research Synthesis Methods* 5(3):221–34.

Marsolek, Wanda, Kristen Cooper, Shannon Farrell, and Julia Kelly. 2018. "The Types, Frequencies, and Findability of Disciplinary Grey Literature within Prominent Subject Databases and Academic Institutional Repositories." *Journal of Librarianship and Scholarly Communication* 6(1): eP2200.

Martinez, Gerardo Sanchez, Eloise Williams, and Shwe Sin Yu. 2015. "The Economics of Health Damage and Adaptation to Climate Change in Europe: A Review of the Conventional and Grey Literature." *Climate* 3(3): 522–41.

Martín-Martín, Alberto, Enrique Orduna-Malea, Juan M. Ayllón, and Emilio Delgado López-Cózar. 2016. "The Counting House: Measuring Those Who Count. Presence of Bibliometrics, Scientometrics, Informetrics, Webometrics and Altmetrics in the Google Scholar Citations, ResearcherID, ResearchGate, Mendeley & Twitter." *EC3* working paper no. 21. Granada: Evaluación de la Ciencia y de la Comunicación Científica, Universidad de Granada and Universidad Politécnica de Valencia. DOI:10.13140/RG.2.1.4814.4402/1.

Martín-Martín, Alberto, Enrique Orduna-Malea, Anne Harzing, and Emilio Delgado López-Cózar. 2016. "Can We Use Google Scholar to Identify Highly-Cited Documents?" *Journal of Informetrics* 11(1): 152–63.

McArthur, Allison. 2016. "How to Find and Use Grey Literature for Scoping Reviews." Presentation. Public Health Ontario. Agency for Health Protection and Promotion. Accessed November 30, 2018. https://www.publichealthontario.ca/en/LearningAndDevelopment/EventPresentations/Grey_Literature_Scoping_Reviews_McArthur_2016.pdf.

McGowan, Jessie, Margaret Sampson, and Carol Lefebvre. 2010. "An Evidence Based Checklist for the Peer Review of Electronic Search Strategies (PRESS EBC)." *Evidence Based Library Information Practice* 5(1): 149–54.

McGowan, Jessie, Margaret Sampson, Douglas M. Salzwedel, Elise Cogo, Vicki Foerster, and Carol Lefebvre. 2016a. "PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement." *Journal of Clinical Epidemiology* 75 (July): 40–46.

———. 2016b. *PRESS Peer Review Electronic Search Strategies: 2015 Guideline Explanation and Elaboration (PRESS E&E)*. Ottawa: CADTH.

McKimmie, Tim, and Joanna Szurmak. 2002. "Beyond Grey Literature: How Grey Questions Can Drive Research." *Journal of Agricultural & Food Information* 4(2):71–79.

Mering, Margaret. 2015. "Preserving Electronic Scholarship for the Future: An Overview of LOCKSS, CLOCKSS, Portico, CHORUS, and the Keepers Registry." *Serials Review* 41(4): 260–65.

Methley, Abigail M., Stephen Campbell, Carolyn Chew-Graham, Rosalind McNally, and Sudeh Cheraghi-Sohi. 2014. "PICO, PICOS and SPIDER: A Comparison Study of Specificity and Sensitivity in Three Search Tools for Qualitative Systematic Reviews." *BMC Health Services Research* 14(1): 579.

National Institutes of Health (NIH). 2015. "Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research." Bethesda, Md.: NIH.

Okoroma, Francisca N. 2011. "Towards Effective Management of Grey Literature for Higher Education, Research and National Development." *Library Review* 60(9): 789–802.

Olson, Curtis A. 2013. "Using the Grey Literature to Enhance Research and Practice in Continuing Education for Health Professionals." *Journal of Continuing Education in the Health Professions* 33(1): 1–3.

Paez, Arsenio. 2017. "Gray Literature: An Important Resource in Systematic Reviews." *Journal of Evidence-Based Medicine* 10(3): 233–40.

Price, Amy. 2016. "Mendeley and More for Systematic Reviews." *The International Network for Knowledge About Wellbeing*. Accessed January 19, 2019. http://www.ithinkwell.org/mendeley-and-more-for-systematic-reviews/.

Pullin, Andrew S., Mukdarut Bangpan, Sarah Dalrymple, Kelly Dickson, Neal R. Haddaway, John R. Healey, Hanan Hauari, et al. 2013. "Human Well-Being Impacts of Terrestrial Protected Areas." *Environmental Evidence* 2(1): 19.

Relevo, Rose, and Howard Balshem. 2011. "Finding Evidence for Comparing Medical Interventions: AHRQ and the Effective Health Care Program." *Journal of Clinical Epidemiology* 64(11): 1168–77.

Rosenthal, Marylu C. 1994. "The Fugitive Literature." In *The Handbook of Research Synthesis*, edited by Harris Cooper and Larry V. Hedges. New York: Russell Sage Foundation.

Rothstein, Hannah R., and Sally Hopewell. 2009. "Grey Literature." In *The Handbook of Research Synthesis and Meta-Analysis*, 2nd ed., edited by Harris Cooper and Larry V. Hedges. New York: Russell Sage Foundation.

Rucinski, Taryn L. 2015. "The Elephant in the Room: Toward a Definition of Grey Legal Literature." *Law Library Journal* 107(4): 543–59.

Rudesill, Dakota S., James Caverlee, and Daniel Sui. 2015. "The Deep Web and the Darknet: A Look Inside the Internet's Massive Black Box." STIP 03. Washington, D.C.: Woodrow Wilson International Center for Scholars.

Saleh, Ahlam A., Melissa A. Ratajeski, and Marnie Bertolet. 2014. "Grey Literature Searching for Health Sciences Systematic Reviews: A Prospective Study of Time Spent and Resources Utilized." *Evidence Based Library and Information Practice* 9(3): 28–50.

Schmucker, Christine, Annette Bluemle, Matthias Briel, Susan Portalupi, Britta Lang, Edith Motschall, Guido Schwarzer et al. 2013. "A Protocol for a Systematic Review on the Impact of Unpublished Studies and Studies Published in the Gray Literature in Meta-Analyses." *Systems Reviews* 2 (May): 24.

Schmucker, Christine, Anette Blümle, Lisa K. Schell, Guido Schwarzer, Patrick Oeller, Laura Cabrera, Erik von Elm, Matthias Briel, and Joerg J. Meerpohl. 2017. "Systematic Review Finds That Study Data Not Published in Full Text Articles Have Unclear Impact on Meta-Analyses Results in Medical Research." *PLoS One* 12(4): e0176210.

Schöpfel, Joachim. 2010. "Towards a Prague Definition of Grey Literature." Paper presented at the Twelfth International Conference on Grey Literature. Prague (December 6–7, 2010).

Schöpfel, Joachim, and Dominic Farace. 2011. "Grey Literature." In *Encyclopedia of Library and Information Sciences*, 3rd ed., edited by Marcia J. Bates and Mary Niles Maack. New York: Taylor and Francis.

Schöpfel, Joachim, and Behrooz Rasuli. 2018. "Are Electronic Theses and Dissertations (Still) Grey Literature in

the Digital Age? A FAIR Debate." *The Electronic Library* 36(2): 208–19.

Schroll, Jeppe Bennekou, Lisa Bero, and Peter C. Gøtzsche. 2013. "Searching for Unpublished Data for Cochrane Reviews: Cross Sectional Study." *British Medical Journal* 346. DOI: 10.1136/bmj.f2231. Accessed November 30, 2018. https://www.bmj.com/content/bmj/346/bmj.f2231.full.pdf.

Shaw, Rachel L., Andrew Booth, Alex J. Sutton, Tina Miller, Jonathan A. Smith, Bridget Young, David R. Jones, and Mary Dixon-Woods. 2004. "Finding Qualitative Research: An Evaluation of Search Strategies." *BMC Medical Research Methodology* 4(1): 1.

Sibbald, Shannon L., Jennifer C. D. MacGregor, Marisa Surmacz, and C. Nadine Wathen. 2015. "Into the Gray: A Modified Approach to Citation Analysis to Better Understand Research Impact." *Journal of the Medical Library Association* 103(1): 49.

Song, Fujian, Sheetal Parekh, Lee Hooper, Yoon K. Loke, J. Ryder, Alex J. Sutton, C. Hing, Chun Shing Kwok, Chun Pang, and Ian Harvey. 2010. "Dissemination and Publication of Research Findings: An Updated Review of Related Biases." *Health Technology Assessment* 14(8): 1–193.

Speirs, Martha A. 2013. "Data Mining for Scholarly Journals: Challenges and Solutions for Libraries." Paper presented at the IFLA World Library and Information Congress 2013, Singapore (August 17–23, 2013). Accessed November 30, 2018. http://library.ifla.org/148/1/165-speirs-en.pdf.

Spencer, Andrew, Brendan Krige, and Sham Nair. 2014. "Digital Doorway: Gaining Library Users Through Wikipedia." Paper presented at the ALIA National Conference, Melbourne (September 15–19, 2014).

Stansfield, Claire, Kelly Dickson, and Mukdarut Bangpan. 2016. "Exploring Issues in the Conduct of Website Searching and Other Online Sources for Systematic Reviews: How Can We Be Systematic?" *Systematic Reviews* 5(1): 191.

Strauss, Anselm, and Juliet Corbin. 1998. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Thousand Oaks, Calif.: Sage Publications.

Thelwall, Mike, and Kayvan Kousha. 2015. "ResearchGate: Disseminating, Communicating, and Measuring Scholarship?" *Journal of the Association for Information Science and Technology* 66(5): 876–89.

Tyndall, Jessica. 2010. "AACODS Checklist." Adelaide: Flinders University.

Vaska, Marcus, and Rosvita Vaska. 2016. "Looking for Information That Is Not Easy to Find: An Inventory of LibGuides in Canadian Post-Secondary Institutions Devoted to Grey Literature." Paper presented at the 18th International Conference on Grey Literature, New York (November 28–29, 2016).

Wade, C. Anne, Herbert M. Turner, Hannah R. Rothstein, and Julia G. Lavenberg. 2006. "Information Retrieval and the Role of the Information Specialist in Producing High-Quality Systematic Reviews in the Social, Behavioural and Education Sciences." *Evidence & Policy: A Journal of Research, Debate and Practice* 2(1): 89–108.

Webster, Jane, and Richard T. Watson. 2002. "Analyzing the Past to Prepare for the Future: Writing a Literature Review." *MIS Quarterly* 26(2): xiii–xxiii.

Weintraub, Irwin. 2000. "The Impact of Alternative Presses on Scientific Communication." *International Journal on Grey Literature* 1(2): 54–59.

Whiting, Penny, Marie Westwood, Margaret Burke, Jonathan Sterne, and Julie Glanville. 2008. "Systematic Reviews of Test Accuracy Should Search a Range of Databases to Identify Primary Studies." *Journal of Clinical Epidemiology* 61(4): 357–64.

Wildridge, Valerie, and Lucy Bell. 2002. "How CLIP Became ECLIPSE: A Mnemonic to Assist in Searching for Health Policy/Management Information." *Health Information and Libraries Journal* 19(2): 113–15.

Wright, Judy M., David J. Cottrell, and Ghazala Mir. 2014. "Searching for Religion and Mental Health Studies Required Health, Social Science, and Grey Literature Databases." *Journal of Clinical Epidemiology* 67(7):800–10.

Yasin, Affan, and Muhammad Ijlal Hasnain. 2012. "On the Quality of Grey Literature and Its Use in Information Synthesis During Systematic Literature Reviews." Master's thesis, Blekinge Institute of Technology.

PART IV

CODING THE LITERATURE

7

INCORPORATING JUDGMENTS ABOUT STUDY QUALITY INTO RESEARCH SYNTHESES

JEFFREY C. VALENTINE
University of Louisville

CONTENTS

7.1 INTRODUCTION

If you were to gather a set of experts on research synthesis and ask them to list the five most important issues affecting the typical systematic review and meta-analysis, study quality would surely be on the list. It would be because it is essentially a truism that study quality can affect study results, which can in turn affect the conclusions drawn by the research synthesist. As a result, developing an explicit, transparent, and reproducible way of assessing study quality is an important goal for all research syntheses. Unfortunately, despite the general assumption that poor study quality positively biases effect sizes, the relationship between study quality and study results is not straightforward, and the implications of study quality for the results of a research synthesis are often not clear.

In this chapter, I make three points. First, a judgment of the quality of any given study is context dependent, in that it is a function of specific characteristics of research design and implementation that might be a threat to validity in the specific context of the studies in question. Second, in part because study quality is context dependent, for many research questions there is very little empirical evidence addressing whether study quality indicators covary with effect sizes, and even less evidence that the two are causally related. These facts have important implications for what we are able to say about the effects of study quality on the results of a research synthesis. The final point is that study quality is a multidimensional construct. Therefore, attempts to arrive at a single number, score, or unitary judgment about the quality of a particular study are probably misguided, and could lead to unwarranted conclusions about the cumulative results of a research literature.

7.2 WHAT IS STUDY QUALITY?

Study quality means different things to different people (Shadish 1989). A university administrator might say that a good study is one that results in a successful grant application. A teacher, human resource manager, or clinical psychologist might say that a good study is one that results in knowledge that improves practice. A journal editor might say that a good study is one that is often cited; a journal peer reviewer might say that a good study is one that makes a theoretical contribution to the literature. All of these are valid in their context. In the context of a research synthesis, a good study is one in which the research methods used are well aligned to the research question under investigation.

The description of study quality shares two characteristics with the notion of measurement validity. First, strictly speaking, the validity of a measure is neither absolute nor final. That is, even a measure that is well accepted in a field does not have perfect validity (so judgments of measurement validity are never absolute) and researcher's beliefs about a measure's validity are never completely settled; they change as new applications of the measure reveal more about the value of its use (so validity is never final). Second, even measures that researchers believe are very likely to be highly valid are believed valid only for particular purposes. For example, experts in human intelligence might believe that a certain IQ test produces a score that is likely to be a highly valid indicator of intelligence; these same experts, however, are unlikely to believe that the score is a highly valid indicator of marriage satisfaction. Like a measure being fit to its use, a study's research design and implementation should be fit for the research question it is trying to answer.

One consequence of describing study quality as the degree of fit between research goals and research questions is that study quality considerations differ as a function of the nature of the research question and the audience for the research. The study quality dimensions relevant to synthesists who are analyzing lab-based experiments in social psychology differ somewhat from those relevant to synthesists analyzing classroom-based experiments, and both differ somewhat from the study quality dimensions relevant to synthesists analyzing public polls. As a result, in this chapter I attempt to strike a balance between the need to speak to readers working in a variety of disciplines who might be working on different types of syntheses, both of which have implications for the specific quality assessments that are done.

7.3 WHY BE CONCERNED ABOUT STUDY QUALITY?

There are two related reasons to be concerned about study quality in a research synthesis. First, it is a certainty that study quality will vary across the studies included in a review. This will occur even if the synthesists' attempt to limit the extent of variability in the quality of included studies. For example, restricting included studies to only those using random assignment does not eliminate variability in study quality due to participant attrition, missing data, and so on.

In addition, empirical evidence indicates that study effect sizes can vary as a function of study quality indicators in some contexts. For example, Margaret Spinelli, Jean Endicott, and Raymond Goetz conducted a trial evaluating a treatment for postpartum depression (2015). They examined mood and global functioning ratings generated by independent assessors relative to those generated by the therapists delivering the intervention. Relative to the control group, the therapists delivering the intervention consistently rated participants in the treatment condition as showing more improvement than the masked independent evaluators, suggesting the possibility that therapist ratings were biased by their expectancies.

Similarly, Will Shadish and his colleagues conducted an experiment in which participants were randomly assigned to be in either a randomized experiment or in a nonrandomized experiment (2008). Participants in the randomized experiment were yet again randomly assigned to either a math or a vocabulary training session, and participants in the nonrandomized experiment were allowed to choose whether they would attend either math or vocabulary training. The results indicated that allowing participants to choose their condition (the nonrandomized experiment) was associated with a positive bias in effect sizes.

The studies by Spinelli and her colleagues (2015) and Shadish and his (2008) investigated the relationship between study quality indicators and effect size within a single sample of participants. Another strategy for examining the relationship between study quality and effect size is to use meta-analysis to compare the effect sizes obtained in studies that score well on a certain quality indicator to those that do not score well. Much of this work has been done in medicine and informed the development of the Cochrane Collaboration's Risk of Bias tool (Higgins et al. 2011). Thomas Chalmers, Raymond Matta, Harry Smith, and Anna-Marie Kunzler offer a very early example (1977). These researchers examined the effect sizes obtained from studies of the use of anticoagulants (blood thinners) for patients who had experienced a heart attack. Although both randomized experiments and nonrandomized experiments suggested that anticoagulants are beneficial, the effect was notably larger in nonrandomized studies.

Since then, many individual studies and meta-analyses have examined the impact of study quality indicators, and this continues to be a fruitful area of research. In fact, Agnes Dechartres and her colleagues conducted a meta-analysis of meta-analyses, which they refer to as meta-epidemiological studies (2016). Collectively, they included fifty-six meta-epidemiological studies that in turn included more than three thousand meta-analyses, which were based on more than twenty-one thousand individual studies. They concluded that two design features were associated with effect sizes across the syntheses. The first was allocation concealment, or making sure that those randomizing participants to conditions cannot predict what the next assignment will be. The range of odds ratios was 0.63 to 1.02 in ten meta-syntheses of binary outcomes, and because the effect sizes were coded such that odds ratios below 1.0 suggested a positive benefit, this result suggests a positive bias in effect sizes (that is, intervention effects were larger in studies that scored low on the allocation concealment dimension). The second was adequate randomized sequence generation. An example of a low-quality sequence generation might be using date of admission as the assignment mechanism (for example, patients admitted on an even date like the second or fourth of the month are assigned to the treatment condition). For this dimension, results again suggested a positive bias, with odds ratios ranging from 0.81 to 1.10.

One might infer from these examples that in medical contexts, low study quality upwardly biases effect size, that is, it makes intervention effects appear larger than they actually are. However, this is not universally true. To take one example, in the Dechartres meta-study, four syntheses of meta-analyses of continuous outcomes suggested very little, if any, bias associated with sequence generation problems (2016). It is not immediately clear why, if sequence generation is associated with effect sizes, the relationship would only hold true for binary outcomes. To take another example from that study, participant dropout was assessed in two meta-analyses. The association between dropout and effect size was statistically significant and negative in one (OR = 1.33), and not statistically significant and negative in another (OR = 1.07). These effect sizes are similar in magnitude, but in a different direction, to those observed for sequence generation. If dropout really is negatively related to effect size (that is, if more dropout is associated with smaller effects), the finding runs counter to the general expectation that low quality results in a positive bias.

As another example of the context dependency of study quality assessments, Cathleen McHugh and I used meta-analysis to examine the effects of attrition in randomized experiments in educational contexts (Valentine and McHugh 2007). We found little evidence that attrition biased effect sizes in these studies. However, the context does matter. Most of the studies included in this review

had three features that made it less likely that attrition biased effect size estimates. First, the studies were experiments that were conducted using young schoolchildren (roughly ages six through twelve) for whom schooling was compulsory. In addition, the studies occurred in contexts in which average attendance rates were high (greater than 95 percent). Finally, most of the studies were conducted in the United States, where it is not required to obtain participation consent for studies of everyday education practice. As a result the students in these studies (and their parents) were likely unaware that they were even participating in an experiment. All three of these characteristics suggest that whether a given student is available for outcome measurement is probably the result of something functionally random (like being sick on a given day) and hence unrelated to whether the student was in the treatment or in the control condition. This empirical example does not mean that attrition does not affect effect sizes. It does suggest that while attrition has the potential to bias estimates of treatment effects, its influence may be stronger or weaker depending on a study's context. Yet, attrition is addressed on almost all quality scales (more on this in the following section), and studies with "too much" attrition will usually be treated skeptically and perhaps even excluded from the review. This will happen even if attrition is actually unrelated to effect size in that context, thereby unnecessarily reducing the amount of information available for meta-analysis and the generalizability of the meta-analytic database. The problem highlighted by these examples is that an aspect of a study's research design that positively biases effect sizes in one context might negatively bias effect sizes in another context and have no effect in another context.

7.4 HOW NOT TO ADDRESS STUDY QUALITY

Given that dimensions related to study quality have the potential to be related to effect sizes, how then should synthesists address it in their reviews? Three common strategies are to deal with it minimally or not at all, to rely on the journal peer review process to weed out low-quality studies (operationally, to use only published studies), and to rely on scores derived from quality scales. All of these strategies have problems.

7.4.1 Not Addressing Study Quality

Blair Johnson and his colleagues examined two hundred research syntheses in the health promotion area, and found that about 20 percent of them did not explicitly mention methodological quality (2015). Similarly, Humam Saltaji and his colleagues examined just under a thousand syntheses on topics related to oral health, and found that more than half (56 percent) did not formally assess study quality (2016). These findings should not be interpreted to mean that many research synthesists fail to deal with study quality at all, because in fact it turns out to be difficult to completely avoid study quality issues in a research synthesis. For example, in their now-classic review of the relationship between class size and academic achievement, Gene Glass and Mary Lee Smith included studies "whether [or not] they employed rigorous empirical methods" (1979, 5), but did require that studies have a comparison group. In doing so, they eliminated from consideration studies that, for example, employed a single group of students and then measured outcomes before and after reducing class size. Of course, this is in essence a study quality consideration because the excluded design leaves open a number of alternative explanations (such as normal student growth). Later I return to the methods that Glass and Smith used to address study quality. For now, the point is that almost all research synthesists address study quality but not all do in an explicit, transparent, and reproducible way.

7.4.2 Peer Review as a Proxy for Study Quality

Historically, perhaps the most common strategy for addressing study quality in a research synthesis has been to rely on the peer review process to weed out low-quality studies. This is an easy to implement and appealing strategy but it also has its drawbacks. It is appealing because one goal of peer review can be to keep suspect research out of academic journals. But it may not reveal the picture of the literature the synthesists want because the peer review process is not a perfect screen; readers of research have seen published studies that they thought were not very good, in part because peer reviewers have other criteria that reports need to meet. As an example, journals are often interested in a study's potential for advancing theory. Similarly, though becoming somewhat less common, some journals have formal or informal policies discouraging replication research, expressing a preference for novelty over knowledge cumulation (see, for example, Kail 2012; Ritchie, Wiseman, and French 2012). In addition, journal editors and peer reviewers exhibit a preference for studies that have statistically significant findings on their main outcomes. This latter concern is particularly

important and is part of the reason that publication bias is a problem (see also chapter 19, this volume).

Another reason that using peer review as a proxy for study quality can be flawed is that much good research gets produced that is never submitted for peer review. Not everyone has publication as a goal. For example, government agencies and nonprofit foundations, such as charities, often commission research that they then make available outside of peer-reviewed journals. Assuming that peer review results in only high-quality studies is a mistake and assuming that all good studies undergo peer review compounds that error (Glass 2000).

7.4.3 Study Quality Scales

Another popular approach is to use a scale that measures study quality. This approach is popular in part because it relieves some of the burden of figuring out what quality means for the research question guiding the synthesis. Most commonly, quality scales result in a single score that represents the quality of a study. This score is used as a weight in meta-analysis, a rare approach and not recommended (see Ahn and Becker 2011), to set study inclusion criteria, that is, the meta-analysis only includes studies that pass a certain score threshold, or used to categorize studies into high and low study quality groups.

All of these approaches are problematic because the scores produced by quality scales are likely error prone. That is, they introduce an additional source of measurement error into analyses because the validity of quality scales has rarely been subject to empirical examination (see, for example, Crowe and Sheppard 2011). For example, blood clots are a common and potentially serious complication of surgery. Peter Jüni and his colleagues found a systematic review and meta-analysis investigating the relative effectiveness of two different versions of a drug (low molecular weight heparin versus standard heparin) on the likelihood that surgery patients would develop post-operative blood clots (1999). They also located twenty-five quality scales, twenty-four of which were published in peer-reviewed medical journals. They then conducted twenty-five different meta-analyses, with each meta-analysis using one of the twenty-five quality scales. Specifically, they selected a quality scale and applied it to all of the studies in the original meta-analysis. They then categorized the studies in the meta-analysis as either high or low quality, and performed meta-analysis within each study quality category. They then examined the relative effectiveness of low molecular weight heparin relative to standard heparin in the high study quality and in the low study quality categories. This process was repeated for all twenty-five quality scales.

Jüni and colleagues (1999) found that in about half of the twenty-five meta-analyses, effect sizes were similar in the high and low study quality categories. In about a quarter of the meta-analyses, the *high-quality* studies suggested that *low molecular weight heparin* was *more effective* than standard heparin, while the low-quality studies suggested that low molecular weight heparin was no more effective than standard heparin. In the remaining meta-analyses, this pattern was reversed: the *high-quality* studies suggested that *low molecular weight heparin* was *no more effective* than standard heparin, while the low-quality studies suggested that low molecular weight heparin was more effective than standard heparin (for a similar example in studies of interventions for students with learning and other disabilities, see also Cook, Dupuis, and Jitendra 2015; for another example from medicine, Brouwers et al. 2005).

Jüni and his colleagues' results suggest that the decision about the relative effectiveness of low molecular weight heparin depended on the quality scale chosen. Imagine surgeons are preparing for surgery and run across the original meta-analysis comparing the effects of low molecular weight versus standard heparin. The surgeons know that study quality is an important issue, so they find a study quality scale published in a recent medical journal and apply it to the studies in the original meta-analysis, and base the decision about whether to give their patients low molecular weight or standard heparin on what the "high-quality" studies suggest is the best course of action. The treatment decision is being driven in part by the quality scale that the surgeons choose to use; this choice was arbitrary, and it may have similarly arbitrary effects on treatment outcomes.

Why did the scales come to different conclusions about study quality? Harris Cooper and I suggest several reasons (Valentine and Cooper 2008). First, the quality scales that Jüni and his colleagues collected differed in their comprehensiveness: the number of items ranged from three to thirty-four (1999). As a result, some scales were quite general and others more detailed. In addition, even when scales have a similar number of items the weights assigned to those items can vary dramatically. For example, one of the quality scales used by Jüni et al. allocated 15 percent of its total points to whether or not the study randomly placed patients into conditions, and 5 percent of its total points to whether the outcome

assessors were unaware of the patients drug condition. Another quality scale with the same number of items reversed these weights (15 percent for masking outcome assessors, 5 percent for randomly assigning patients to conditions). This highlights a critical problem: the weights used in quality scales are highly arbitrary and almost certainly have no empirical justification. At best, they reflect judgments by the scale's authors about the relative importance of the various dimensions included on the scale. These judgments are probably influenced by the particular contexts in which the authors work, so even if the weights are reasonable for the contexts with which the authors are familiar, they are less likely to be reasonable for other contexts. At worst, the weights applied to study quality dimensions are essentially idiosyncratic to the scale's authors.

As worrisome as these two issues are, an even bigger problem is that most study quality scales result in a single number that represents study quality. This means that two studies with very different strengths and weaknesses might receive the same score on the same scale. For example, a study that implements a very solid research design on a sample that is quite unrepresentative of the population of individuals for whom an intervention was developed might receive the same score as a study with a weaker research design but a much more representative sample. It is hard to imagine how both of these scores could be valid representations of study quality.

7.5 HOW TO ADDRESS STUDY QUALITY

When considering how to address study quality in your research synthesis, there are a few principles that should be kept in mind. Most can be inferred from the preceding discussion. First, study quality is a multidimensional construct. As a result, it is best to refer to study quality dimensions and their associated indicators rather than study quality, which implies a unitary judgment. In addition, in many research fields, it is difficult to determine whether a particular study quality indicator will positively bias effect size estimates, negatively bias effect size estimates, or have no consistent effect on effect size estimates.

These two principles suggest that synthesists proceed very cautiously when making judgments about study quality, and explicitly justify their reasoning to readers. They should exercise restraint when using study quality as a reason to exclude studies from your synthesis and making claims about the relationship between study quality indicators and effect size.

Finally, whether a given study quality indicator will be important depends on the research question asked, which is to say that study quality is context dependent. As a result, synthesists need to think deeply about how study quality indicators might operate in the context of their research question. Notions of internal, construct, and statistical conclusion validity can be especially helpful when thinking about study quality (Shadish, Cook, and Campbell 2002).[1] In this tradition, validity refers to the approximate truth of an inference or claim about a relationship (Cook and Campbell 1979; Shadish, Cook, and Campbell 2002). Different characteristics of a study's design and implementation lead to inferences that have more or fewer challenges to valid along one or more dimensions. Factors that might lead to an incorrect inference are termed "threats to validity" or "plausible rival hypotheses." *Internal validity* refers to the validity of inferences about whether some intervention or experimental manipulation has caused an observed change in an outcome. Threats to internal validity include any mechanism that might plausibly have caused the observed outcome even if the manipulation had never occurred. *Construct validity* refers to the extent to which the operational characteristics of manipulations and outcome measures used in a study adequately represent the intended abstract categories. Researchers most often think of construct validity in terms of the outcome measures. However, construct validity also refers to the adequacy of other labels used in the study, such as the manipulations and the labels applied to participants (for example, at-risk). As an example, a measure that purports to measure intelligence but actually measures academic achievement is mislabeled and hence, presents a construct validity problem. *Statistical conclusion validity* refers to evidence on the covariation between two variables (for example, a treatment assignment indicator and an outcome) arising from a study. In the context of a research synthesis, the specific concern is about the effect size estimate and its precision (for example, the latter can be greatly affected by violating the assumption of statistical independence; see Konstantopoulos and Hedges, this volume).

The task of a synthesist is to think deeply about the relevant validity considerations that might be operating in the context of the research question. An extended example might help. Recall that Will Shadish and his colleagues randomly assigned participants to be in either a randomized experiment or in a nonrandomized experiment: specifically, participants in the nonrandomized part of the study were allowed to choose which tutorial to par-

ticipate in (2008). Doing so allowed them to estimate the direction and magnitude of bias associated with the nonrandomized experiment. However, their primary interest was in identifying the conditions under which the nonrandomized experiment might generate results that approximated the randomized experiment. To do so, they collected data about participants' math and vocabulary pretest scores, their topic preference, several personality characteristics, levels of math anxiety, and depression. They chose these variables because the theoretical and empirical literature suggested that these characteristics might be good predictors of which group a given participant would choose. Shadish and colleagues concluded that including these variables in the estimation model successfully removed most of the bias associated with self-selection (as opposed to random assignment) into study conditions.

From the standpoint of a research synthesis, the important point is that there are indicators related to the biases introduced by poor study design. However, because these will vary from research question to research question, at least one synthesis team member should be highly expert about the research problem.

That said, an important caveat is that it is possible that even within a particular research question, study quality indicators might play out in different ways in different studies in the same review. For example, consider a synthesis that examines the effectiveness of a new drug relative to standard treatment. In some studies, patient socioeconomic status is related to drug choice and in others it is not. Socioeconomic status may be important because the higher price for the new drug will lead poorer people to choose the standard treatment. It might then be appropriate for the synthesists to downgrade nonrandomized studies if they do not control for well-measured socioeconomic status in the estimation model. However, if some nonrandomized studies were conducted in more affluent communities, then the likelihood that socioeconomic status biased results is much lower than it is in poorer communities, and the resulting quality judgments might be in error. This is yet another reason to exercise restraint when making claims about the effects of study quality on study outcomes.

Ideas about which study quality indicators are most likely to matter can come from a number of sources, including relevant empirical research examining the relationship between study quality indicators and effect size, previous research (including qualitative research) on the research question, research methods textbooks, and study quality instruments like ROBINS-I (Sterne et al. 2016), the Study DIAD (Valentine and Cooper 2008), the Cochrane Risk of Bias tool (Higgins et al. 2011), and the RTI item bank (Vishwanathan and Berkman 2012). Furthermore, scholars may have written about study quality indicators in a relevant context. As examples, see the What Works Clearinghouse's evidence standards for education (What Works Clearinghouse 2017), and the advice designed for specific types of education research, such as writing (see Graham and Harris, 2014).

Perhaps more important, research synthesists will often have formal hypotheses and informal hunches about the causes and consequences of variation on study quality dimensions. As an example, assume that a team of synthesists believe that, for a synthesis on the effects of some drug, that adverse reactions to the active ingredient in a medication might lead to participant attrition and adverse reactions are more likely to occur among less healthy individuals. If true, it suggests that attrition should positively bias effect size (less healthy individuals are dropping out of the treatment group at a rate greater than in the control group), and the synthesists would be well justified in scrutinizing attrition rates and how attrition was addressed in the analysis. Such hypotheses should be informed not only by statistical theory but also by empirical research from the discipline in which the research synthesists are working.

In general, there are two types of quality indicators. First, some indicators relate to the choices made by researchers while designing their studies. For example, whether the study is a randomized or a nonrandomized experiment, and whether the researchers used outcome measures that produce reasonably valid scores are two often used indicators of quality. Another category of items relates to the choices that researchers made while analyzing the results of their studies. How missing data were handled and the covariates included in an estimation model are examples of considerations in this category.

Occasionally, authors may report information in a way that allows synthesists to approximate or even exactly recreate the original data, which could then possibly allow them to reanalyze the data in a way they believe is more appropriate. For example, assume the synthesists believe that socioeconomic status is an important covariate in nonrandomized experiments. A study involved collecting this information but it was not included in the estimation model because the SES difference between the treatment and control groups was not statistically significant. If the authors reported sufficient statistics the

synthesists will be able to recreate the underlying data and estimate the model with SES included. Similarly, and perhaps more likely, when a good pretest of an outcome exists it can often be helpful to compute a simple "difference in differences" effect size measure (that is, the post-pre change in the intervention group minus the post-pre change in the comparison group) to reduce bias, and the needed descriptive statistics are frequently available.[2]

I have three general suggestions for addressing study quality in a research synthesis: set a few study inclusion criteria based on highly defensible study quality indicators; examine the interrelations between study quality indicators and other study characteristics; and test the relationship between study quality indicators and the effect size revealed in included studies.

7.5.1 Set Inclusion Criteria

Virtually all synthesists use at least some study quality criteria as screening tools. Depending on the stringency of these criteria, doing so can be thought of as either weeding out the weakest evidence or excluding all but the strongest evidence. As an example of weeding out the weakest evidence, a synthesist investigating the effects of an intervention might include all designs that have an intervention and a comparison group, but exclude one group pretest-posttest designs. As an example of excluding all but the strongest evidence, a synthesist might include only randomized experiments that experienced low attrition.

To begin, synthesists should be as inclusive as possible when setting study inclusion criteria based on study quality indicators. The inclusion criteria should be highly defensible, so that few researchers will question these decisions. This is because in many research contexts there is insufficient information to make a confident prediction about the likely direction and magnitude of the relationship between a particular study quality indicator and effect size. Furthermore, assume there is good reason to believe that a certain study quality indicator does bias effect sizes. Including studies that score high and studies that score low on the study quality indicator will allow a test of this belief. And, there is nothing that compels synthesists to group all of their included studies into a single meta-analysis. In many cases it might be reasonable to conduct separate meta-analyses by study quality indicator (for example, conduct separate meta-analyses for randomized and nonrandomized experiments).

Three considerations temper the use of an inclusive approach. First, although this is relatively rare, it might be the case that for a given research question, a very large number of studies scoring high on certain study quality indicators are available. If so, limiting the review to studies that score high on these study quality indicators is a viable option. Another circumstance that merits limiting included studies occurs when a field has settled questions about the direction of bias associated with a certain study quality indicator. For example, assume the primary outcome is one for which experts in the field have agreed should be measured in a certain way (for example, the field has agreed on a "gold standard" measurement or measurements). In this case it seems reasonable to restrict the review to studies that follow these guidelines. Finally, all syntheses consume resources, and sometimes including a wide variety of studies can overwhelm those resources. In this case it might make sense to be more restrictive. That said, restrictions putatively done in the name of study quality should not serve to make things easier on the synthesists at the expense of good work with a reasonable expenditure of resources.

7.5.2 Interrelations Between Indicators and Other Characteristics

Mark Lipsey and David Wilson assert that studies have "personalities" (2001). This is true in the sense that study characteristics tend to cluster together in a trait-like way. This fact is not limited to the interrelations between study quality indicators, but also extends to the relationships between study quality indicators and substantively important characteristics of the studies. This problem relates to what Cooper refers to as the distinction between study-generated evidence and review-generated evidence (2017). Characteristics such as the nature of the sample and the duration of the intervention are either selectively chosen by researchers or dictated by context. They are not randomly assigned. As a result, any observed covariation between study characteristics (including study quality indicators) and study outcomes could be the result of a causal relationship or evidence of a noncausal association (including a spurious relationship caused by an unmeasured third variable).

The interrelations between study quality indicators and study characteristics can be examined in a number of ways. One is to construct a correlation matrix with relevant study characteristics. All variables need to be scaled so that the correlation coefficient is interpretable. For example, variables with multiple categories such as race/ethnicity will not work—they need to be reduced to single

Table 7.1. Example of Study Characteristics Cross-Tabulated by Study Design

Study Characteristic	Randomized	Not Randomized
Intervention duration (mean weeks)	10.2	16.5
Intervention intensity (mean minutes per week)	55.8	56.1
Intervention fidelity (percentage scoring high)	41	22
Low SES (percentage of sample)	66	45
Age (mean of sample)	10.2	10.2
Performing at grade level (percentage of sample)	72	84

SOURCE: Authors' tabulation.

degree of freedom contrasts. And, the statistical significance of tests should not be overinterpreted and even might be ignored; this is a case in which the size of the correlations matters much more than the statistical significance of the relationships. Another strategy is to create a cross-tabulation table that illustrates how important study features vary as a function of study quality indicators (for an example, see table 7.1).

7.5.3 Statistical Tests

In some cases, it will be possible to statistically test the relationships between study quality indicators and effect size. This is often referred to as moderator analysis, and it is the strategy that Glass and Smith argued ought to be used to address study quality (1979). Conceptually, moderator analysis in a meta-analysis can be thought of as an extension of regression. Meta-analysts can take a univariate approach by testing relationships between study quality indicators and effect size one at a time (analogous to simple regression), or a multivariate approach by testing relationships between multiple study quality characteristics and potentially other study characteristics (for example, variations in treatment) simultaneously (analogous to multiple regression). Neither approach should be used when the number of studies is small, and the multivariate approach requires a relatively large number of studies. Unfortunately, it is not possible to quantify the terms *small* and *relatively large*. The definition of these terms depends on analytic choices, such as model choice, the within-study sample sizes, though the number of studies is more important, and on the spread of the covariates, for example, if testing study design as a moderator, a 90–10 split in designs will require more studies than a 50–50 split (see Hedges and Pigott 2004).

If using a univariate approach, perhaps because the number of studies is not large enough to support a multivariate analysis, the number of tests should be limited to avoid type I error inflation. One option is to test only the most important characteristics. It is best to identify these characteristics before looking at the data, and if identified after looking at the data, this should be disclosed to readers. Another strategy is to implement a correction for multiple hypothesis tests. Of course, both strategies can be used simultaneously. Conceptually, the univariate approach involves testing one relationship at a time, and this increases the likelihood that a particular relationship is either wholly or partly spurious.

Though requiring more data, the multivariate approach has the advantage of yielding effect size estimates that reflect statistical control over multiple specified study characteristics. As in multiple regression, the analysis can be done simultaneously (all covariates are modeled at the same time) or hierarchically. For example, the synthesists might be interested in testing the effect of a collection of implementation indicators while controlling for several indicators of the validity of the randomization process, so they enter the randomization indicators as a block, and then the implementation indicators as another block. Doing so will allow the calculation the proportion of variance in study effect sizes explained by adding the implementation block, and will yield regression coefficients for the implementation indicators that control for the other study quality indicators.

Regardless of the approach taken, three additional points are worth making. First, determining the proportion of variance explained (R^2) in meta-regression is not straightforward. In a primary study, theoretically 100 percent of the variance in observed outcomes is explainable. In a meta-analysis, some of the variance in observed

outcomes (effect sizes) is due to random sampling error, and hence, is explainable by sampling theory only. As a result the R^2 index has to be modified for use in meta-analysis by estimating and removing the sampling variance in effect sizes (for details, see Aloe, Becker, and Pigott 2010). One implication of this is that the R^2 reported by many statistical programs is not usable because it is an underestimate of the explainable variance.

Another important point is that low statistical power is often a concern in these analyses. This is another reason not to overinterpret the hypothesis tests. In this case, failure to reject the null hypothesis that different levels of a moderator have different population values, it is possible that low statistical power concealed a positive result (for introductory treatments of statistical power in meta-analysis, see Hedges and Pigott 2001, 2004; Valentine, Pigott, and Rothstein 2010).

Finally, when examining whether study indicators are associated with study outcomes it is best seen as a primarily descriptive (as opposed to inferential) exercise. In part this assertion rests on Cooper's distinction between study-generated and review-generated evidence discussed earlier (2017). We are almost always working with review-generated evidence, and thus any resulting inferences are tentative at best. But, describing the differences associated with study quality indicators is important on its own. For example, assume that a meta-analysis reveals that randomized experiments were associated with smaller effects than nonrandomized experiments, and that the difference was statistically significant. If the effect sizes were $d = +0.30$ for randomized experiments and $d = +0.40$ for nonrandomized experiments, the two effect sizes probably have a roughly similar meaning. And even if the difference was not statistically significant, if the effect sizes were $d = +0.01$ for randomized experiments and $d = +0.30$ for nonrandomized experiments, this could be an important finding. Focusing on the underlying effect sizes (as opposed to the inferential test) is the best way to make sure that these differences are clear to readers.

7.6 CONCLUSION

This chapter provides an overview of several ways of addressing study quality in a research synthesis and offers three main points. First, study quality is context dependent. A feature of research design and implementation that might be a threat to validity in one context might be irrelevant in a different context. Second, in many research areas very little empirical evidence addresses whether study quality indicators covary with effect sizes, and even less that the two are causally related. Finally, study quality is a multidimensional construct. Taken together, these considerations suggest that synthesists are often left to proceed without strong, context-independent guidance about how to address study quality. This uncertain situation can be uncomfortable, and lead synthesists to adopt strategies (such as using study quality scales to arrive at a single quality score) that are less helpful than they seem to be and may even be counterproductive.

A better approach is to carefully consider the study quality indicators likely to be important in the context of the research question. The most important of these indicators might be used as study inclusion criteria. For study quality indicators not used as inclusion criteria, it will be important to carefully describe how they are related to each other and to other study characteristics. If possible, the relationship between study quality indicators and study outcomes should be tested (remember that there must be a sufficient number of studies to do this). Taken together, this collection of strategies is the best way to balance the need to limit the potentially biasing effect of studies that score poorly on quality indicators with the reality that for most research questions the evidence that study quality does bias study outcomes is not strong.

7.7 NOTES

1. In general, considerations related to external validity should not be part of an assessment of the quality of different study dimensions. Although there are exceptions, in most cases external validity indicators will not bias effect size estimates. Instead, external validity indicators are usually addressed by including studies that match the desired population, intervention, and outcomes of interest.
2. Two related points are worth highlighting. First, always be aware of the possibility of ecological bias, which refers to the situation in which a relationship found at one level, such as the meta-analytic, might not hold at another, such as the individual study (for more, see Berlin et al. 2002; Cooper and Patall 2009). A second point is that study authors may be willing to reanalyze data for you, for example by using your preferred missing data technique or by running a custom estimation model. Increasingly, datasets are publicly available. For example, the Association for Psychological

Science awards an Open Data Badge to authors for publicly warehousing data and software code to allow others to attempt to reproduce the results (https://www.psychologicalscience.org/publications/badges).

7.8 REFERENCES

Ahn, Soyeon, and Betsy J. Becker. 2011. "Incorporating Quality Scores in Meta-Analysis." *Journal of Educational and Behavioral Statistics* 36(5): 555–85. DOI: 10.3102/1076998610393968.

Aloe, Ariel M., Betsy J. Becker, and Therese D. Pigott. 2010. "An Alternative to R^2 for Assessing Linear Models of Effect Size." *Research Synthesis Methods* 1(3–4): 272–83.

Berlin, Jesse A., Jill Santanna, Chris H. Schmid, Linda A. Szczech, and Harold I. Feldman. 2002. "Individual Patient-Versus Group-Level Data Meta-Regression for the Investigation of Treatment Effect Modifiers: Ecological Bias Rears Its Ugly Head." *Statistics in Medicine* 21(3): 371–87.

Brouwers, Melissa C., Mary E. Johnston, Manya L. Charette, Steve E. Hanna, Alejandro R. Jadad, and George P. Browman. 2005. "Evaluating the Role of Quality Assessment of Primary Studies in Systematic Reviews of Cancer Practice Guidelines." *BMC Medical Research Methodology* 5(1): 8. DOI: 10.1186/1471-2288-5-8.

Chalmers, Thomas C., Raymond J. Matta, Harry Smith Jr., and Anna-Marie Kunzler. 1977. "Evidence Favoring the Use of Anticoagulants in the Hospital Phase of Acute Myocardial Infarction." *New England Journal of Medicine* 297(20): 1091–96.

Cook, Thomas D., and Donald T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston, Mass.: Houghton Mifflin.

Cook, Bryan G., Danielle N. Dupuis, and Asha K. Jitendra. 2015. "A Preliminary Investigation of the Empirical Validity of Study Quality Appraisal." *Journal of Learning Disabilities*. DOI: 10.1177/0022219415581178.

Cooper, Harris. 2017. *Research Synthesis and Meta-Analysis*, 5th ed. Thousand Oaks, Calif.: Sage Publications.

Cooper, Harris, and Erika A. Patall. 2009. "The Relative Benefits of Meta-Analysis Using Individual Participant Data and Aggregate Data." *Psychological Method* 14(2): 165–76.

Crowe, Michael, and Lorraine Sheppard. 2011. "A Review of Critical Appraisal Tools Show They Lack Rigor: Alternative Tool Structure Is Proposed." *Journal of Clinical Epidemiology* 64(1): 79–89. doi:10.1016/j.jclinepi.2010.02.008.

Dechartres, Agnes, Ludovic Trinquart, Timor Faber, and Philippe Ravaud. 2016. "Empirical Evaluation of Which Trial Characteristics Are Associated with Treatment Effect Estimates." J*ournal of Clinical Epidemiology* 77(1): 24–37.

Glass, Gene V. 2000. "Meta-Analysis at 25." Accessed December 3, 2018. http://www.gvglass.info/papers/meta25.html.

Glass, Gene V., and Mary Lee Smith. 1979. "Meta-Analysis of Research on Class Size and Achievement." *Educational Evaluation and Policy Analysis* 1(1): 2–16.

Graham, Steve E., and Karen R. Harris. 2014. "Conducting High Quality Writing Intervention Research: Twelve Recommendations." *Journal of Writing Research* 6(2): 89–123. DOI: 10.17239/jowr-2014.06.02.1.

Hedges, Larry V., and Therese D. Pigott. 2001. "The Power of Statistical Tests in Meta-Analysis." *Psychological Methods* 6(3): 203–17.

———. 2004. "The Power of Statistical Tests for Moderators in Meta-Analysis." *Psychological Methods* 9(4): 426–45.

Higgins, Julian P. T., Douglas G. Altman, Peter C. Gøtzsche, Peter Jüni, David Moher, Andy D. Oxman, and Jonathan A. Sterne. 2011. "The Cochrane Collaboration's Tool for Assessing Risk of Bias in Randomised Trials." *British Medical Journal* 343: d5928.

Johnson, Blair T., Robert E. Low, and Hayley V. MacDonald. 2015. "Panning for the Gold in Health Research: Incorporating Studies' Methodological Quality in Meta-Analysis." *Psychology and Health* 30(1): 135–52.

Jüni, Peter, Anne Witshci, Ralph Bloch, and Matthias Egger. 1999. "The Hazards of Scoring The Quality of Clinical Trials for Meta-Analysis." *Journal of the American Medical Association* 282(11): 1054–60.

Kail, Robert V. 2012. "Reflections on Five Years as Editor." *Observer* 25(9). Accessed December 3, 2018. https://www.psychologicalscience.org/observer/reflections-on-five-years-as-editor.

Lipsey, Mark W., and David B. Wilson. 2001. "The Way in Which Intervention Studies Have 'Personality' and Why It Is Important to Meta-Analysis." *Evaluation and the Health Professions* 24(3): 236–54.

Ritchie, Stuart. J., Richard Wiseman, and Christopher C. French. 2012. "Replication, Replication, Replication." *The Psychologist* 25(5): 346–48.

Saltaji, Humam, Maria B. Ospina, Susan Armijo-Olivo, Shruti Agarwal, Greta G. Cummings, Maryam Amin, and Carlos Flores-Mir. 2016. "Evaluation of Risk of Bias Assessment of Trials in Systematic Reviews of Oral Health Interventions, 1991–2014." *Journal of the American Dental Association* 147(9): 720–28. DOI: 10.1016/j.adaj.2016.03.017

Shadish, William R. 1989. "The Perception and Evaluation of Quality in Science." In *Psychology of Science: Contributions to Metascience,* edited by Barry Gholson, William R.

Shadish Jr., Robert A. Niemeyer, and Arthur C. Houts. Cambridge: Cambridge University Press.

Shadish, William R., M. H. Clark, and Peter Steiner. 2008. "Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments." *Journal of the American Statistical Association* 103(484): 1334–44.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston, Mass.: Houghton Mifflin.

Spinelli, Margaret G., Jean Endicott, and Raymond R. Goetz. 2015. "Disagreement Between Therapist Raters and Independent Evaluators in a Controlled Clinical Trial of Interpersonal Psychotherapy for Depressed Pregnant Women." *Journal of Psychiatric Practice* 21(2): 114–23.

Sterne, Jonathan A. C., Miguel A. Hernán, Barnaby C. Reeves, Jelena Savović, Nancy D. Berkman, Meera Viswanathan, David Henry, Douglas G. Altman, Mohammed T. Ansari, et al. 2016. "ROBINS-I: A Tool for Assessing Risk of Bias in Non-Randomized Studies of Interventions." *British Medical Journal* 355:i4919. DOI: 10.113/bmj.i4919.

Valentine, Jeffrey C., and Harris Cooper. 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(2): 130–49.

Valentine, Jeffrey C., and Catherine McHugh. 2007. "The Effects of Attrition on Baseline Group Comparability in Randomized Experiments in Education: A Meta-Analytic Review." *Psychological Methods* 12(3): 268–82.

Valentine, Jeffrey C., Therese D. Pigott, and Hannah R. Rothstein. 2010. "How Many Studies Do You Need? A Primer on Statistical Power for Meta-Analysis." *Journal of Educational and Behavioral Statistics* 35(2): 215–47.

Viswanathan, Meera, and Nancy D. Berkman. 2012. "Development of the RTI Item Bank on Risk of Bias and Precision of Observational Studies." *Journal of Clinical Epidemiology* 65(2): 163–78. DOI: 10.1016/j.jclinepi.2011.05.008.

What Works Clearinghouse. 2017. *Standards Handbook, Version 4.0*. Washington: Institute of Education Sciences. Accessed January 3, 2019. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf.

8

IDENTIFYING POTENTIALLY INTERESTING VARIABLES AND ANALYSIS OPPORTUNITIES

MARK W. LIPSEY
Vanderbilt University

CONTENTS

8.1 INTRODUCTION

Research synthesis relies on information reported in a selection of studies on a topic of interest. When that information is captured in quantitative form, it can be used in a meta-analysis that subjects it to statistical analysis. This chapter examines the types of variables that can be coded from the source studies, outlines the kinds of relationships that can be examined in the meta-analysis of the resulting data, and identifies a range of analysis opportunities to stimulate thinking about what data might be used in a research synthesis and what sorts of questions might be addressed.

In particular, the variables that can be coded from the studies included in a meta-analysis can be divided into those representing study results in the form of effect sizes and those describing various characteristics of the studies that produced the results (study descriptors). These variables provide numerous analysis opportunities for the meta-analyst and, for discussion purposes in this chapter, are organized into four categories. One is descriptive analysis, which portrays the nature of the studies included in the meta-analysis with regard to the distributions of effect sizes and the profile of study characteristics. The other three involve examination of interrelationships among the coded variables: among study descriptors, between study descriptors and effect sizes, and among effect sizes. Each of these has the potential to yield informative findings about the body of research synthesized in the meta-analysis.

8.1.1 Potentially Interesting Variables

Meta-analysis revolves around effect-size statistics that summarize the findings of each study of interest on such relationships as those between treatment conditions and outcomes or associations between variables. Effect sizes, therefore, provide one or more variables (typically dependent variables) that are sure to be of interest in most meta-analyses. However, research studies have other characteristics that also may be of interest. These include, for instance, the research designs and procedures used; the attributes of the participant samples; and features of the settings, personnel, activities, and circumstances involved. Such characteristics constitute a second major category of variables of potential interest. As a group, they are referred to as study descriptors.

8.1.2 Study Results

The quantitative findings of bodies of research appropriate for meta-analysis may involve only a single variable examined in each study, such as the proportion of survey respondents who self-identify as political conservatives. Most often, however, they involve relationships between pairs of variables. These relationships may represent associations between variables measured on a sample of respondents, such as the relationship between religious affiliation and political conservatism. One of the variables may also represent a differentiation of respondents that is of particular interest, such as between treatment and control groups in experimental studies or between males and females in studies of gender differences. It is those relationships that are captured in the effect-size statistics commonly used in meta-analysis. As detailed elsewhere in this volume (see chapter 11), effect sizes come in a variety of forms—correlation coefficients, standardized differences between means, odds ratios, and so on—depending on the nature of the quantitative study results they represent.

Whatever the issues to be addressed, one of the first challenges a meta-analyst faces is the likelihood that many of the studies of interest will report multiple quantitative findings relevant to those issues and thus yield multiple effect sizes that might be included in the analysis. If we define a study as a set of observations taken on a subject sample on one or more occasions, three possible forms of multiple effect sizes may be available: different measures, different subsamples, and different times of measurement. Each of these has implications for conceptualizing, coding, and analyzing effect sizes.

8.1.2.1 Effect Sizes for Multiple Measures Each study in a meta-analysis may report relationships of interest on more than one measure. These relationships may involve different constructs, that is, different things being measured, such as academic achievement, attendance, and attitudes toward school. They may involve multiple measures of the same construct, such as achievement measured both by standardized achievement tests and grade point averages. A study of predictors of school achievement thus might report the correlations of age, gender, one or more measures of peer relations, and one or more measures of family structure with one or more measures of achievement. Each such correlation can be coded as a separate effect size. Similarly, a study of gender differences in aggressive behavior might compare males and females on physical aggression measured two ways, verbal aggression measured three ways, and relational aggression measured one way, yielding six possible effect sizes. Moreover, the various studies eligible for a meta-analysis may differ among themselves in the type, number, and mix of measures that contribute to effect sizes.

A meta-analyst must decide what constructs to define and what measures will be viewed as representing those constructs. Within that framework, a key question is what effect sizes to analyze for each relationship involving those constructs and measures. The basic options are threefold. One approach is to analyze effect sizes on all measures for all constructs in relatively undifferentiated form. This strategy would yield, in essence, a single global category of study results. For example, the outcome measures used in research on the effectiveness of psychotherapy show little commonality from study to study. A meta-analyst might, for some purposes, treat all these outcomes as the same—that is, as instances of results relevant to some general construct of personal functioning. An average over all the resulting effect sizes addresses the global question of whether psychotherapy has generally positive effects on the mix of measures typical to psychotherapy research. This is the approach that was used in the classic meta-analysis of research on psychotherapy effects (Smith and Glass 1977).

At the other extreme, effect sizes may be analyzed only when they relate to a particular measure of a construct of interest, that is, a specific operationalization of that construct or, perhaps, a few such specific operationalizations. Effect sizes would then be computed and analyzed separately for each such measure. Bradley Erford, Erin Johnson, and Gerta Bardoshi, for instance, conducted a meta-analysis of the psychometric properties of the Beck Depression Inventory that was appropriately limited to studies of that specific instrument (2016). Similarly, Michiel van Vreeswijk and Erik de Wilde focused their meta-analysis on differences between clinical and nonclinical groups on the Autobiographical Memory Test (AMT) (2004).

An intermediate strategy is for the meta-analyst to define one or more sets of constructs that include different operationalizations but distinguish different content domains of interest. Effect sizes for relationships involving those constructs would be coded and analyzed, but study results involving other constructs would be ignored. In a meta-analysis asking if physical activity attenuates the relationship between sedentary time and mortality, for example, Ulf Ekelund and his colleagues focused exclusively on the constructs of physical activity, sitting time, and mortality (2016). They accepted a range of measures of each of these constructs, but their meta-analysis did not include any of the other constructs that also happened to be reported in the contributing studies.

Whether global or highly specific construct categories are defined for coding effect sizes, the meta-analyst is likely to find some studies that report results on multiple measures within a given construct category. In such cases, one of several approaches can be taken. The analyst can simply code and analyze all the effect sizes contributed to a category, including multiple ones from the same study. This permits some studies to provide more effect sizes than others and, thus, to be overrepresented in the synthesis results. It also introduces statistical dependencies among the effect sizes because some of them are based on the same subject samples. These issues will have to be dealt with in the analysis (see chapter 13). Alternatively, criteria can be developed for selecting the single most appropriate measure within each construct category and ignoring the remainder (despite the loss of data that entails). Or multiple effect sizes within a construct category can be averaged (perhaps using a weighted average) to yield a single mean or median effect size within that category for each study.

The most important consideration that must inform the approach a meta-analyst takes to the matter of study results embodied in multiple constructs and measures is the purpose of the meta-analysis. For some purposes the study results of interest may involve only one construct, as when a meta-analysis of research evaluating educational innovations focuses exclusively on achievement test results. For other purposes the meta-analyst may be interested in a broader range of study results and welcome diverse constructs and measures. For instance, achievement, attendance, and social-emotional outcomes may all be of interest in a meta-analysis of research on an educational intervention. Having a clear view of the purposes of a meta-analysis and adopting criteria for selecting and categorizing effect-size statistics consistent with those purposes are fundamental requirements of a good meta-analysis. These steps define the effect sizes the meta-analyst will be able to analyze and, hence, shape the potential findings of the meta-analysis itself.

8.1.2.2 Effect Sizes for Subsamples In addition to overall results, many studies may report breakdowns for one or more participant subsamples. For example, a validation study of a personnel selection test might report test-criterion correlations for participants in different racial groups, or a study of the effects of drug counseling for teenagers might report results separately for males and females. Coding effect sizes for each subsample potentially permits a more discriminating analysis of the relationship between participant characteristics and study findings than can be obtained from overall study findings alone.

Two considerations bear on the potential utility of effect sizes from subsamples. First, the variable distinguishing

subsamples must represent a dimension of potential interest to the meta-analyst. Effect sizes computed separately for male and female subsamples, for instance, will be of concern only if there are theoretical, practical, or exploratory reasons to examine gender differences in the magnitude of effects. Second, a participant breakdown that is of interest must be reported widely enough to yield a large enough number of effect sizes to permit meaningful analysis. If most studies do not report results separately for males and females, for example, the utility in coding effect sizes for those subsamples is limited.

As with effect sizes from multiple measures within a single study, effect sizes for multiple subsamples within a study pose issues of statistical dependency if the subsamples are not mutually exclusive—that is, if the subsamples share participants. Although the female subsample in a breakdown by gender would not share participants with the male subsample, it would almost certainly share participants with any subsample resulting from a breakdown by age. In addition, however, any feature shared by subsamples within a study—for example, being studied by the same investigator—can introduce statistical dependencies into the effect sizes computed for those subsamples. Analysis of effect sizes based on subsamples, therefore, must be treated like any analysis that uses more than one effect size from each study and thus potentially violates assumptions of statistical independence. This may be addressed by separately meta-analyzing effect sizes from each overlapping subsample, or by using one of the specialized techniques for handling statistically dependent effect sizes (see chapter 13)

8.1.2.3 Effect Sizes for Different Times of Measurement Some studies may report findings on the same variables measured on the same participant sample at different times. For example, a study of the effects of a smoking cessation program might report the outcome immediately after treatment and at six-month and one-year follow-ups. Such time-series information potentially permits an interesting analysis of temporal patterns in outcome—decay curves for the persistence of treatment effects, for example. Another situation that often produces effect sizes measured at different times on the same participant samples is meta-analysis of relationships reported in longitudinal studies. A meta-analysis of risk factors for adolescent depression, for example, might involve effect sizes for the relationships of risk factors measured at different times during childhood to depression measured in adolescence.

As with effect sizes for subsamples, the meta-analyst must determine whether the relationships of interest reported at different times vary widely enough to yield enough effect-size data to be worth analyzing. Even in favorable circumstances, only a portion of the studies in a synthesis may provide measures at different time points. A related problem is that the intervals between times of measurement may vary widely from study to study, making these results difficult to summarize across studies. If this occurs, one approach is for the meta-analyst to establish broad timing intervals and code each effect size in the category it most closely fits. Another approach is to code the time of measurement for each result as an interval from some common reference point—for instance, the termination of treatment. Still another approach is to capture the time trends with a summary statistic such as the slope on the effect sizes over time. These techniques allow the meta-analyst to then analyze the functional relationship between the effect sizes and the time that has passed since the completion of treatment.

Because the participants at time 1 will be much the same (except for attrition) as those measured at a later time, the effect sizes for the two occasions will not be statistically independent. As with multiple measures or overlapping subsamples, these effect sizes cannot be included together in an analysis that assumes independent data points unless special adjustments are made.

8.1.2.4 The Array of Study Results As the previous discussion indicates, the quantitative results from a study selected for meta-analysis may be reported for multiple constructs, for multiple measures of each construct, for the total subject sample, for various subsamples, and for multiple times of measurement. To effectively represent key study results and support interesting analyses, the meta-analyst must establish conceptual categories for each of these dimensions. These categories will group effect sizes judged to be substantially similar for the purposes of the meta-analysis, differentiate those believed to be importantly different, and ignore those judged to be irrelevant or uninteresting. Once these categories are developed, it should be possible to classify each effect size that can be coded from any study according to the type of construct, subsample, and time of measurement it represents. When enough studies yield comparable effect sizes, that set of results can be analyzed separately. Some studies will contribute only a single effect size (one measure, one sample, and one time) to the synthesis. Others will report more differentiated results and may yield a number of useful effect sizes.

It follows that different studies may provide effect sizes for different analyses on different categories of effect sizes. Any analysis based on effect sizes from only a subset of the studies under investigation raises a question of the generalizability of the findings to the entire set because of the distinctive features that may characterize that subset. For instance, in a meta-analysis of the correlates of effective job performance, the subset of studies that break out results by age groups might be mostly those conducted in large corporations. What is learned from them about age differences might not apply to employees of small businesses. The meta-analyst needs to be attentive to such distinctions when interpreting the results of the analysis. Studies that report differentiated results, nonetheless, offer the meta-analyst the opportunity to conduct more probing analyses of the issues of interest. Including such differentiation in a meta-analysis can add rich and useful detail to its findings.

8.1.3 Study Descriptors

Whatever the set of effect sizes under investigation, it is usually of interest to also examine matters related to the characteristics of the studies that yield those results. To accomplish this, the meta-analyst must identify and code information from each study about the particulars of its participants, methods, treatments, context, and the like. Meta-analysts have shown considerable diversity and ingenuity in the study characteristics they have coded and the coding schemes they have used. Specific discussion of the development, application, and validation of coding schemes is provided elsewhere in this volume (see chapter 9). What is described here are the general types of study descriptors that the meta-analyst might consider coding into the meta-analytic database for use in one or more of the various kinds of analysis described later in this chapter.

To provide a rudimentary conceptual framework for study descriptors, we first assume that study results are determined conjointly by the nature of the substantive phenomenon under investigation and the nature of the methods used to study it. The variant of the phenomenon selected for study and the particular methods applied, in turn, may be influenced by the characteristics of the researcher and the research context. These latter characteristics will be labeled as extrinsic factors because they are extrinsic to the substantive and methodological factors assumed to directly shape study results. In addition, the actual results and characteristics of a study must be fully and accurately reported for them to be validly coded in a synthesis. Various aspects of study reporting thus constitute a second type of extrinsic factor because they too may be associated with the study results represented in a synthesis though they are not assumed to directly shape those results.

This scheme, albeit somewhat crude, identifies four categories of study descriptors that may interest the meta-analyst: substantive features of the matter the studies investigate; particulars of the methods and procedures used to conduct those investigations; characteristics of the researcher and the research context; and attributes of the way the study is reported and the way it is coded. The most important of these, of course, are the features substantively pertinent to characterizing the phenomenon under investigation. In this category are such matters as the nature of the treatment provided; the characteristics of the participants; the cultural, geographical, or temporal setting; and those other influences that might moderate the relationships under study. It is these variables that permit the meta-analyst to identify the extent to which differences in the results of studies are associated with differences in the substantive characteristics of the situations they investigate. From such information the analyst may be able to determine that one treatment variation is more effective than another (or more effective for one type of outcome), that a relationship holds for certain types of participants or circumstances but not for others, or that there is a developmental sequence that yields different results at different time periods. Findings such as these, of course, result in better understanding of the nature of the phenomenon under study and are the objective of most syntheses.

Those study characteristics not related to the substantive aspects of the phenomenon involve various possible sources of distortion, bias, or artifact in study results as they are presented in the original research or the coding for the meta-analysis. The most important features of this sort are methodological or procedural aspects of the manner in which the studies were conducted. These include variations in the designs, research procedures, quality of measures, and forms of data analysis that might yield different results even if every study were investigating exactly the same phenomenon. A meta-analyst may be interested in examining the influence of method variables for two reasons. First, analysis of the relationships between method choices and study results provides useful information about which aspects of research procedures make the most difference and, hence, should be most carefully selected. Second, method differences confound

substantive comparisons among study results and may thus need to be included in the analysis as statistical control variables to help disentangle the influence of methods on those results.

Factors extrinsic to both the substantive phenomenon and the research methods include characteristics of the researcher (such as gender and disciplinary affiliation), the research circumstances (such as the nature of study sponsorship), and reporting (such as the form of publication). Such factors would not typically be expected to directly shape study results but nonetheless may be related to something that does and thereby will correlate with coded effect sizes. Whether a researcher is a psychologist or a sociologist, for example, should not itself directly determine study results. Disciplinary affiliation may, however, influence methodological practices or the selection of variants of the phenomenon to study, which in turn, could affect study results.

Another type of extrinsic factor involves aspects of the reporting of study methods and results that might yield different values in the meta-analyst's coding even if all studies were investigating the same phenomenon with the same methods. The available study documents may not report important details of the procedures, measures, treatments, or results. This requires coders to use coding conventions designed to yield the most likely option for the missing information (for example, if a study report does not say that random assignment was used, it probably was not) or to record missing values on items that, with better reporting, could be coded accurately. Insufficiencies in the information reported in a study, therefore, may influence the effect sizes coded or the representation of substantive and methodological features closely related to effect sizes in ways that distort the relationships of interest to the meta-analysis. At the extreme, publication bias, in which results from entire studies are not reported or the reports are not accessible to the meta-analyst, can remove studies with distinctive characteristics from the meta-analysis leaving an unrepresentative sample available for analysis (see chapter 18 in this volume).

A meta-analyst who desires a full description of study characteristics will want to make some effort to identify and code all those factors thought to be potentially related to the study results of interest. From a practical perspective, the decision about what specific study characteristics to code will have to reconcile two competing considerations. First, research synthesis provides a service by documenting in detail the nature of the body of research bearing on a given issue. This consideration motivates a coding of a broad range of study characteristics for descriptive purposes. On the other hand, many codable features of studies have limited value for anything but descriptive purposes: they may not be widely reported in the available research or may show little variation from study to study. Finally, of course, some study characteristics will simply not be germane to the meta-analyst's purposes (though it may be difficult to specify in advance what will prove relevant and what will not). Documenting study characteristics that are poorly reported, lacking in variation, or irrelevant to present purposes surely has some value, as the next section argues. However, given the time-consuming and expensive nature of coding, the meta-analyst inevitably must find some balance between coding broadly for descriptive purposes and coding narrowly around the specific target issues of the particular meta-analysis.

8.2 ANALYSIS OPPORTUNITIES

A researcher embarks on a meta-analysis to answer certain questions by systematically coding the effect sizes and characteristics of studies selected to be relevant to those questions and then statistically analyzing the resulting data. Many technical issues are of course involved in the statistical analysis; they are discussed elsewhere in this volume. This chapter simply provides an overview of the various types of analysis and the kinds of insights they might yield. With such an overview in mind, the meta-analyst should be in a better position to know what sorts of questions might be answered and what data must be coded to address them.

Four generic forms of analysis are outlined here. The first, descriptive analysis, uses the coded variables to provide an overall picture of the nature of the research studies included in the meta-analysis. The other three forms of analysis examine relationships among coded variables. As set out, variables coded in a synthesis can be divided into those that describe study results (effect sizes) and those that describe study characteristics (study descriptors). Three general possibilities emerge from this scheme: analysis of the relationships among study descriptors, analysis of the relationships between study descriptors and effect sizes, and analysis of the relationships among effect sizes.

8.2.1 Descriptive Analysis

By its nature, meta-analysis involves the collection of information describing key results and various important

attributes of the studies under investigation. Descriptive analysis of these data can provide an informative picture of the nature of the research selected for inclusion in the meta-analysis. It can also help identify issues that have already been sufficiently studied and gaps that need additional study. Further, it can highlight common methodological practices and provide a basis for assessing the areas in which improvement is warranted. The descriptive information obtained in a meta-analysis deals with either study results or study descriptors. Each of these is considered in turn in the following sections.

8.2.1.1 Study Results Descriptive information about study results is almost universally reported in meta-analyses. The central focus, of course, is the distribution of effect sizes across studies. As noted earlier, there may be numerous categories of effect sizes; correspondingly, there may be numerous effect-size distributions to describe and compare (see chapters 11 and 12 in this volume).

A useful illustration of descriptive analysis of effect sizes occurs in a meta-analysis of eighty-four studies on the association of alcohol consumption and cardiovascular disease outcomes (Ronksley et al. 2011). The authors display the distributions of relative risk effect sizes in forest plots for six outcomes along with the mean effect sizes for each outcome. These displays allow the reader to assess the uniformity of the effects across studies as well as their central tendency. Furthermore, the associated numbers of effect sizes and confidence intervals, which were also reported, allow an appraisal of the depth and statistical precision of the available evidence for each mean effect size.

8.2.1.2 Study Descriptors Meta-analysts all too often report relatively little information about studies other than their results. It can be quite informative, however, to provide breakdowns of the coded variables that describe substantive study characteristics, study methods, and extrinsic factors such as publication source. This disclosure accomplishes a dual purpose. First, it informs readers about the specific nature of the research that has been chosen for the meta-analysis so they may judge its comprehensiveness and biases. For example, knowing the proportion of unpublished studies, or studies conducted before a certain date, or studies with a particular type of design might be quite relevant to interpreting the findings of the meta-analysis. Also, meta-analysts frequently discover that studies do not adequately report some information that is important to the synthesis. It is informative for readers to know the extent of missing data on various coded variables.

Second, summary statistics for study characteristics provide a general overview of the nature of the research available on the topic addressed. This overview allows readers to ascertain whether researchers have tended to use restricted designs, measures, samples, or treatments and what the gaps in coverage are. Careful description of the research literature establishes the basis for critique of research practices in a field and helps identify characteristics desirable for future research.

Valerie Henderson and her colleagues used this descriptive approach to examine the methodological quality of preclinical research that uses animal models in cancer drug development (2015). Focusing on the preclinical efficacy studies of sunitinib, a drug widely used in cancer treatment, their meta-analysis summarized the characteristics of the studies with regard to sample sizes, statistical power, random assignment, blinding to condition, statistical analysis, and a range of items related to external validity. This information was then used to assess the adequacy of the available studies and identify needed improvements in methodology and reporting.

It is evident that much can be learned from careful description of study results and characteristics in meta-analysis. Indeed, it can be argued that providing a broad description and appraisal of the nature and quality of the body of research under examination is fundamental to all other analyses that the meta-analyst might wish to conduct. Proper interpretation of those analyses depends critically on a clear understanding of the character and limitations of the primary research on which they are based. It follows that conducting and reporting a full descriptive analysis should be routine in meta-analysis.

8.2.2 Relationships Among Study Descriptors

The various descriptors that characterize the studies in a meta-analysis may themselves have interesting interrelationships. It is quite unlikely that study characteristics will be randomly and independently distributed over the studies in a given research literature. More likely they will fall into patterns in which certain characteristics tend to occur together. Many methodological and extrinsic features of studies may be of this sort. We might find, for example, that published studies are more likely to be federally funded and have authors with academic affiliations. Substantive study characteristics may also cluster in interesting ways, as when certain types of treatments are more frequently applied to certain types of participants.

A more probing analysis of the interrelationships among study descriptors could focus on certain key study features and determine if they were functions of other temporally or logically prior features. For example, an analyst might use sample size as an outcome variable and examine the extent to which it is predictable from characteristics of the researcher and the nature of the research setting. Similarly, the meta-analyst could examine whether different treatments or variants of a treatment used in different studies were a function of the characteristics of the participants to whom they were applied, the treatment setting, and so forth. Analysis of cross-study interrelationships among study descriptors has not often been explored in meta-analysis. Nonetheless, such analysis should be useful both for data reduction—that is, creation of composite moderator variables—and to more fully describe the nature and patterns of research practices in an area of inquiry.

A distinctive aspect of analysis of the interrelations of study descriptors is that it does not involve the effect sizes that are central to most other forms of analysis in meta-analysis. An important implication of this feature is that the sampling unit for such analyses is the study itself, not the individual participant within a study. That is, features such as the type of design chosen, nature of treatment applied, publication outlet, and other similar characteristics usually describe the study, not the participants, and thus are not influenced by participant-level sampling error. This simplifies much of the statistical analysis when investigating these types of relationships because attention need not be paid to the varying number of participants represented in different studies.

8.2.3 Relationships Between Study Effect Sizes and Study Descriptors

The effect sizes that represent study results often show more variability within their respective categories than would be expected on the basis of sampling error alone (see chapter 12 of this volume). A natural and quite common question for the analyst is whether certain study descriptors are associated with the magnitude of the effects, that is, whether they are moderators of effect size. It is almost always appropriate for a synthesis to conduct a moderator analysis to identify at least some of the circumstances under which effect sizes are larger and smaller. Moderator analysis is essentially correlational, examining the covariation of selected study descriptors and effect sizes, though it can be conducted in various statistical formats. In such analysis, the study effect sizes become the dependent variables and the study descriptors become the independent or predictor variables.

The three broad categories of study descriptors identified earlier—substantive aspects, study methods and procedures, and extrinsic matters of research circumstances—are all potentially related to study effect sizes. The nature and interpretation of relationships involving descriptors from these three categories, however, is quite different.

8.2.3.1 Relationships with Extrinsic Variables Extrinsic variables, as defined earlier, are not generally assumed to directly shape actual study results even though they may differentiate studies that yield different effect sizes. In some instances, they may be marker variables associated, in turn, with research practices that do exercise direct influence on study results. It is commonly found in synthesis, for example, that published studies have larger effect sizes than unpublished studies (Rothstein, Sutton, and Borenstein 2005). Publication of a study, of course, does not itself inflate the effect sizes, but it may reflect the selection criteria and reporting proclivities of the authors, peer reviewers, and editors who decide whether and how a study will be published. Analysis of the relationship of extrinsic study variables to effect size, therefore, may reveal interesting aspects of research and publication practices in a field, but is limited in its ability to reveal why those practices are associated with different effect-size magnitudes.

A more dramatic example appeared in the literature some years ago. In a meta-analysis of gender differences in conformity, Alice Eagly and Linda Carli found that studies with a higher percentage of males among the authors were more likely to report greater conformity among females (1981). One might speculate that this reflects some bias exercised by male researchers, but it is also possible that the different results occurred because they tended to use different research methods than female researchers. Indeed, Betsy Becker demonstrated that male authorship was confounded with characteristics of the outcome measures and the number of confederates in the study (1986). Both of these method variables, in turn, were correlated with study effect sizes.

8.2.3.2 Relationships with Method Variables Research methods and procedures constitute a particularly important category of moderator variables too often underrepresented in synthesis. Experience has shown that variation in study effect sizes is often associated with methodological variation among the studies (Lipsey 2003). One reason these relationships are interesting is that much can be learned from synthesis about the connection

between researchers' choice of methods and the results those methods yield. Such knowledge provides a basis for examining research methods to discover which aspects introduce the greatest bias or distortion into study results.

Investigation of a key method variable in their meta-analysis of studies of perceptual processing of human faces using sequential matching tasks, for example, allowed Jennifer Richler and Isabel Gauthier to identify differences in results for two forms of that task, referred to as the partial design and the complete design, that favored the complete design (2014). Similarly, Laim Dougherty and David Shuker demonstrated that the results of mate selection studies in behavioral ecology depended in part on whether respondents were given sequential options presented individually or multiple options presented simultaneously (2015). On measurement issues, Andres De Los Reyes and his colleagues documented the low levels of agreement between reports by patients, parents, and teachers on the mental health symptoms of children and adolescents (2015). Findings such as these provide invaluable methodological information to researchers seeking to design valid research in their respective areas of study.

Another reason to be concerned about the relationship between methodological features of studies and effect sizes is that method differences may be correlated with substantive differences. In such confounded situations, a meta-analyst must be very careful not to attribute effect-size differences to substantive factors when those factors are also related to methodological factors. For example, in a meta-analysis of interventions for juvenile offenders, Mark Lipsey (2003) found that such substantive characteristics as the gender mix of the sample and the amount of treatment provided were not only related to the recidivism effect sizes but also to the type of research design used in the study (randomized or nonrandomized). Such confounding raises a question about whether the differential effects associated with gender and amount of treatment reflect real differences in treatment outcomes or only spurious effects stemming from the correlated methodological differences. Meta-analysts may have some ability to disentangle such confounding in their statistical analysis, but in any event should be alert to its occurrence and attempt to assess its influence on the results of the meta-analysis.

8.2.3.3 Relationships with Substantive Variables

Once disentangled from method variables, the category of substantive variables is usually of most interest to the meta-analyst. Determining that effect size is associated with type of subjects, treatments, or settings often has considerable theoretical or practical importance. A case in point is a meta-analysis of the effects of school-based interventions for promoting social and emotional learning (Durlak et al. 2011). Moderator analysis revealed, first, that larger effect sizes were associated with programs for which no implementation problems were reported. In regard to the characteristics of the programs, effect sizes were larger for interventions that included four recommended practices: a sequential coordinated set of activities, active forms of learning, at least one program component devoted to developing personal or social skills, and targeting specific social-emotional skills.

When multiple classes of effect-size variables as well as multiple classes of substantive moderator variables are in play, quite a range of analyses of their interrelations is possible. For example, investigation can be made of the different correlates of effects on different outcome constructs. Stephen Leff and his colleagues used this approach in a meta-analysis of the effects of different housing arrangements for persons with mental illness (2009). They examined four different housing models in relation to seven different outcomes and then explored differential effects for subgroups distinguished by gender, race, and presence of co-occurring substance disorders. Subgroup differences themselves may be fruitful terrain for differentiation by substantive moderator variables. The meta-analysis conducted by Emily Grijalva and her colleagues, for instance, investigated gender differences in narcissism (2015). They confirmed the widely held belief that men are more narcissistic than women and found that pattern to be stable over time and age. They then went further and differentiated four forms of narcissism and found that the gender difference was pronounced for only two of them.

The opportunity to conduct systematic analysis of the relationships between study descriptors and effect sizes is one of the most attractive aspects of meta-analysis. However, certain limitations must be kept in mind. First, a meta-analysis can examine only those variables that can be coded from primary studies in essentially complete and accurate form. Second, a variable must show enough variation across studies for such analysis to be feasible and meaningful; for example, if nearly all the studies were conducted on males, an analysis of sex differences would not be possible. Third, this aspect of meta-analysis is a form of observational study, not experimental study. Many of the moderator variables that show relationships with effect size may also be related to each other. Confounded relationships of that sort are inherently ambiguous with regard to which variables are the more influential ones or, indeed, whether their influence stems from other

unmeasured and unknown variables correlated with those examined. As noted, this situation is particularly problematic when method variables are confounded with substantive ones. Even without confounded characteristics, however, the observational nature of meta-analysis tempers the confidence with which the analyst can describe the causal influence of study characteristics on their findings (on the distinction between study-generated and synthesis-generated evidence, see chapter 2).

8.2.4 Relationships Among Effect Sizes

Study findings in meta-analysis are reported most often either as an overall mean effect size or as mean effect sizes for various categories of results. The availability of multiple classes of effect sizes offers the potential for informative analysis of the patterns of covariation among the effect sizes themselves. This can be especially fruitful when different effect sizes from the same studies represent quite different constructs. In that situation, a meta-analyst can examine whether the magnitude of effects on one construct is associated across studies with the magnitude of effects on another construct. Some studies of educational interventions, for instance, may measure effects on both achievement and student attitude toward instruction. Under such circumstances, a meta-analyst might wish to ask whether the intervention effects on achievement covary with the effects on attitude. That is, if an intervention increased the attitude scores for a sample of students, did it also improve their achievement scores and vice versa? However, a cross-study correlation, though potentially interesting, does not necessarily imply a within-study relationship. Thus, achievement and attitude might covary across studies, but not covary among the students in the samples within the studies.

A simple example of relationships between effect sizes is found in the meta-analysis by van Vreeswijk and de Wilde (2004) of differences between clinical and nonclinical groups on their scores on the AMT and on measures of depression. The AMT yields two scores, one on specific positive memories and one on negative overgeneral memories, and effect sizes were coded for both. Cross-study correlations between the effect sizes showed that studies reporting large differences on specific positive memories between clinical and nonclinical groups tended to report small differences in depression scores. On the other hand, studies reporting large differences on negative overgeneral memories tended to also report large differences on depression scores. This was taken as evidence that depressed mood might well mediate performance on the AMT.

An even more ambitious form of meta-analysis may be used to construct a correlation matrix that represents the interrelations among many variables in a research literature. Each study in such a synthesis contributes one or more effect sizes representing the correlation between two of the variables of interest. The resulting synthetic correlation matrix can then be used to test multivariate path or structural equation models. An example of such an analysis is a meta-analysis of bivariate correlations coded from studies that conducted mediational analysis for the effects of mindfulness interventions on mental health and well-being (Gu et al. 2015). The authors then synthesized correlation matrices for the studies contributing relevant correlations and used meta-analytic structural equation modeling to test the mediational relationships of interest. Of seven candidate mediators, mindfulness and repetitive negative thinking were found to be significant mediators of mental health outcomes.

A potential problem in analyzing multiple categories of study results is that all categories of effects will not be reported by all studies. Thus, each synthesized effect size examining a relationship will most likely draw on a different subset of studies creating uncertainty about their generality and comparability. This and related issues having to do with analyzing synthesized correlation matrices are discussed elsewhere in this volume (see chapter 16).

8.3 CONCLUSION

Meta-analysis can be thought of as a form of survey research in which the participants interviewed are not people but research reports. The meta-analyst prepares a questionnaire of items of interest, collects a sample of research reports, interacts with those reports to determine the appropriate response on each item, and analyzes the resulting data. The kinds of questions that can be addressed by meta-analysis of a given research literature are thus determined by the variables that the meta-analyst is able to code and the kinds of analyses that are possible given the nature of the resulting data.

This chapter examines the types of variables potentially available from research studies, the kinds of questions that might be addressed using those variables, and the forms of analysis that investigate those questions. Its purpose has been to provide an overview that will help the prospective meta-analyst select appropriate variables for coding and plan probing and interesting analyses of the resulting data. Contemporary meta-analysis often neglects important

variables and analysis opportunities. As a consequence, we learn less from such work than we might. Even worse, what we learn may be erroneous if confounds and alternative analysis models have been inadequately probed. Current practice has only begun to explore the breadth and depth of knowledge that well-developed meta-analyses can potentially yield.

8.4 REFERENCES

Becker, Betsy J. 1986. "Influence Again: An Examination of Reviews and Studies of Gender Differences in Social Influence." In *The Psychology of Gender: Advances through Meta-Analysis*, edited by Janet S. Hyde and Marcia C. Linn. Baltimore: Johns Hopkins University Press.

De Los Reyes, Andres, Tara M. Augenstein, Mo Wang, Sarah A. Thomas, Deborah A. G. Drabick, Darcy E. Burgers, and Jill Rabinowitz. 2015. "The Validity of the Multi-informant Approach to Assessing Child and Adolescent Mental Health." *Psychological Bulletin* 141(4): 858–900.

Dougherty, Liam R., and David M. Shuker. 2015. "The Effect of Experimental Design on the Measurement of Mate Choice: A Meta-Analysis." *Behavioral Ecology* 26(2): 311–19.

Durlak, Joseph A., Roger P. Weissberg, Allison B. Dymnicki, Rebecca D. Taylor, and Kriston B. Schellinger. 2011. "The Impact of Enhancing Students' Social and Emotional Learning: A Meta-Analysis of School-Based Universal Interventions." *Child Development* 82(1): 405–32.

Eagly, Alice H., and Linda L. Carli. 1981. "Sex of Researchers and Sex-Typed Communications as Determinants of Sex Differences in Influenceability: A Meta-Analysis of Social Influence Studies." *Psychological Bulletin* 90(1): 1–20.

Ekelund, Ulf, Jostein Steene-Johannessen, Wendy J. Brown, Morten W. Fagerland, Neville Owen, Kenneth E. Powell, Adrian Bauman, and I-Min Lee. 2016. "Does Physical Activity Attenuate, or Even Eliminate, the Detrimental Association of Sitting Time with Mortality? A Harmonised Meta-Analysis of Data from More than 1 Million Men and Women." *Lancet* 388(10051): 1302–10.

Erford, Bradley T., Erin Johnson, and Gerta Bardoshi. 2016. "Meta-analysis of the English Version of the Beck Depression Inventory—Second Edition." *Measuring and Evaluation in Counseling and Development* 49(1): 3–33.

Grijalva, Emily, Daniel A. Newman, Louis Tay, M. Brent Donnellan, P. D. Harms, Richard W. Robins, and Taiyi Yan. 2015. "Gender Differences in Narcissism: A Meta-Analytic Review." *Psychological Bulletin* 141(2): 261–310.

Gu, Jenny, Clara Strauss, Rod Bond, and Kate Cavanagh. 2015. "How Do Mindfulness-based Cognitive Therapy and Mindfulness-Based Stress Reduction Improve Mental Health and Well-being? A Systematic Review and Meta-Analysis of Mediation Studies." *Clinical Psychology Review* 37(1): 1–12.

Henderson, Valerie C., Nadine Demko, Amanda Hakala, Nathalie MacKinnon, Carole A. Federico, Dean Fergusson, and Jonathan Kimmelman. 2015. "A Meta-Analysis of Threats to Valid Clinical Inference in Preclinical Research of Sunitinib." *eLife* 4:e08351.

Leff, H. Stephen, Clifton M. Chow, Renee Pepin, Jeremy Conley, I. Elaine Allen, and Christopher A. Seaman. 2009. "Does One Size Fit All? What We Can and Can't Learn from a Meta-Analysis of Housing Models for Persons with Mental Illness." *Psychiatric Services* 60(4): 473–82.

Lipsey, Mark W. 2003. "Those Confounded Moderators in Meta-Analysis: Good, Bad, and Ugly." *Annals of the American Academy of Political and Social Science* 587 (May): 69–81.

Richler, Jennifer J., and Isabel Gauthier. 2014. "A Meta-analysis and Review of Holistic Face Processing." *Psychological Bulletin* 140(5): 1281–302.

Ronksley, Paul E., Susan E. Brien, Barbara J. Turner, Kenneth J. Mukamal, and William A. Ghali. 2011. "Association of Alcohol Consumption with Selected Cardiovascular Disease Outcomes: A Systematic Review and Meta-Analysis." *British Medical Journal* 342: d671.

Rothstein, Hannah R., Alexander J. Sutton, and Michael Borenstein, eds. 2005. *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. New York: John Wiley & Sons.

Smith, Mary L., and Gene V. Glass. 1977. "Meta-Analysis of Psychotherapy Outcome Studies." *American Psychologist* 32(9): 752–60.

van Vreeswijk, Michiel F., and Erik J. de Wilde. 2004. "Autobiographical Memory Specificity, Psychopathology, Depressed Mood, and the Use of the Autobiographical Memory Test: A Meta-Analysis." *Behavior Research and Therapy* 42(6): 731–43.

9

SYSTEMATIC CODING FOR RESEARCH SYNTHESIS

DAVID B. WILSON
George Mason University

CONTENTS

9.1 INTRODUCTION

A research synthesist has collected a set of studies that address a similar research question and wishes to code the studies to create a dataset suitable for systematic review and meta-analysis. This task is analogous to interviewing, but rather than a person being interviewed the study is. Interrogating might be a more apt description, though it is usually the coder who loses sleep because of the study and not the other way around. The goal of this chapter is to provide synthesists with practical advice on designing and developing a coding protocol suitable to this task. A coding protocol is both the coding forms, whether paper or computerized, and the coding manual providing instructions on how to apply coding form items to studies.

A well-designed coding protocol will describe the characteristics of the studies included in the research synthesis and will capture the pertinent findings relevant to the research question in a fashion suitable for comparison and synthesis, ideally using meta-analysis. Stated differently, the goal of a good coding protocol is to encode information about the methodological aspects of the study, the characteristics of the participants, interventions or experimental manipulations, measured variables, and other theoretically relevant features of the study or its content. A coding protocol typically focuses on ways in which the studies may differ from one another. The research context may vary across studies, for example. Studies may use different operationalizations of one or more of the critical constructs. Capturing variations in settings, participants, methodology, experimental manipulations, and measured variables is an important goal of the coding protocol not only for careful description but also for use in the analysis to explain variation in study findings (see chapter 8). More fundamentally, the coding protocol serves as the foundation for the transparency and replicability of the synthesis.

9.2 IMPORTANCE OF TRANSPARENCY AND REPLICABILITY

A core feature of the scientific method is study replicability and reproducibility. Setting aside the possibility of fraud, any given study may have results that are simply inaccurate (that is, arrived at the incorrect conclusion) through no fault of the scientists conducting the work. The ability to replicate a study, either with identical methods or through intentional variation in the methodological approach, is key to establishing whether the results of a study are reproducible because it is through replication that the collective body of literature becomes more accurate. Such replication requires transparency and full reporting of the methods used by a study.

The importance of replicability and reproducibility has taken center stage in recent years in both the physical and social sciences. The presidential address by Ralph Cicerone at the annual National Academy of Sciences meeting in 2015 was on the topic of research reproducibility, replicability, and reliability. In this address, Cicerone discussed the recent efforts at reproducing important findings in several fields and the high failure rates of these efforts. The recent attempt to replicate one hundred experimental and correlational studies in psychology by the Open Science

Collaboration is an important example (2015). This study showed that replicated effects were, on average, half the size of original effects. Furthermore, only 39 percent of the findings from the original hundred studies were subjectively judged as having been reproduced. Rather than an indictment on the field of psychology (or science more generally), this work reinforces the need for replications. As Cicerone points out, science is self-correcting but the process requires replications.

Research synthesis is critical to self-correction because it provides a credible method of assessing the reproducibility of results across a collection of studies that are either pure replications of one another, or are conceptual replications (that is, studies examining the same basic empirical relationship, albeit with different operationalizations of key constructs and variation in method). That said, research synthesis is a scientific method and should itself also be held to the same standard. This can be achieved only through the use of transparent methods that are fully reported in sufficient detail to allow others to critically assess the procedures used (transparency) and to conduct an identical or similar study (replication) to establish the veracity of the original findings (reproducibility). This should not be interpreted to imply that research synthesis is merely the technical application of a set of procedures. The numerous decisions made in the design and conduct of a research synthesis requires thoughtful scholarship. A high-quality research synthesis, therefore, explicates these decisions sufficiently to allow others to fully understand the basis for the findings and to be able to replicate the synthesis if they so desire.

The coding protocol that synthesists develop is a critical aspect of the transparency needed to enable replication. Beyond detailing how studies will be characterized, the coding protocol documents the procedures used to extract information from studies and gives guidance on how to translate the narrative information of a research report into a structured and (typically) quantitative form. Although few journals will publish a full coding protocol, these should be made available either electronically from the author or via a journal's online supplemental materials.

Coding studies is challenging and requires subjective decisions. A decision that may on the surface seem straightforward, such as whether a study used random assignment to experimental conditions, often requires judgments based on incomplete information in the written reports. One goal of a good coding protocol is to provide guidance to coders on how to handle these ambiguities. As a scientific activity, appropriately skilled and adequately trained coders should be able to independently code a study and produce essentially the same coded data. Replication is not possible if only the author of the coding protocol can divine its nuances. In the absence of explicit coding criteria, researchers who claim to know a good study when they see one, or to know a good measure of leadership style by looking at the items, or to determine whether an intervention was implemented well by reading the report may be a good scholar, but they are not engaged in scientific research synthesis. Because it is impossible to know what criteria were invoked in making these judgments, her decisions lack transparency and cannot be replicated by others. The expertise of an experienced scholar should be reflected in the coding protocol (the specific coding items and decision rules for how to apply those items to specific studies) and not in opaque coding that cannot be scrutinized by others. For example, a "yes" and "no" coding item that asks whether a treatment intervention was well implemented is vague and will almost certainly be coded inconsistently across coders. In contrast, a set of items that ask whether the reported discussed implementation fidelity, whether any data were collected regarding the integrity of the treatment, whether critical aspects of the treatment were delivered to at least a specified proportion of the participants, and so on increases transparency of the coding process and improves replicability across coders. Detailed coding guidelines should be developed for any item that requires judgment on the part of the coders.

9.3 CODING ELIGIBILITY CRITERIA

The potentially eligible studies that a synthesist identifies through the search and retrieval process should be coded against explicit eligibility criteria. Assessing each retrieved study against the elements of the eligibility criteria and recording this information on an eligibility screening form or into an eligibility database provides an important record of the nature of studies excluded from a research synthesis. Based on this information, the synthesist can report on the number of studies excluded and the reasons for the exclusions. Many journals now require a flow diagram, such as the CONSORT, that details the number of studies excluded at various decision points (Moher et al. 2001). I have found this type of information to be useful when responding to a colleague who questions why a particular study was not included in the synthesis, enhancing the transparency of the review.

The eligibility criteria should flow naturally from the research question or purpose of the synthesis (see chapter 2). Your research question may be narrow or broad and this will be reflected in the eligibility criteria. For example, a research synthesis might focus narrowly on social psychological laboratory studies of the relationship between alcohol and aggressive behavior that use the competitive reaction time paradigm, or on the effectiveness of a specific drug relative to a placebo for a clearly specified patient population and outcome. At the other end of the continuum would be a research syntheses focused on broad research questions such as the effect of class size on achievement or juvenile delinquency interventions on criminal behavior. In each case, the eligibility criteria should clearly elucidate the boundaries for the synthesis, indicating the characteristics studies must have to be included and characteristics that would exclude studies.

9.3.1 Study Features to Be Explicitly Defined

Eligibility criteria should address several issues, including the defining feature of the studies of interest, the eligible research designs, any sample restrictions, required statistical data, geographical and linguistic restrictions, and any time frame restriction. I briefly elaborate on each.

9.3.1.1 Defining Features of the Empirical Relationship of Interest Research questions typically address a specific empirical relationship or empirical finding that is the defining feature of the studies included in the synthesis. Eligibility criteria should clearly elaborate on the nature of the empirical finding that defines the parameters for study inclusion. For example, a meta-analysis of the effectiveness of an intervention, such as cognitive-behavioral therapy for childhood anxiety, needs to clearly delineate the characteristics that an intervention must have to be considered cognitive-behavioral. Similarly, a meta-analysis of group comparisons, such as the relative math achievement of boys and girls, needs to define the parameters of the groups being compared and the dependent variables of interest. The prior examples focused on an empirical relationship, such as the effect of an intervention or social policy on one or more outcomes or the correlation between an employment test and job performance. In some cases, a single statistical parameter may be of interest, such as the prevalence rate of alcohol or illegal drugs in homicide victims, although meta-analyses of this type are less common.

Specifying the defining features of the eligible studies involves specifying both the nature of an independent variable, such as an intervention or employment test, and one or more dependent variables. The level of detail needed to adequately define these will vary across syntheses. The more focused your synthesis, the easier it will be to clearly specify the construct boundaries. Broad syntheses will require more information to establish the guidelines for determining which studies to include and which to exclude.

9.3.1.2 Eligible Designs and Required Methods What is the basic research design that the studies have in common? If multiple design types are allowable, what are they? A research synthesis of intervention studies may be restricted to experimental (randomized) designs or may include quasi-experimental comparison group designs. Similarly, a synthesis of correlational studies may be restricted to cross-sectional designs, or only longitudinal designs, or it may include both types. Some research domains have a plethora of research designs examining the same empirical relationship. The eligibility criteria need to specify which of these designs will be included and which will not, and to provide a justification for both inclusion and any restrictions.

It is rarely adequate to simply list the eligible research designs, because design names may have different meanings for scholars from different disciplinary backgrounds. For example, I have seen authors state that quasi-experimental designs were eligible with no elaboration of what is meant. Are only comparison group designs eligible? What about one-group, before-and-after designs? Some view the latter as pre-experimental whereas others view the latter as a quasi-experimental design. Rather than quibbling over the correct definition, simply provide a description of what is meant by each type of eligible design unless the meaning is unmistakably clear. This is consistent with guidance in the *Cochrane Handbook* (Higgins and Green 2011, section 13.2.2).

It is easy to argue that only the most rigorous research designs should be included in a synthesis. There are tradeoffs, however, between rigor and inclusiveness. Tradeoffs need to be evaluated within the context of the research question and the broader goals of the synthesis. For example, including only well-controlled experimental evaluations of a social policy intervention may reduce external validity because the level of control needed to mount such a study may alter the nature of the intervention and the context within which the intervention occurs. In this situation, quasi-experimental comparison group studies may provide valuable information about the likely effects of the social policy in more natural settings (see chapter 7).

9.3.1.3 Key Sample Features This criterion addresses the critical features of the sample or study participants. For example, are only studies involving humans eligible? In conducting a research synthesis of the effects of alcohol on aggressive behavior, Mark Lipsey and his colleagues find many studies involving fish—intoxicated fighting fish to be more precise (1997). Most syntheses of social science studies will not need to specify that they are restricted to studies of humans, but may well need to specify which age groups are eligible for inclusion and other characteristics of the sample, such as diagnostic status. A decision rule may need to be developed for handling studies with samples that include a small number of participants outside the target sample. For example, studies of juvenile delinquency programs are not always clearly restricted to youth age twelve to eighteen. Some studies include a small number of nineteen- and twenty-year-olds. The sample criterion should specify how these ambiguous cases are to be handled. The researchers could decide to exclude these studies, to include them only if the results for youth age twelve to eighteen are presented separately, or to include them if the proportion of nineteen- and twenty-year-olds does not exceed some threshold value.

9.3.1.4 Required Statistical Data A frustrating aspect of conducting a meta-analysis is missing statistical data. It is common to find an otherwise eligible study that does not provide the statistical data needed for computing an effect size. It is often possible, particularly for more recent studies, to obtain missing information directly from the original authors. With some work, it is also often possible to reconfigure the data provided to compute an effect size, or at least a credible estimate (see Lipsey and Wilson 2001). However, at times the necessary statistical data are simply not available.

Eligibility criteria must indicate how to handle studies with inadequate data. There are several possibilities: excluding the study from the synthesis, including it either in the main analysis or as a sensitivity analysis (see chapter 12) with some imputed effect-size value, or including it in the descriptive aspects of the synthesis but excluding it from the effect-size analyses. The advantages and disadvantages of different options are beyond the scope of this chapter but missing data are dealt with elsewhere in this volume (see chapter 17). It is recommended that missing data be included in some fashion even if simply in the descriptive systematic review rather than the meta-analytic synthesis.

9.3.1.5 Geographical and Linguistic Restrictions It is important to specify any geographical or linguistic restrictions for the synthesis. For example, will only English language reports be considered? Does the locale for the study matter? Youth engaged in delinquency is a universal problem: delinquent youth are found in every country and region of the world. However, the meaning of delinquency and the societal response to it are culturally embedded. Because they are, a research synthesis of juvenile delinquency programs needs to consider the relevance of studies conducted outside the synthesist's cultural context. Any geographical or linguistic restriction should be based on theoretical and substantive issues. Many non-English journals provide an English abstract, enabling the identification of potentially relevant studies in these journals by those who are monolingual. Restricting a review to English-language reports may introduce a source of bias to the review because studies from non-English locales that are published in English may differ systematically from those published in the native language. For example, Peter Jüni and his colleagues find that non-English language trials tended be have a smaller mean effect (2002). However, Andra Morrison and her colleagues fail to find a consistent English language bias (2012). Free online translation programs may also be useful, at least for assessing eligibility (such as http://www.freetranslation.com or http://translate.google.com).

9.3.1.6 Time Frame It is important that a synthesist specify the time frame for eligible studies. Further, the basis for this time frame should not be arbitrary or what is convenient for the authors. Issues to consider are whether studies prior to a particular date generalize to the current context and whether constructs with similar names have shifted in meaning enough over time to reduce comparability. Similarly, social problems may evolve in ways that reduce the meaningfulness of older research to the present. In some research domains, the decision regarding a time-frame restriction is easy: research on the topic began at a clearly defined historical point and only studies from that point forward are worth considering. In other research domains, a more theory-based argument may need to be put forth to justify a somewhat arbitrary cut point.

Whether the year of the report or year the data were collected is the critical issue. For some research domains, the two can be quite different. Unfortunately, it is not always possible to determine when data were collected, necessitating a decision rule for these cases.

I have often seen published meta-analyses that restrict the time frame to studies conducted or published after the date of a prior meta-analysis or widely cited synthesis. The logic seems to be to establish what the newer literature

"says" on the topic. This is sensible if the research in this area changed in some meaningful way following the publication of the synthesis. The goal should be to only exclude studies based on some meaningful substantive or methodological basis, not an arbitrary one such as the date of a prior synthesis.

I did not include publication type as an eligibility dimension in need of specification. A quality research synthesis will include all studies meeting the explicit eligibility criteria independent of publication status (for a complete discussion of the publication selection bias issue, see chapter 18).

9.3.2 Refining Eligibility Criteria

It can be difficult to anticipate all essential dimensions to be specified in the eligibility criteria. I have had the experience of evaluating a study for inclusion and found that it does not fit with the literature being examined yet meets the eligibility criteria. This situation generally arises when the criteria are not specific enough in defining the essential features of the studies. In such cases, the criteria can be modified and reapplied to studies already evaluated. It is important, however, that any refinement of the eligibility criteria not be motivated by a desire to include or exclude a particular study based on its findings. It is for this reason that the Cochrane Collaboration, an organization devoted to the production of systematic reviews in health care, requires that any modifications to eligibility criteria made once study selection has begun be clearly explained and justified in the final report.

9.4 DEVELOPING A CODING PROTOCOL

Developing a good coding protocol is much like developing a good survey questionnaire: it requires a clear delineation of what is important to measure and a willingness to revise and modify initial drafts. Before opening your favorite word processor and beginning to write, you need to list the constructs or study characteristics that are important to measure, much as one would when constructing a questionnaire. Many of these constructs may require multiple items. An overly general item is difficult to code reliably, as the generality allows each coder to interpret the item differently. For example, asking coders to evaluate the internal validity of a study on a 5-point scale is a more difficult coding task than asking a series of low inference questions related to internal validity, such as the nature of assignment to conditions, extent of differential attrition, the use of statistical adjustments to improve baseline equivalence, and the like. One can, of course, create composite scales from the individually coded items (for cautions regarding the use of quality scales, see chapter 7). After delineating the study characteristics of interest, a synthesist should draft items for the coding protocol and circulate it to other research team members or research colleagues (particularly those with substantive knowledge of the research domain being synthesized) for feedback. The protocol should be refined based on this feedback. The synthesist should also explore whether existing coding protocols have coding items and measures that fit the constructs of interest and, more important, whether they report any psychometrics for these items. Borrowing quality items from other synthesist may not only help avoid reinventing the wheel but also enhance the overall quality of the coding items.

Sharon Brown, Sandra Upchurch, and Gayle Acton propose an alternative approach that derives the coding categories from the studies themselves (2003). Their method begins by listing the characteristics of the studies into a matrix reflecting the general dimensions of interest, such as participant type, variations in the treatment, and the like. This information is then examined to determine how best to categorize studies on these critical dimensions. That is, rather than start with a priori notions of what to code based on the synthesist's ideas about the literature, Brown and her colleagues recommend reading through the studies and extracting descriptive information about the essential characteristics of interest. The actual items and categories of the coding protocol are then based on this information. This approach may be particularly helpful in developing items designed to describe the more complex aspects of studies, such as treatment variations and operationalizations of dependent measures. However, doing so might miss theoretically relevant items that may not be evident from a reading of the eligible studies.

It is useful to think of a coding protocol in much the same way as any measure developed for a primary research study. Ideally, it will be valid and reliable: measure what it is intended to measure and do so consistently across studies. Validity and reliability can be enhanced by developing coding items that require as little inference as possible on the part of the coders and allow coders to record the requested information in a way that corresponds closely to the way in which the characteristic of interest is typically described within the literature. For example, studies conducted in a school context may be more likely to report the grade range than the age of the

students. Thus, a coding form structured around the grades rather than the age of the sample will thus be easier to code and likely include fewer coding errors. The number of inference that a coder needs to make can be reduced by breaking apart the coding of complex study characteristics into a series of focused and simpler decisions. For example, rather than having a single item reflecting degree of selection bias, a coding protocol could include a series of questions about each important aspect of selection bias, such as method of assignment to conditions, attrition (both overall and differential), similarity of groups at baseline, and so forth. Each of these individual items requires fewer inferences. For closed-ended coding items (that is, items with a fixed set of options), the correspondence between the items and manner in which the item is described in the literature can be increased by developing the items from a sample of studies. Alternatively, an open-ended question can be used. With this approach, coders record the relevant information as it appears in the study report and specific coding items or categories are created from an examination of these open-ended responses (that is, the verbatim text extracted from the reports).

It is important to keep in mind the purpose that each coded construct serves as part of the research synthesis, as it is possible, and potentially wasteful, to code too many items. Some constructs may be important to code simply for their descriptive value, providing a basis for effectively summarizing the characteristics of the collection of studies. Other constructs are primarily coded for use as moderators in effect-size analyses. The number of possible moderator analyses is often limited unless the meta-analysis has a large number of studies. Thus it is wise to consider carefully the potential value of each item you plan to code. You can always return to the studies and code an additional item if it becomes necessary.

Although the specific items included in the coding protocol of each research synthesis vary, most address several general categories: report identification, study setting, participants, methodology, treatment or experimental manipulation, dependent measures, and effect-size data. I briefly discuss the types of information you might consider for each of these categories.

9.4.1 Types of Information to Code

9.4.1.1 Report Identification Although it is rather mundane, you will want to assign each study an identification code. A complication is that some research studies are reported in multiple manuscripts (such as a journal article and technical report or multiple journal articles), and some written reports present the results from multiple studies or experiments. The primary unit of analysis for a research synthesis will be an independent research study, that is, a study sample that does not overlap with the sample in another study in the review. For example, one report on an experimental study may provide data at posttest, whereas another report on the same sample may report data at follow-up. These two reports should be considered a single study, as should multiple publications that report different outcome measures on the same study sample. Because of the potential for multiple reports for a single study, the synthesist needs to create a system that allows tracking each manuscript and each research study. The coding forms should record basic information about the manuscript, such as the authors' names, type of manuscript (journal article, dissertation, technical report, unpublished manuscript, and so on), year of publication, and year or years of data collection (if possible).

9.4.1.2 Study Setting Research is always conducted in a specific context and a synthesist may wish to code aspects of this setting. For example, it may be of value to record the geographic location where the study was conducted, particularly if the synthesis includes studies from different countries or continents. The degree of specificity with regard to geographic location will depend on the research question and nature of the included studies. Related to this would be the nature of the institution in which the study was conducted, such as a school, correctional institution, or hospital. In some research areas, such as intervention studies, the research may vary in the degree of naturalness of the research setting. For example, an experimental test of child psychotherapy may occur entirely in a lab setting, such as the research lab of a clinical psychology professor, and rely on participants recruited through newspaper ads or be set in a natural clinic setting involving participants seeking mental health services (for an example of a meta-analysis that compared the results from children's psychotherapy studies in the lab against those in mental health clinics, see Weisz, Weiss, and Donenberg. 1992). Other potentially interesting contextual issues are the academic affiliation of the researcher (such as, disciplinary affiliation, in-house researcher or outside academic, and the like) and source of funding, if any. Each of these contextual issues may be related in some way to the findings. For example, researchers with a strong vested interest in the success of the program may report more positive findings than other researchers. Dennis Gorman provides a recent discussion

of this issue within the context of prevention research and provides evidence that conflicts of interest are related to reporting of more positive findings and the use of more flexible statistical methods (2016).

9.4.1.3 Participants Only in a very narrowly focused research synthesis will all of the subject samples be identical on essential characteristics. Thus, the coding protocol should include items that address the variations in participant samples that occur across the included studies. The aggregate nature of the data at the study level limits the ability to capture essential subject features. For example, age of the study sample is likely to be of interest. Study authors may report this information in a variety of ways, including the age range, the school grade or grades, the mean age, the median age, or they may make no mention of age. The coding protocol needs to be able to accommodate these variations and have clear guidelines for coders on how to characterize the age of the study sample. For example, a synthesist may develop guidelines that indicate that if the mean age is not reported, it should be estimated from the age range or school grade (such as, a sample of kindergarten students in the United States could be coded as having an average age of 5.5). The specific sample characteristics coded should depend on the research question and the nature of the studies included, but other variables to consider are gender mix, racial mix, risk or severity level and how this was assessed or specified, any restrictions placed on subject selection (such as, clinical diagnosis), and motivation for participation in the study (such as, course credit, volunteers, court mandated, paid participants, and the like). It is also important to consider whether the summary statistics adequately measure characteristic of interest and whether they are potentially biased. For example, mean age for samples of college students may be skewed by a small number of older students. Knowing how best to capture sample characteristics will depend on substantive knowledge of the research domain being synthesized.

9.4.1.4 Methodology Research studies included in your research synthesis will often vary in certain methodological features. Research synthesists will want to develop coding items to reflect that variability so that they can assess the influence of method variation on observed results. The importance of methodological features in explaining variability in effects across studies has been well established empirically (Cheung and Slavin 2015; Lipsey and Wilson 1993; Heinsman and Shadish 1996; Shadish and Ragsdale 1996; Weisburd, Lum, and Petrosino 2001; Wilson and Lipsey 2001).

The focus in coding should be on the ways in which the studies in the sample differ methodologically. Things to consider coding include the basic research design, nature of assignment to conditions, subject and experimenter blinding, attrition from baseline (both overall and differential), crossovers, dropouts, other changes to assignment, the nature of the control condition, and sampling methods. Each research domain will have unique method features of interest, and as such the features that should be coded will differ. For example, in conducting a meta-analysis of analog studies of the relationship between alcohol consumption and aggressive behavior, it would be important to code the method used to keep the participants unaware of their experimental condition. In other meta-analyses, coding the type of quasi-experimental design or whether the study used a between-subjects or within-subjects design might be essential. It is also critical to code items specifically related to methodological quality and risk of bias. The issues involved in assessing methodological quality are addressed in detail elsewhere in this volume (chapter 7).

9.4.1.5 Treatment or Experimental Manipulation If studies involve a treatment or experimental manipulation, the synthesist will want to develop coding items to fully capture any differences between the studies in these features. It is useful to have one or more items that capture the general type or essential nature of the treatment or experimental manipulation in addition to more detailed codes. For example, the studies included in our meta-analysis of experimental studies on the effects of alcohol on aggressive behavior used one of two basic types of experimental manipulations: a competitive-reaction time paradigm or a teacher-learner paradigm (Lipsey et al. 1997). An initial coding item captured this distinction and additional items addressed other nuances within each paradigm. As an additional example, a meta-analysis of school-based drug and delinquency prevention programs first categorized the programs into basic types, such as instructional, cognitive-behavioral or behavioral modeling, and the like (Wilson, Gottfredson, and Najaka 2001). Additional coding items were used to evaluate the presence or absence of specific program elements. Other issues to consider are the amount and intensity of treatment, who delivers the treatment, and how effective implementation of the treatment or experimental manipulation was assessed. In my experience meta-analyzing treatment effectiveness research, study authors inadequately report on the implementation integrity of treatment being evaluated. However, it may still be of interest

to document what is available and if possible examine whether effect size is related to the integrity of the treatment or other experimental manipulation.

9.4.1.6 Dependent Measures Studies will have one or more measured variable. Not all measures used in a study will necessarily be relevant to the research question. The coding manual therefore needs to clearly elaborate a decision rule for determining which measures to code and which to ignore.

For each measure to be coded, a synthesist will want to develop items that capture the essential construct measured and the specific measure used. Note that different studies may operationalize a common construct differently. For example, the construct of depression may be operationalized as scores on the Hamilton Rating Scale for Depression, the Beck Depression Inventory, or a clinician rating. A coding protocol should be designed to capture both the characteristics of the construct and its operationalization. It also may include other information about the measure, such as how the data were collected (such as self-report, observer rating, or clinical interview), psychometric properties (such as reliability and validity), whether the measure was a composite scale or based on a single item, and the scale of the measure (that is, dichotomous, discrete ordinal, or continuous). In many situations it is also essential to know the timing of the measurement relative to the experimental manipulation or treatment (such as, baseline, three months, six months, twelve months, and the like). For measures of past behavior, the reporting time frame would also be considered relevant. For example, participants in drug prevention studies are often asked to self-report their use of drugs over the past week, past month, and past year.

9.4.1.7 Effect Sizes The heart of a meta-analysis is the effect size because it encodes the essential research findings from the studies of interest. Often, the data reported in a study need to be reconfigured to compute the effect size of interest. For example, a study may only report data separately for boys and girls, but you might only be interested in the overall effect across both genders. In such cases you will need to properly aggregate the reported data to be able to compute your effect size. Similarly, a study may use a complex factorial design, of which you are only interested in a single factor. If so, you may need to compute the marginal means if they are not reported. In treatment effectiveness research, some primary authors will remove treatment dropouts from the analysis. However, you may be interested in an intent-to-treat analysis and therefore may need to compute a weighted mean between the treatment completers and treatment dropouts to obtain the mean for those assigned to the treatment group. The coding protocol cannot anticipate all of the potential statistical manipulations that you may need to make. A copy of these computations should be kept for reference, ideally in some computerized form such as a spreadsheet, an R script, or a simple text file. If the effect size was computed using software external to your main analysis software, such as an online effect-size calculator, notes should be kept on which computation method was used.

I have found it helpful to code the data on which the effect size is based in addition to the computed effect size. For example, for the standardized mean, difference effect size this would require recording the means, standard deviations or standard errors, and sample sizes for each condition. Other common data on which this effect size is based include the *t*-test and sample sizes and the exact *p*-value for the *t*-test and sample sizes. If dichotomous outcomes are being coded, the frequencies or proportions of successes or failures for each condition would need to be recorded. This does not exhaust the possibilities but captures the common configurations for the computation of the standardized mean difference (see Lipsey and Wilson 2001). For the remaining instances based on more complex situations, I recommend simply keeping track of the computational method. To assist in data cleanup, it is useful to record the page number of the manuscript where the effect size data can be found. Figure 9.1 shows an example of an effect size coding form for a meta-analysis based on the standardized mean difference. Variations of this form could easily be created for other effect size types, such as the correlation coefficient, odds ratio, or risk ratio.

9.4.1.8 Confidence Ratings The inadequate reporting of information in studies will regularly frustrate synthesists. Often this will simply mean that they must code an item as "cannot tell" or "missing." However, at times information is provided but is inadequate or unclear. For example, a study may report the overall sample size but not the sample sizes for the individual groups. Based on the method of constructing the conditions, it may be reasonable to assume that conditions were of roughly equal size. In a meta-analysis of correctional boot-camps (MacKenzie, Wilson, and Kider 2001), several studies failed to mention whether the sample was all boys, all girls, or a mix. Given the context, that the samples were all boys seemed reasonably certain. A synthesist may wish to track the degree of uncertainty or confidence in

EFFECT SIZE LEVEL CODING FORM

Code this sheet separately for each eligible effect size.

Identifying Information:

1. Study (document) identifier — StudyID |___|___|___|
2. Treatment-Control identifier — TxID |___|___|___|
3. Outcome (dependent variable) identifier — OutID |___|___|___|
4. Effect size identifier — ESID |___|___|___|
5. Coder's initials — ESCoder |___|___|___|
6. Date coded — ESDate |___/___/___|

Effect Size Related Information:

7. Pretest, posttest, or follow-up (1 = pretest; 2 = posttest; 3 = follow-up) — ES_Type |___|
8. Weeks Post-Treatment Measured (code 999 if cannot tell) — ES_Time |___|___|___|
9. Direction of effect (1 = favors treatment; 2 = favors control; 3 = neither) — ESDirect |___|

Effect Size Data—All Effect Sizes:

10. Treatment group sample size — ES_TxN |___|___|___|___|___|
11. Control group sample size — ES_CgN |___|___|___|___|___|

Effect Size Data—Continuous Type Measures:

12. Treatment group mean — ES_TxM |___|___.___|___|
13. Control Group mean — ES_CgM |___|___.___|___|
14. Are the above means adjusted (for example, ANCOVA adjusted)? (1 yes; 0 no) — ES_MAdj |___|
15. Treatment group standard deviation — ES_TxSD |___|___.___|___|
16. Control group standard deviation — ES_CgSD |___|___.___|___|
17. *t*-value — ES_t |___.___|___|___|

Effect Size Data—Dichotomous Measures:

18. Treatment group; number of failures (recidivators) — ES_TxNf |___|___|___|___|___|
19. Control group; number failures (recidivators) — ES_CgNf |___|___|___|___|___|
20. Treatment group; proportion failures — ES_TxPf |___.___|___|
21. Control group; proportion failures — ES_CgPf |___.___|___|
22. Are the above proportions adjusted for pretest variables? (1 = yes; 0 = no) — ES_PAdj |___|
23. Logged odds-ratio — ES_LgOdd |___.___|___|___|
24. Standard error of logged odds-ratio — ES_SELgO |___.___|___|___|
25. Logged odds-ratio adjusted for covariates? (1 = yes; 0 = no) — ES_OAdj |___|

Figure 9.1. Sample Effect-Size Level Coding Form

SOURCE: Author's tabulation.

Effect Size Data—Hand Calculated:

26. Hand calculated *d*-type effect size ES_Hand1 |___.___|___|___|
27. Hand calculated standard error of the *d*-type effect size ES_Hand2 |___.___|___|___|

Effect Size Data Location

28. Page number where effect size data found ES_Pg |___|___|___|___|

Effect Size Confidence

29. Confidence in effect size value ES_Conf |___|
    1. Highly Estimated—have *N* and crude *p*-value only, for example, $p<.10$, or other limited information
    2. Some Estimation—have complex but complete statistics; some uncertainty about precision of effect size or accuracy of information
    3. No Estimation—have conventional statistical information and am confident in accuracy of information

Figure 9.1. (*Continued*)

the information coded for key items, such as the data on which the effect size is based. For example, coders could rate on a 3-point or 5-point scale their confidence in the data on which the effect size is based, as shown in figure 9.1.

9.4.2 Iterative Refinement

An important component of good questionnaire development is pilot testing. Similarly, pilot tests of the coding protocol should be conducted on a sample of studies to identify problem items and lack of fit between the coding categories and the characteristics of the studies. Synthesists should not start simply with the studies at the top of the pile. These are likely to be those that were easy to retrieve or were from the same bibliographic source, such as PsycNet or ERIC. Synthesists want to ensure that the pilot test of the coding protocol is conducted on studies that represent the full range of characteristics of the sample, to the extent that this is possible. Although throwing out data coded as part of the pilot testing is not necessary, synthesists do need to recode any items that were refined or added as a result of the pilot test. The number of studies that need to be coded as part of the pilot phase will depend on the complexity of the research area and the skill and experience of the research team.

9.4.3 Structure of the Data

A complexity of meta-analytic data is its nested or hierarchical nature: multiple effect sizes nested within studies. The coding protocol must allow for this complexity. There are two general approaches to handling the nested nature of the data that can be adapted to any situation: a flat-file and a hierarchical or relational file structure. In the flat-file approach, a synthesist specifies or knows beforehand the nature and extent of the nesting and repeats the effect size variables the necessary number of times: for example, one set of variables for the first effect size of interest and another set for the second. This works well for a limited and highly structured nesting. In my experience, however, the number of effect sizes of interest per study is typically not known beforehand. A hierarchical or relational data structure creates separate data tables for each level of the hierarchy. In its simplest form, this would involve a study-level data table with one row per study and a related effect-size-level data table with one row per effect size. A study identifier links the two data tables. I illustrate each approach.

9.4.3.1 Flat File Approach The flat file approach creates a single data table with one row per study and as many columns as variables or coded items. It is suitable for situations in which the synthesist knows the maximum number of effect sizes per study that will be coded and can place each effect size in a distinct category. For example, suppose a synthesist is interested in meta-analyzing a collection of studies examining a new teaching method. The outcomes of interest are math and verbal achievement as measured by a well-known standardized achievement test. In this case the synthesist could develop a coding protocol that has one set of effect-size items that addresses the math achievement outcome and another set that addresses the verbal achievement outcome. This approach has the advantage that all the data are contained within one data table. Also, with one row per study, only

one effect size per study can be included in any single analysis, by design. The disadvantage is that the flat file approach becomes cumbersome as the number of effect sizes per study increases: wide data tables are difficult to work with. Also, any statistical manipulation of effect sizes, such as the application of the small sample size bias correction to standardized mean difference effect sizes needs to be performed separately for each additional effect size, because each effect size within a study is contained in a different column or variable within the dataset. Furthermore, even in this example with only two effect sizes, the number of columns can become rather large if it accommodates all of the data needed for computing the two effect sizes.

9.4.3.2 Hierarchical or Relational File Approach The hierarchical or relational file approach creates separate data tables for each conceptually distinct level of data. Data that can be coded only a single time for a given study, such as the characteristics of the research methods, sampling procedures, study context, and experimental manipulation, are entered into a single study level data table. Data that change with each effect size are entered into a separate data table with one row per effect size. This data structure can accommodate any number of effect sizes per study.

In my experience, a minimum of three data tables is needed: a study level, a variable (measurement) level, and an effect-size level. In many research domains, a single variable may be measured at multiple time points, such as at baseline, posttest, and follow-up. Rather than code the features of the variable multiple times for each effect size, a synthesist can separate the variable coding items into their own data table and link that information to the effect-size data table with a study identifier and a variable identifier. This simplifies coding and eases data cleanup.

In a more complex meta-analysis, additional data tables may be useful. For example, a separate sample level data table may be useful if you are interested in coding effects for subgroups of the total study sample, such as males and females. Not only does this structure reduce duplication of coding, it also facilitates the identification of all effects that are associated with a distinct subgroup, facilitating appropriate analysis.

This approach is shown in figure 9.2 for a meta-analysis of wilderness challenge programs for juvenile delinquents (Lipsey and Wilson 2001). These studies may each include any number of measures and each may be measured at multiple time points. The coding protocol was divided into three sets of items: those at the study level, that is, do not change for different outcomes; those that describe each variable used in the study; and those that encode each effect size. The effect-size data table includes a study identifier (Study_ID) and a variable identifier (DV_ID) that link each effect size row with the appropriate data in both the study-level and variable-level data tables. This approach also works well for meta-analyses of correlational studies. However, rather than each effect size being associated with a single variable, it will be associated with two. In this way, you can efficiently code all correlations in a correlation matrix by first coding the characteristics of each measured variable and then simply indicate which pair of variables is associated with each correlation.

The hierarchical approach has the advantage of flexibility, simplified coding, and easier data cleanup. The disadvantage is the need to manipulate the data tables prior to analysis. The steps involved are merging the individual data tables into a single flat file, selecting a subset of effect sizes for analysis, and confirming that the selection algorithm yields or results in a single effect size per study. Because the effect size is in a single variable (one column of the data table), it is far easier to create a composite effect size within each study to maintain the one effect size per study restriction. For example, you may wish to perform an analysis based on any achievement effect size within a study, averaging multiple achievement effect sizes if they exist. This can easily be accomplished with a hierarchical data structure.

The principle of the hierarchical approach is to avoid coding the same information more than once. This helps reduce coding effort and coding errors. It also means that when doing data cleanup, corrections do not need to be repeated for each instance of the same bit of information.

9.4.4 Coding Forms and Coding Manual

It is often advantageous to develop both a coding manual and coding forms. The former is a detailed documentation of how to code each item and includes notes about decision rules developed as the synthesis team codes studies. A coding form, however, is designed for efficient recording of coded data. Figure 9.3 provides an example of the contents of a code manual for two items and the corresponding items on the coding form. Note that the coding form is decidedly light on information, providing just enough information for trained coders to know where to record relevant study data.

This approach offers several advantages. First, the synthesists can annotate the coding manual as the team codes studies and that information does not become "trapped" in the coding form of a single study. Second, coding is

STUDY_ID	PUBYEAR	MEANAGE	TX_TYPE
001	1992	15.5	2
002	1988	15.4	1
003	2001	14.5	1

STUDY_ID	DV_ID	CONSTRUCT	SOURCE
001	01	2	2
001	02	6	1
002	01	2	2
003	01	2	2
003	02	3	1
003	03	6	1

STUDY_ID	ES_ID	DV_ID	FU_MONTHS	TX_N	CG_N	ES
001	02	01	0	44	44	–.39
001	02	01	6	42	40	–.11
001	03	02	0	44	44	.10
001	04	02	6	42	39	.09
002	01	01	0	30	30	.34
002	02	01	12	30	26	.40
003	01	01	6	52	52	.12
003	02	02	6	51	50	.21
003	03	03	6	52	49	.33

Figure 9.2. Example of a Hierarchical Data Structure

SOURCE: Author's tabulation.
NOTE: The boxes and arrows show the related data for study 001. The variables are: STUDY_ID = study identifier; PUBYEAR = publication year; MEANAGE = mean age of sample; TX_TYPE = treatment type; DV_ID = dependent variable identifier; CONSTRUCT = construct measured by the dependent variable; SOURCE = source of measure; ES_ID = effect size identifier; FU_MONTHS = months post-intervention; TX_N = treatment sample size; CG_N = control group sample size; ES = effect size.

more efficient because coders need to consult the manual only for cases where they are uncertain as to the correct way to code an item for a given study. Finally, coders can condense the coding forms in a matrix layout, as shown in figure 9.4, for efficient coding of measures and effect size data, or any other data items that may need to be coded multiple times for a single study. Notice that in figure 9.4, four effect sizes from a single study can be coded on a single page, simplifying the transfer of data from tables in reports onto the coding form.

9.5 CODING MECHANICS

At a mechanical level, studies can be coded on paper coding forms or directly into the computer in some fashion. The discussion so far has presumed that paper coding forms are used. I recommend that even if a synthesist does develop a database complete with data entry forms, they start with paper coding forms as the template. Paper coding forms are also easier to share with others who request a copy of your coding protocol. These paper forms should also provide clear information about the structure of the data files, including the field (or variable) names for each data table. These field names can be recovered from the database but are usually not visible on data entry forms.

9.5.1 Paper Forms

Paper coding forms, such as those shown in figure 9.1 and 9.4, have the advantage of being simple to create and to use. There is also no need to worry about computer backups or

Sample Coding Items from Coding Form

____ ____ . ____ ____ 4. Mean age of sample [MEANAGE]

____ 19. Occur in a wilderness setting?
(1 = yes; 0 = no) [WILDNESS]

Sample Coding Items from Coding Manual

4. Mean age of sample. Specify the approximate or exact mean age at the beginning of the intervention. Code the best information available; estimate mean age from grade levels if necessary. If mean age cannot be determined, enter 99.99.

19. Did the program occur in a wilderness setting? Code as yes if the activities took place outdoors, even if the participants were camping in cabins or other buildings. Code as no if the activities took place indoors or used man-made contraptions. (1 = yes; 0 = no)

Figure 9.3. Sample Coding Items

SOURCE: Author's tabulation.
NOTE: Variable names are shown in brackets. (Extracted from example used in Lipsey and Wilson, 2001, of a meta-analysis of challenge programs for juvenile delinquents.)

Effect Size Data				
1. Study ID	|___|___|___|			
2. Outcome ID	|___|___|___|	|___|___|___|	|___|___|___|	|___|___|___|
3. ES ID	|___|___|___|	|___|___|___|	|___|___|___|	|___|___|___|
. . .				
30. Treatment *N*	|___|___|___|	|___|___|___|	|___|___|___|	|___|___|___|
31. Control *N*	|___|___|___|	|___|___|___|	|___|___|___|	|___|___|___|
32. Treatment mean	|___|___.___|___|	|___|___.___|___|	|___|___.___|___|	|___|___.___|___|
33. Control mean	|___|___.___|___|	|___|___.___|___|	|___|___.___|___|	|___|___.___|___|
34. Adjusted?	|___|	|___|	|___|	|___|
35. Treatment SD	|___|___.___|___|	|___|___.___|___|	|___|___.___|___|	|___|___.___|___|
36. Control SD	|___|___.___|___|	|___|___.___|___|	|___|___.___|___|	|___|___.___|___|
37. *t*-value	|___|___.___|___|	|___|___.___|___|	|___|___.___|___|	|___|___.___|___|

Figure 9.4. Example of a Matrix Layout

SOURCE: Author's tabulation.
NOTE: For a subset of items from an effect size coding form allowing up to four effect sizes per form.

database corruption. If the research synthesis involves a small number of studies, such as twenty or fewer, then paper coding forms are likely to be a good choice. However, with paper coding forms, coders must enter the data into a computer and manage all of the completed paper coding forms. Also, the comparison of double-coding for a large synthesis is more tedious when using paper forms.

When creating paper coding forms, a synthesist should always include the variable names that will be used in the statistical software program on the forms. Memory fades quickly. It can be a tedious task reconstructing which variables in a data file corresponded to specific items in a coding protocol. I also recommend that a synthesist place the fields for recording the data either down the left or right margin of the coding forms rather than scattered here and there across the page. This positioning has two advantages. First, it allows easy visual inspection of the forms to ensure that each item has been completed. Second, data entry proceeds more smoothly because all data are in a column along either the left or the right margin.

9.5.2 Coding Directly into a Database Databases provide an alternative to paper coding forms and allow direct coding of studies into a computer, eliminating the need to transfer data from paper forms to computer files. Almost any modern database program can be used, such as FileMaker Pro and Microsoft Access. Using a database for study coding has several advantages. First, a synthesist can specify allowable values for any given field (coded item), helping to enhance data accuracy. Second, databases save data as coders work, helping reduce the likelihood of data loss. Third, the data in most databases can easily be exported to any number of statistical software programs. Fourth, the synthesist can create data entry forms that look similar to paper coding forms, providing all of the same information about how to code an item that might be on a paper coding form. A simple example of a coding form in a database program is shown in figure 9.5. Fifth, with a bit more effort, a synthesist can have the database compute effect sizes, at least for the more common statistical configurations. And sixth, data queries can be constructed to assess data consistency, aiding in the process of data cleanup. The disadvantages of using a database are the time, effort, and technical skills needed to create the database and design the data entry forms. For any reasonably large meta-analysis, the result is generally well worth the effort.

Synthesists may be willing to share databases that they have created. Even though some customization is generally needed, developing a database from an existing template can be extremely useful, saving both time and effort.

9.5.3 Disadvantages of a Spreadsheet Spreadsheets, such as MS Excel or LibreOffice Calc, are extremely useful programs. They are, however, not well suited for systematically coding meta-analytic data. The spreadsheet framework is rows and columns, columns representing different variables and rows representing different studies or effect sizes. This is similar to the typical flat file data table familiar to anyone accustomed to data analysis, and as such seems a natural fit with the task at hand. However, coding directly into a spreadsheet is different from basic data entry, where the data are already neatly organized on a survey form or other data instrument. Even in such a case, data entry using a spreadsheet is typically inefficient relative to creating a simple ASCII data file in a text editor or to using a database program. In coding studies, the order of items on a coding form never corresponds to the order of the information in a report. Thus coders generally "bounce around" the coding form, recording information as they find it. In a spreadsheet, doing so involves moving back and forth across the various columns used for the different coding items, and the number of columns needed typically far exceeds what can be easily displayed on a single computer screen. Furthermore, the column headings are often cryptic given limited column width, providing scant information on how to code an item. For example, it is difficult to display the values associated with a nominal scale in a spreadsheet (such as, 1 = randomized; 2 = quasi-experimental with baseline adjustment; and so forth). When using a spreadsheet, it is also quite easy to inadvertently change cells, overwriting previously coded data.

The disadvantages of the spreadsheet in this context lead me to strongly discourage its use. I recommend that you either use paper coding forms and then enter the data into your favorite statistical or meta-analytic software package (directly or by creating a raw ASCII datafile), or that you create a database with data entry forms. A spreadsheet may be serviceable for a meta-analysis with few studies and a smaller number of coded items, but it is still less than ideal.

9.6 TRAINING OF CODERS

Coding studies is a challenging, time-consuming, and tedious task. Most research syntheses will involve a research team, even if that team consists of only two people. The training of the coders is important to ensure the consistency and accuracy of the information extracted from the studies. Practice coding, regular meetings, coder specialization, and assessment of coder reliability all contribute to coder training and the quality of the final synthesis.

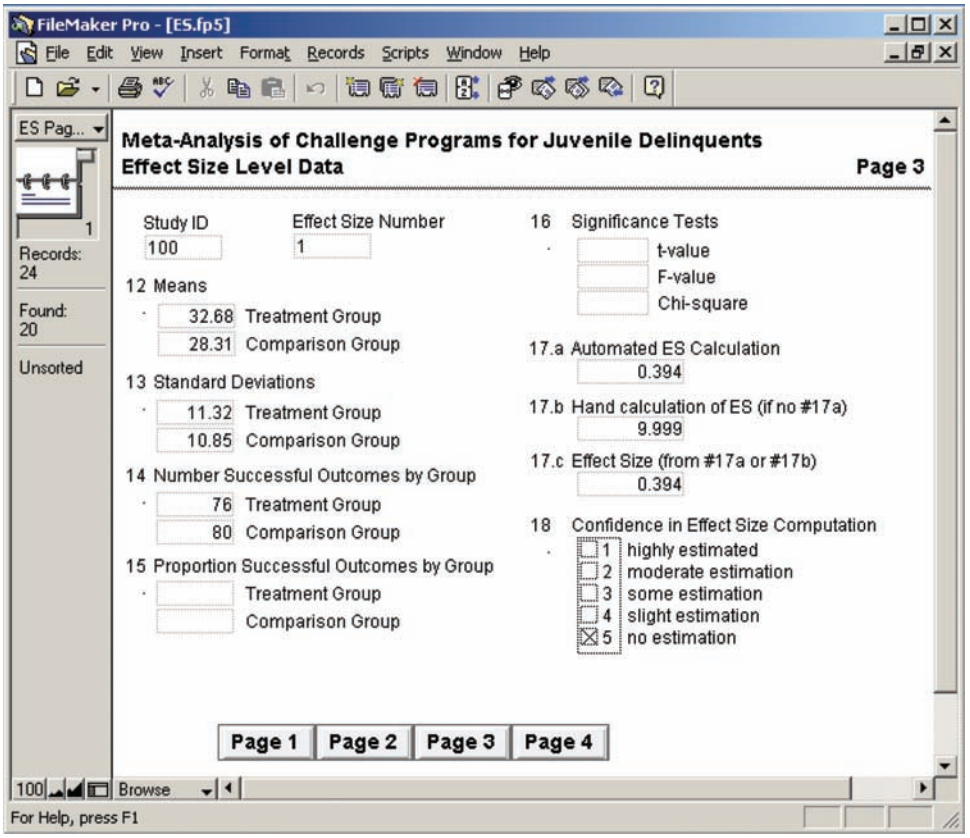

Figure 9.5. Example of a Database Coding Form

SOURCE: Author's tabulation.

9.7 PRACTICE CODING

William Stock recommends a training process that involves eight steps (1994). A central feature of this process is practice coding of a sample of studies, comparing the results among all members of the research team, modifying coding items if necessary, and repeating this process until good consistency across coders is achieved. Stock's eight steps are as follows:

- The principal investigator provides an overview of the synthesis.
- Each item on a form and its description in the code book is read and discussed.
- The process for using the forms is described. This is the method chosen to organize forms so that the transition from coding to data entry and data management is facilitated, and should take into account studies that include reports for subsamples or multiple occasions or measures.
- A sample of five to ten studies is chosen to test the forms.
- A study is coded by everyone, each coder recording how long it takes to code each item. These time data are used to estimate how long it will take to code individual items, entire study reports, and the complete synthesis.
- Coded forms are compared and discrepancies are identified and resolved.
- The forms and code book are revised as necessary.

- Another study is coded and so on. Steps four through eight are repeated until apparent consensus is achieved. (134–35)

This process should be adapted to your specific needs. A small meta-analysis of five to fifteen studies may abbreviate this process by practice coding only two or three studies and then simply double-coding all studies to ensure coding reliability.

9.7.1 Regular Meetings

During the coding phase of a meta-analysis, coders should meet regularly to discuss coding difficulties. Discussing the difficult coding decisions provides an opportunity to develop a normative understanding of the application of coding rules among all coders. Synthesists should also annotate the coding manual during these meetings to ensure that decisions made about how to handle various study ambiguities are not forgotten and that similar studies encountered later in the coding process can be handled in a similar way.

9.7.2 Using Specialized Coders

A common mode of operation is for each coder to read through an entire study and code all aspects required by the synthesis. This approach presumes that all members of the research team have the same skills and can accurately complete the entire protocol. However, colleagues and research assistants working on the meta-analysis may not have all of these skills. For example, a colleague with exceptional substantive and theoretical knowledge of the relevant research domain may not be facile with the computation of effect sizes, particularly when it requires heavy manipulation of the statistical data presented. Similarly, a research assistant with strong quantitative abilities may not have adequate background training and experience to code substantive issues that require complex judgments. In these situations, it may be advantageous for coders to specialize, each completing a different section of the coding protocol. Coder specialization has the advantage of exploiting strengths to improving coding quality. It can also help address the problem of coder drift (that is, a change in the interpretation of coding items over time) because each coder will proceed more quickly through the studies, focusing on a smaller set of items. Finally, it is easier to keep coders who only code effect size data blind to other potentially important aspects of the study, such as the type of intervention, study authorship, and institutional affiliation.

9.7.3 Assessing Coding Reliability

The reliability of coding is critical to the quality of the research synthesis. Both intra- and interrater reliability are important. Intrarater reliability is the consistency of a single coder from occasion to occasion. A common problem in a research synthesis is coder drift: as more studies are coded, items are interpreted differently, resulting in coding inconsistencies. This is most likely to affect items that require judgment on the part of coders, such as an assessment of the similarity of groups at baseline. Purely factual items are less likely to suffer from coder drift. Intrarater reliability can be assessed by having coders recode studies they coded at some earlier point. Items that show low intrarater agreement might need to be recoded to improve accuracy.

Interrater reliability is the consistency of coding between different coders. It is assessed by having at least two different members of the synthesis team code a sample of the studies. Studies should be coded in a different order by the two coders. Once a sample of studies has been double coded, the two sets of results are compared. Reliability is often computed as the percentage of agreement across the sample of studies. A weakness of this agreement is that it is affected by the baserate, resulting in artificially high for items shared by most studies. Other reliability statistics, such as kappa or the intra class correlation, can also be computed and are likely to be more informative (see chapter 10).

An important decision is whether to double code all studies or only a random sample of studies. Historically, the practice within the field of research synthesis has been to double-code all studies for a small meta-analysis but to only double code a random sample for a large meta-analysis. However, systematic reviews conducted for either Cochrane or the Campbell Collaboration are required to double code all information related to the outcomes (that is, everything related to the effect sizes) and are strongly recommended to double-code all other study characteristics (Chandler et al. 2013). A study by Nina Buscemi and her colleagues finds that single data extraction produced more errors than double data extraction (2006). In this study, single data extraction still had two coders; the second coder, though, simply verified the coding of the first coder; that is, the verifier compared the coded information to information in the written manuscript.

Double data extraction was fully independent separate coding by two individuals. Best practice would thus dictate double coding for all studies, particularly for all coded items relevant to the research synthesis and use a consensus process to resolve all coding differences.

A complication of determining the reliability of coded items stems from the dependent nature of some questions. How the synthesist answers one question often restricts the options for another. William Yeaton and Paul Wortman recommend that reliability estimates take into account this contingent nature of some items (1993). Ignoring this hierarchy of items may result in underestimating the reliability of these contingent items. Therefore, they recommend that for contingent items, reliability be assessed only when agreement has been reached on the higher order or more general item. Although this idea has merit, I am unaware of any research synthesis that has adopted it.

Items with low coder agreement are problematic because they reflect the difficulty coders had in assessing the item given the information available in the written report. The double coding and resolution of any coding differences enhance the reliability of the item, but the synthesist should consider whether the item needs to be rewritten to address coder confusion or to simplify coder judgments, possibly breaking a complex judgment into simpler judgments. This may require a another pass through the eligible studies to code any modified items. At a minimum, items with problematic coder agreement should be acknowledged when reporting the results of the meta-analysis.

9.7.4 Masking of Coders

In primary research it is often advisable to have participants unaware of the conditions of assignment. For example, in a randomized clinical trial of a new drug, it is common for participants to be kept uninformed as to whether they are receiving the new drug or a placebo. Similarly, keeping raters unaware of conditions, such as medical personnel evaluating a patient's status, is also often advisable. A study may involve both of these methods (a double-blind study) to protect against both sources of potential bias.

There is an analog in meta-analysis. Might not coders be influenced in making quality judgments or other coding decisions about a study by knowledge of who conducted the study, the author's institutional affiliation, or the study findings? Well-known research teams might benefit from a positive bias and their counterparts may be challenged by a negative bias. Two decades ago, Alejandro Jadad and his colleagues examined the effect of masking on the ratings of study quality and found that masked assessment was more consistent than open assessment and resulted in lower quality ratings, suggesting that some coders were affected by knowledge of the authors (1996). In another study, the University of Pennsylvania Meta-Analysis Blinding Study Group randomly assigned pairs of coders to one of two conditions: unaware of author and author affiliation information or aware of such information (Berlin 1997). Of interest was the effect of masking on effect-size data. The study found no statistically or clinically meaningful difference in computed effect sizes. Thus, masked assessment may be beneficial for assessment of study quality but does not appear to matter for the coding of effect-size information. The latter is likely to generalize to the coding of other factual information.

At present, empirical evidence to provide clear guidance on the value of masking study authorship and affiliation is scant. Unless the coders are already familiar with key studies in research domain of interest, then masking may be beneficial and fairly easy to implement. Clearly, more work is needed in this area.

9.8 COMMON MISTAKES

Novice research synthesists make several mistakes in designing coding protocols. These errors often stem from a failure to fully plan the analysis of the meta-analytic data. The most common and serious error is a failure to recognize the implications of the hierarchical nature of meta-analytic data. A large data file with a single row per effect size without codes to allow for the selection of independent subsets of effect sizes for analysis leads to an unmanageable data structure. If synthesists are going to code multiple effect sizes per study, they need to think through how they are going to categorize them such that only one effect size (or a composite of multiple effect sizes) is selected for any given analysis. If this is not assured, the dependency in the results yields biased meta-analytic results, such as standard errors that are too small (for a discussion of this issue and a method for analyzing statistically dependent effect sizes, see chapter 13).

Another common error in coding is to underestimate the time, effort, and difficulty of coding studies for meta-analysis. Although it may be possible to code some studies in thirty minutes to an hour, I have spent as long

eight hours coding a single study. Dissertations and technical reports are often more time consuming than journal articles given their length and more complete reporting of analyses. Resolving differences between two (or more) versions of a study also adds to the coding burden.

Finally, as I mentioned earlier, individuals new to meta-analysis overestimate the level of detail that will be available in the individual studies and often try to code more items than are useful or ultimately used. During the development phase of the coding protocol, items for which information is rarely adequate should be eliminated unless it is of enough theoretical interest to document the inadequacy (for a different take on the issue, see chapter 8). Also, the synthesist should keep in mind that their ability to use items for moderator analyses will be limited by the size of the meta-analysis. Complex moderator analyses are simply not possible for a small meta-analysis, reducing the utility of a large number of variables coded for that purpose.

9.9 CONCLUSION

The goal of this chapter is to provide guidance on how to achieve a coding protocol that is replicable and transparent and that captures essential information about the studies being synthesized. No matter how fancy the statistical analyses of effect sizes are or impressive the forest plots, the scientific credibility of a research synthesis rests of the ability of others to scrutinize how the data on which the results of the synthesis were generated. Without this transparency, a meta-analysis cannot be replicated or adequately scrutinized. Science progresses not from an authoritarian proclamation of what is known on a topic but rather from a thoughtful, systematic, and empirical taking of stock of the evidence that is open to debate, discussion, and reanalysis. A detailed coding protocol is a key component of this process.

9.10 REFERENCES

Berlin, Jesse A. 1997. "Does Blinding Readers Affect the Results of Meta-Analysis?" *Lancet* 350(9072): 185–86.

Brown, Sharon A., Sandra L. Upchurch, and Gayle J. Acton. 2003. "A Framework for Developing a Coding Scheme for Meta-Analysis." *Western Journal of Nursing Research* 25(2): 205–22.

Buscemi, Nina, Lisa Hartling, Ben Vandermeer, Lisa Tjosvold, and Terry P. Klassen 2006. "Single Data Extraction Generated More Errors than Double Data Extraction in Systematic Reviews." *Journal of Clinical Epidemiology* 59(7): 697–703.

Chandler, Jackie, Rachel Churchill, Julian Higgins, Toby Lasserson, and David Tovey. 2013. "Methodological Standards for the Conduct of New Cochrane Intervention Reviews." Version 2.3. Accessed December 4, 2018. http://www.editorial-unit.cochrane.org/mecir.

Cheung, Alan C. K., and Robert E. Slavin. 2015. *How Methodological Features Affect Effect Sizes in Education. Best Evidence Encyclopedia (BEE): Empowering Educators with Evidence on Proven Programs.* Baltimore, Md.: Johns Hopkins University Press.

Gorman, Dennis M. 2016. "Can We Trust Positive Findings of Intervention Research? The Role of Conflict of Interest." *Prevention Science* 19(3): 295–305.

Heinsman, Donna T., and William B. Shadish 1996. "Assignment Methods in Experimentation: When Do Nonrandomized Experiments Approximate Answers from Randomized Experiments?" *Psychological Methods* 1(2): 154–69.

Higgins Julian P. T., and Sally Green, eds. 2011. "Cochrane Handbook for Systematic Reviews of Interventions, Version 5.1.0." The Cochrane Collaboration. Accessed December 4, 2018. http://www.handbook.cochrane.org.

Jadad, Alejandro R., R. Andrew Moore, Dawn Carroll, Crispin Jenkinson, D. John M. Reynolds, David J. Gavaghan, and Henry J. McQuay. 1996. "Assessing the Quality of Reports of Randomized Clinical Trials: Is Blinding Necessary?" *Controlled Clinical Trials* 17(1): 1–12.

Jüni, Peter, Franziska Holenstein, Jonathan Sterne, Christopher Bartlett, and Matthias Egger. 2002. "Direction and Impact of Language Bias in Meta-Analyses of Controlled trials: Empirical Study." *International Journal of Epidemiology* 31(1): 115–23.

Lipsey, Mark W., and David B. Wilson. 1993. "The Efficacy of Psychological, Educational, and Behavioral Treatment: Confirmation from Meta-Analysis." *American Psychologist* 48(12): 1181–209.

———. 2001. *Practical Meta-Analysis.* Thousand Oaks, Calif.: Sage Publications.

Lipsey, Mark W., David B. Wilson, Mark A. Cohen, and James H. Derzon. 1997. "Is There a Causal Relationship Between Alcohol Use and Violence? A Synthesis of Evidence." In *Recent Developments in Alcoholism*, vol. 13: *Alcohol and Violence*, edited by Marc Galanter. New York: Plenum.

MacKenzie, Doris L., David B. Wilson, and Susan Kider. 2001. "Effects of Correctional Boot Camps on Offending." *Annals of the American Academy of Political and Social Science* 578:126–43.

Moher, David, Kenneth F. Schulz, Douglas G. Altman, and Consort Group. 2001. "The CONSORT Statement: Revised Recommendations for Improving the Quality of Reports of

Parallel-Group Randomised Trials." *The Lancet* 357(9263): 1191–94.

Morrison, Andra, Julie Polisena, Don Husereau, Kristen Moulton, Michelle Clark, Michelle Fiander, Monika Mierzwinski-Urban, Tammy Clifford, Brian Hutton, and Danielle Rabb. 2012. "The Effect of English-Language Restriction on Systematic Review-Based Meta-Analyses: A Systematic Review of Empirical Studies." *International Journal of Technology Assessment in Health Care* 28(2): 138–44.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349(6251): p.aac4716.

Shadish, William R., and Kevin Ragsdale. 1996. "Random versus Nonrandom Assignment in Psychotherapy Experiments: Do You Get the Same Answer?" *Journal of Consulting and Clinical Psychology* 64(6): 1290–305.

Stock, William A. 1994. "Systematic Coding for Research Synthesis." In *The Handbook of Research Synthesis*, edited by Harris Cooper and Larry V. Hedges. New York: Russell Sage Foundation.

Weisburd, David, Cynthia Lum, and Anthony Petrosino. 2001. "Does Research Design Affect Study Outcomes in Criminal Justice?" *Annals of the American Academy of Social and Political Sciences* 578: 50–70.

Weisz, John R., Bahr Weiss, and Geri R. Donenberg. 1992. "The Lab Versus the Clinic: Effects of Child and Adolescent Psychotherapy." *American Psychologist* 47(12): 1578–85.

Wilson, David B., Denise C. Gottfredson, and Stacy S. Najaka. 2001. "School-Based Prevention of Problem Behaviors: A Meta-Analysis." *Journal of Quantitative Criminology* 17(3): 247–72.

Wilson, David B., and Mark W. Lipsey. 2001. "The Role of Method in Treatment Effectiveness Research: Evidence from Meta-Analysis." *Psychological Methods* 6(4): 413–29.

Yeaton, William H., and Paul M. Wortman. 1993. "On the Reliability of Meta-Analytic Reviews: The Role of Intercoder Agreement." *Evaluation Review* 17(3): 292–309.

10

EVALUATING CODING DECISIONS

JACK L. VEVEA
University of California, Merced

NICOLE A. M. ZELINSKY
University of California, Merced

ROBERT G. ORWIN
Westat

CONTENTS

10.1 INTRODUCTION

Coding is a critical part of research synthesis. It is an attempt to reduce a complex, messy, context-laden, and quantification-resistant reality to a matrix of numbers. Thus it will always remain a challenge to fit the numerical scheme to the reality, and the fit will never be perfect. Systematic strategies for evaluating coding decisions enable the synthesist to control for much of the error inherent in the process. When used in conjunction with other strategies, they can help reduce error as well. This chapter discusses strategies to reduce error and to control for error and suggests further research to advance the theory and practice of this particular aspect of the synthesis process. To set the context, however, it is first useful to describe the sources of error in synthesis coding decisions.

10.1.1 Standards for Reporting and Coding Evaluations

Various agencies that oversee or make recommendations for the reporting of meta-analyses have added standards that encourage or require stringent reporting of information related to coding. For example, the American Psychological Association (2010) now includes in its publication manual an appendix on meta-analysis reporting standards (MARS) that requires rigorous specification of coding categories, inclusion of information about the number of

coders and their qualifications, coding reliability and agreement, and how discrepancies in coding are resolved (Journal Article Reporting Standards Working Group 2007). A document detailing Preferred Reporting Items for Systematic Reviews and Meta-Analysis (the PRISMA statement), endorsed by numerous medical associations and journals, similarly encourages reporting on the number of people involved in data extraction, and whether consensus, coding training, or formal reliability assessments were used (Moher et al. 2009). Policies for reviews sanctioned by the Campbell Collaboration encourage at least two independent coders and comparison of results (Steering Group of the Campbell Collaboration 2015). The Cochrane Collaboration includes similar recommendations in its handbook (Higgins and Green 2011). It is clear, then, that evaluation of coding decisions has become a standard element in responsible reporting of meta-analyses.

10.2 SOURCES OF ERROR IN CODING DECISIONS

10.2.1 Deficient Reporting in Primary Studies

Reporting deficiencies in original studies present an obvious problem for the synthesist, to whom the research report is the sole documentation of what was done and what was found. Reporting quality of primary research studies has variously been called "shocking" (Light and Pillemer 1984), "deficient" (Orwin and Cordray 1985), and "appalling" (Oliver 1987).[1] Inaccessible reporting can force the abandonment of a synthesis.[2] Moreover, reporting of inaccurate information can result in a synthesis that distorts the truth.

Virtually all write-ups will report some information poorly, but some will be so vague as to obscure what took place entirely (Oliver 1987). The absence of clear or universally accepted norms undoubtedly contributes to the variation, but other factors do as well: different emphases in training, scarcity of journal space, statistical mistakes, and poor writing. The consequences are differences in the completeness, accuracy, and clarity with which empirical research is reported.

Treatment regimens and subject characteristics cannot be accurately transcribed by the coder when inadequately reported by the original author. Similarly, methodological features cannot be coded with certainty when research methods are poorly described. The immediate consequence of coder uncertainty is coder error. One way to address such shortcomings is to agree on conventions for imputing guesses at the values of poorly reported data. For example, in one analysis, when an investigator was remiss in reporting the length of time the therapist had been practicing, the guessing convention for therapist experience assumed five years (Smith, Glass, and Miller 1980). Such a device helps standardize decisions under uncertainty and therefore increases intercoder agreement. It is unknown whether it reduces coder error, however, because there is no external way to validate the accuracy of the convention. Furthermore, a guessing convention carries the possibility of bias in addition to error. Unlike pure observational error, which presumably distributes itself around the true value across coders, the convention-generated errors may not balance out, but consistently over- or underestimate the true value. This would happen, for example, if in reality the average PhD therapist had been practicing eight years, rather than five years as in the guessing convention. Guessing conventions artificially deflate true variance in coded variables, thereby diminishing the sensitivity of the analysis to detect relationships with other variables. Moreover, any systematic over- and underestimation of true values by guessing conventions exacerbates matters. Specifically, it creates a validity problem (for example, therapist experience as coded would not be a valid indicator of therapist experience). The use of a guessing convention based on mean imputation rather than a convenient choice presents similar problems (Enders 2010). For that reason, imputation following guessing conventions should be considered a last resort. Modern multiple imputation approaches using maximum likelihood or Bayesian methods in this context would be a development worth considering (Schafer and Graham 2002).

An alternative to guessing conventions that often provides a better solution is simply to code the information as undetermined. When the variable is categorical, this has the effect of adding one more category to the possible levels of the moderator (for example, sex of sample predominantly female, predominantly male, mixed, or unknown). In the somewhat rarer case when the problematic variable is continuous, the situation may be handled by dummy coding the availability of the information (0 = not available, 1 = available) and adding to the explanatory model both the dummy variable and the product of the dummy and the continuously coded variable. This has the effect of allowing the estimation of the variable's impact on effect size when good information about the variable is available. Such practice is in keeping with the principle of preference for low-inference coding (see chapter 9).

Through their influence on coding accuracy, deficiencies in reporting quality degrade the integrity of later analyses in predictable ways. Errors of observation in effect sizes destabilize parameter estimates and decrease statistical power, whereas errors of observation in independent variables cause parameter estimates to be biased (Kerlinger and Pedhazur 1973). The severity of the negative impact from reporting problems depends on the nature of the variable being coded. Errors in effect sizes clearly have the potential for severe impact. The impact of errors in independent variables depends on the centrality of the variable in the synthesist's interests. For example, errors in type of experimental manipulation or severity of disease are likely more consequential than problems with incidental codes such as year of publication or institution of the first author.

10.2.2 Ambiguities in the Judgment Process

Coding difficulty reflects the complexity of the item to be coded, not simply the clarity with which it is reported. There is some relationship between reporting quality and the need for judgment calls, in that deficient reporting of study characteristics increases the need for judgment in coding. Many other variables, however, intrinsically require judgment regardless of reporting quality. In fact, more judgment can sometimes be necessary when reporting is thorough, because the coder has more information to weigh.

Numerous variables pose judgment problems for the coder. Consider, for example, treatment integrity (that is, the extent to which the delivered treatment measures up to the intended treatment). In the psychotherapy literature that Mary Smith, Gene Glass, and Thomas Miller synthesized in 1980, attrition from treatment (but not from measurement), spotty attendance, failure to meet advertised theoretical requirements, and assorted other implementation problems all potentially degraded treatment integrity. The authors did not attempt to code treatment integrity per se, though they did code their degree of confidence that the labels placed on therapies by authors described what actually transpired. In a reanalysis, Robert Orwin and David Cordray (1985) attempted to code integrity more globally, but without much success. The following examples (representing actual cases) point up some of the difficulties:

Case 1: The efficacy of ego therapy was tested against a placebo treatment and no-treatment controls. Several participants did not attend every session; only those attending four or more sessions out of seven were posttested.

Case 2: The comparative efficacy of therapist-administered desensitization and self-administered desensitization was tested (relative to various control groups) for reducing public speaking anxiety. The therapists were advanced graduate students in clinical psychology who had been trained in the use of desensitization but were inexperienced in its application.

Case 3: The comparative effect of implosive therapy, eclectic verbal therapy, and bibliotherapy was tested for reducing fear of snakes. All eclectic verbal therapy was performed by Gene Glass, who has published several articles on implosive therapy.

Case 1 exemplifies by far the most common treatment integrity problem Orwin and Cordray encountered: the potential dilution of the treatment regimen by nonattendance (1985). Here the coder must judge whether participants with absentee rates up to 43 percent can be said to have received the intended treatment. If not, the coder needs to determine by how much was it degraded, and how this degradation affects the estimated effect size (a question made still more difficult by the authors' failure to report the number of participants with nonattendance problems). That approach addresses the question "How effective was the therapy for people who completed treatment?" A better question might focus on how effective the treatment is for those who begin it (intention to treat). If the interest is in outcomes for those participants who intended to complete the treatment, estimation of the amount of degradation is a secondary goal that may be appropriate for sensitivity analysis, but not for addressing the primary question about intent to treat. In case 2, the treatment as advertised is potentially degraded by the use of inexperienced treatment providers. The coder must judge whether the lack of practice made a difference and, if so, by how much. In case 3, the treatment provider's motivation to maintain the integrity of the treatment comes into question. The coder must judge whether the apparent conflict of interest degraded the treatment and, if so, by how much (such as uninspired compliance or outright sabotage).

Theoretically, the need for judgment calls in coding such complex constructs could be eliminated by a coding algorithm that considered every possible contingency and provided the coder with explicit instructions in the event of each one, singly and in every combination. The

Smith, Glass, and Miller algorithm for internal validity suggests an attempt at this (1980, 63–64). The components of this decision rule represent generally accepted internal validity concerns. Yet in trying to apply it, Orwin and Cordray frequently find that it fails to accommodate several important contingencies and thus puts their own sense of the study's internal validity in contradiction to the score yielded by the algorithm (see Orwin 1985). But the fault is not with the failure of the algorithm to include and instruct on all contingencies, for in practice that would not be possible. With a construct as complex as treatment integrity or internal validity, no amount of preliminary work on the coding algorithm will eliminate the need for judgment calls. Indeed, the contingency instructions are judgment calls themselves, so at best the point of judgment has only been moved, not eliminated.[3] Nevertheless, when it is practical to modify the coding protocol so as to accommodate exceptional contingencies, the practice of striving for low-inference definitions of complex constructs may be advantageous, if only as an aid to coding reliability.

Other examples abound. David Terpstra reports perfect coding reliability in his synthesis of organization development research (1981). R. J. Bullock and Daniel Svyantek's replication reports numerous problems with both reliability and validity (1985). Specifically, they report that problems occurred in the coding of the dependent variable, where coding required a great deal of subjectivity. Similarly, Joanmarie McGuire and her colleagues were unable to achieve adequate intercoder agreement on methodological quality despite having methodologically sophisticated coders, written coding instructions, and clearly reported study methods (1985; see also chapter 7). As noted previously, coding difficulty reflects the complexity of the item to be coded, not just the clarity with which it is reported. One solution to this problem that has gained favor is using low-inference criteria to assess such issues as study quality or implementation fidelity.[4] To the degree that constructs such as study quality of implementation fidelity can be reduced to unambiguous, directly observable criteria, both the validity and the reliability of these measures may be enhanced (see, for example, Valentine and Cooper 2008).

Inherent ambiguities affect more fundamental coding decisions than what values to code, such as what effect sizes to include. Bert Green and Judith Hall observe that synthesists are divided as to whether to use multiple outcomes per group comparison (1984). As others show, the so-called conceptual redundancy rule—the process used in the original Smith, Glass, and Miller psychotherapy synthesis for determining which of multiple effect sizes within a given study should be counted in determining overall effect size—can be interpreted quite differently by different coders (Orwin and Cordray 1985; Matt 1989; Smith, Glass, and Miller 1980).[5] In recoding a twenty-five-study sample from the original Smith, Glass, and Miller psychotherapy data, four independent coders extracted 81, 159, 172, and 165 effect sizes, respectively (Smith, Glass, and Miller 1980; Orwin and Cordray 1985; Matt 1989). The corresponding average effect sizes were $d = 0.90$, 0.47, 0.68, and 0.49. Thus, although the coders were attempting to follow the same decision rule for extracting effect sizes, both the number of effect sizes and the resulting findings varied by a factor of two. Using the same set of printed guidelines on conceptual redundancy, the coders still disagreed substantially. Not surprisingly, coder disagreement on which effect sizes to include led to more discrepant results than coder disagreement on effect size computations once the set of effect sizes had been decided on (Matt 1989). Again, more clarity in the decision rule might have better guided the judgment process, but it is unlikely that it could have eliminated the need for judgment.

Additional inclusion rules can compensate for the lack of agreement stemming from the conceptual redundancy rule. In this case, for instance, the synthesis can be restricted to that subset of nonredundant effect sizes on which all or at least most coders agree (Orwin and Cordray 1985). This has the effect of eliminating potentially questionable effect sizes. In still another variation, David Shapiro and Diana Shapiro retained all measures except those permitting only a "relatively imprecise" effect-size estimate (1982, 589). These were discarded if, in the coders' view, the study provided more complete data on enough other measures. Georg Matt presents additional rules for selecting effect sizes (1989; see also chapter 8).

10.2.3 Coder Bias

An additional error source is coder bias (such as for or against a given therapy). A coder with an agenda is not a good coder, especially for items that require an inference. The ideal coder is totally unbiased and expert in the content area, but such a coder is difficult to find. Some would argue that by definition, it is impossible. Expertise carries the baggage of opinions, and keeping those opinions out of coding decisions—many of which are by definition

judgmental—is difficult. Ambiguities in the judgment process and coder bias are related in that ambiguity creates a hospitable environment for bias to creep in unnoticed.

Sometimes bias is more blatant. In a synthesis of the effects of school desegregation on African American achievement, a panel of six experts convened by the National Institute of Education independently analyzed the same set of studies, obtaining different results and reaching different conclusions (Wortman and Bryant 1985). Panelists excluded studies, substituted alternative control groups, and sought missing information from authors in accordance with their prior beliefs on the effectiveness of desegregation. Even after discussion, the panel "disbanded with their initial views intact" (315).

One approach to reducing bias is to keep coders selectively blind to information that triggers bias. Thomas Chalmers and his colleagues had two coders independently code papers in random order, with the methods sections separated from the results (1987).[6] In addition, papers were photocopied in such a way that the coders could not determine their origins. Harold Sacks and his colleagues suggest that this is an ideal way to control for this type of bias, but note that it is rarely done (1987). In the eighty-six meta-analyses of randomized clinical trials analyzed, none were successfully blinded (three showed evidence of attempts). The rationale for such masking is exactly the same as in primary studies: to reduce experimenter expectancy effects and related artifacts. It is consistent with the central theme of this book, that research synthesis is a scientific endeavor subject to the same rigorous standards as primary research.

In practice, such masking procedures are difficult to implement, and hence are rarely followed. Two other approaches can help to minimize the impact of coder bias. First, to the degree possible, it is best to identify how issues such as implementation fidelity and study quality will be defined in advance of data collection; this helps prevent creating definitions in such a way that particular individual studies supporting a particular viewpoint will be favored. Second, there is growing agreement that a preference for low-inference coding is helpful with coder bias. To the degree that issues such as study quality can be objectified by criteria like experimenter masking, psychometric properties of measures, impact rating of journal, Carnegie status of primary authors' institutions, and so on, subjectivity is abated, leaving less opportunity for bias (for more discussion of low-inference coding, see chapters 7 and 9).

10.2.4 Coder Mistakes

Of course, coders can be unbiased in the sense of holding no prior views about the likely outcomes of the research being coded, and still make systematic mistakes—systematic in the statistical sense of nonrandom error. The problem is by no means unique to synthesis coding. In their analysis of errors in the extraction of epidemiological data from patient records, Ralph Horwitz and Eunice Yu find that most coding errors occurred because the data extractor simply failed to find information that was present in the medical record (1984). Additional errors were made when information was correctly extracted but the coding criteria were incorrectly applied.

The synthesis coding process is also subject to the same range of simple coder mistakes as any other coding process, including slips of the pencil and keyboard errors. It is particularly vulnerable to the effects of boredom, fatigue, and so on. In a synthesis of any size, many hours are required, often over a period of months, to code a set of studies.

10.3 STRATEGIES TO REDUCE ERROR

Here we discuss nine strategies that potentially reduce error: contacting original investigators, consulting external literature, training coders, pilot testing the coding protocol, revising the coding protocol, possessing substantive expertise, improving primary reporting, using averaged ratings, and seeking coder consensus. Although reducing coding error is distinct from evaluating coding decisions, it is the higher purpose that evaluation of coding decisions serves, and hence merits discussion in this context.

10.3.1 Contacting Original Investigators

An apparent solution to the problem of missing or unclear information is to contact the original investigators in the hope of retrieving or clarifying it. This becomes labor intensive when the number of studies is large, so it is prudent to consider the odds of successful retrieval. The investigators have to be alive; they need to be located; they have to have collected the information in the first place; they need to have kept it; and they have to be willing and able to provide it. Janet Hyde's 1981 synthesis of cognitive gender differences is informative here. Rather than trying to estimate effect sizes from incomplete information, she wrote to the authors. Of the fifty-three studies in her database, eighteen lacked the necessary means and

standard deviations. Although all eighteen authors were located, only seven responded, and only two were able to supply the information. Furthermore, the successful contact rate may have been atypically high because the topic of cognitive gender differences at that time was relatively young; studies that predated the synthesis by more than fifteen years were rare. Of course, Hyde's final success rate could have been still worse had more esoteric information than means and standard deviations been sought. As Richard Light and David Pillemer note, the chance of success probably depends quite idiosyncratically on the field, the investigators, and other factors such as the dates of the studies (1984).[7]

10.3.2 Consulting External Literature

A subset of the information of interest to the synthesist is theoretically obtainable from other published sources if omitted or obscured in the reports. For one variable, experimenter affiliation (training), Smith, Glass, and Miller exercised this option in their original 1980 psychotherapy synthesis, and others have since then. When the experimenter's affiliation was not evident from the report, the American Psychological Association directory was consulted. In their reanalysis of the same data, Orwin and Cordray attempted to get the reliability of the outcome measure, when not extractable from the report, from the *Mental Measurements Yearbook* (Orwin and Cordray 1985; Buros 1978). The strategy was unsuccessful for a number of reasons. Many measures were not included in the yearbook (for example, less-established personality inventories, experimenter-developed convenience scales); when measures were included, a discussion of reliability was frequently omitted; when measures were reviewed and a discussion of reliability included, the range of estimates was sometimes too wide to be useful; and when measures were included, reliability discussed, and an interpretable range of values provided, they were not always generalizable to the population sampled for the psychotherapy study. Contacting test developers directly would provide more information than consulting the *Mental Measurements Yearbook*, but reliabilities of experimenter-developed convenience scales are not obtainable this way. Nor could the problem of generalizing to different populations be resolved.

The use of external sources may be more successful with other variables. For example, detailed information needed for effect-size calculations may be available in a dissertation, but not in the final, shorter publication of the research. A technical report may contain more detail than the published paper. Manuals for large data sets such as the National Assessment of Educational Progress or the Early Childhood Longitudinal Study are likely to present details not reported in a paper. However, the proportion of variables that are potentially retrievable via those strategies will typically be small. For example, of the more than fifty variables coded in the Smith, Glass, and Miller study, experimenter allegiance appeared to be the only variable other than experimenter affiliation that might be deducible through an external published source.

10.3.3 Training Coders

Given the importance and complexity of the coding task, the need for solid coder training as an error-reduction strategy is self-evident (for more, see chapter 9). The training should include a phase in which coders are made familiar with the project and the coding protocol. Ideally, all coders who will be involved in the project should independently code a subset of studies selected to exemplify particularly challenging coding problems. The process of jointly examining the results sensitizes the coders to the possibility of coder variability and provides a vehicle for training in how to resolve the issues that have been identified as most likely to present difficulties. It may sometimes be practical to merge this training phase with the process of developing and pilot testing the coding protocol. Moreover, if the protocol is revised, the need for appropriate retraining is evident.

10.3.4 Pilot Testing the Coding Protocol

Piloting the coding protocol for a synthesis is no less essential than piloting any treatment or measurement protocol in primary research. First, it supplies coders with direct experience in applying the conventions and decision rules of the process before coding the main sample, which is necessary to minimize learning effects. Second, it assesses whether a coder's basic interpretations of conventions and decision rules are consistent with the synthesist's intent and with the other coders' intent. This is necessary to preclude underestimating attainable levels of interrater reliabilities. Third, it can identify inadequacies in the protocol, such as the need for additional categories for particular variables or additional variables to adequately map the studies.

Bullock and Svyantek took the concept of pilot testing a step further in their reanalysis of a prior synthesis of the

organization development literature (1985). After dividing the sixteen years of the study period in half, each author independently coded the first half so as to develop reliability estimates as well as resolve preliminary coding problems. After examining their work, along with each instance of disagreement, the authors attempted to improve interrater reliability by developing more explicit decision rules for the same coding scheme. The more explicit rules were then applied to the studies from the second half of the time period. (The first half was also recoded in accordance with the revised protocol.) Agreement was higher in the second half on six of seven variables tested, sometimes remarkably so (for example, agreement on sample size increased from 70 to 94 percent). It is not clear why the authors halved the sample by period rather than randomly, given that using period introduced a potential confound into the explanation of improvement. Improved quality of reporting in studies from the second half of the time period could account for much of the improvement in interrater reliability.[8] The distinction between what these authors did from what is typically done is that the preliminary codings were performed on a large enough subset of studies to yield reliable quantitative estimates of agreement. This enabled an empirical validation that agreement had indeed improved on the second half.

10.3.5 Revising the Coding Protocol

On occasion, inadequacies in the coding protocol will not be identified in the pilot testing phase. Indeed, the *Cochrane Handbook of Systematic Reviews* states that it is rare that a coding form does not require modification after piloting (Higgins and Green 2011). It may be, for example, that in the course of coding the fiftieth study, a coder becomes aware of an ambiguity in a categorical coding scheme that requires attention. It may come to the attention of those monitoring the coding process as a result of checks on coder reliability, such as a check for coder drift or a consensus discussion (for more, see chapter 9). When a coder modifies the way in which the coding protocol is used, it is important to identify that change and assess whether the modification is appropriate. If it is deemed appropriate, then the coding protocol is changed, and one must take steps to ensure the uniform (across coders and studies) and retroactive implementation of the change. Following such a practice will ultimately result in higher quality, more consistent coding than would be obtained if coders continued to use the original coding form.

10.3.6 Possessing Substantive Expertise

Substantive expertise will not reduce the need for judgment calls but should increase their accuracy. Numerous authors have stressed the need for substantive expertise, and with good reason: The synthesist who possesses it makes more informed and thoughtful judgments at all levels. Still, scholars with comparable expertise disagree frequently on matters of judgment in the social sciences and elsewhere, particularly when they bring preexisting biases, as noted earlier. Substantive expertise informs judgment, but will not guarantee that the right call was made. Employing low-inference coding protocols will reduce, but not eliminate, the need for expertise. We also note that low-inference codes often require a great deal of expertise to develop.

10.3.7 Improving Primary Reporting

Although the individual synthesist can do nothing about improving primary reporting, social science publications can do something to reduce error in synthesis coding. There is evidence of progress in this regard. A recent effort on the part of the American Psychological Association to encourage full reporting of descriptive statistics and effect sizes (Journal Article Reporting Standards Working Group 2007) resulted in the addition of an appendix on reporting standards to the current edition of the Publication Manual (American Psychological Association 2010). A similar policy has been adopted by a number of prominent journals in the medical field, in the form of the Consolidated Standards of Reporting Trials (CONSORT) (see Begg et al. 1996; Moher et al. 2001).

Research about the impact of these standards on actual practice is sparse at best. Lucy Turner and her colleagues report improved reporting in journals that have endorsed the CONSORT statement, but also note room for considerable further improvement (2012). Similar empirical work evaluating the guidelines suggested in the Journal Article Reporting Standards appears to be lacking.

10.3.8 Using Averaged Ratings

It is a psychometric truism that averages of multiple independent ratings will improve reliability (and therefore reduce error) relative to individual ratings. In principle, then, it is desirable for the synthesist to use such averages whenever possible. Two practical problems limit the applicability of this principle. First, the resources required to

double- or triple-code the entire set of studies can be substantial, particularly when the number of studies is large or the coding form is extensive. Second, many variables of interest in a typical synthesis are categorical, and therefore cannot be readily averaged across raters. For example, "therapy modality" from Smith, Glass, and Miller (1980) could be coded as individual, group, family, mixed, automated, or other. It is not clear how a mean rating would be possible for such a variable (with at least three coders, a median or a modal rating might make sense).

The first problem can be ameliorated by targeting for multiple coding only those variables that are known in advance to be both important to the analysis and prone to high error rates. Effect size and internal validity, for example, are two variables that would meet both criteria. The foreknowledge to select the variables can frequently be acquired through a targeted review of existing research, as well as through the synthesist's own pilot test (Cordray and Sonnefeld 1985). The second problem is more structural, being a property of the categorical measure. A commonsense solution might be, again, to target only those variables that are both important to the analysis and error-prone and have each coded by three raters. Three raters would permit a "majority rule" in the event that unanimity was not achieved. If the three coders selected three different responses—that is, there was no majority—the synthesist should probably consider revising how it is treated in the protocol.

10.3.9 Using Coder Consensus

Many meta-analyses are relatively limited in their scope, involving only a few effect sizes and potential explanatory variables. This situation is particularly common for syntheses of medical clinical trials, where the tendency is to focus on narrowly defined questions. Often under such circumstances it is possible to arrange for every study to be coded by two or more independent coders. When that occurs, it is appropriate to have periodic discussions during which consensus is sought if it is not already present. It may be the case that, where there is disagreement, one coder has noticed a detail that the others have missed, and that when the detail is described, all will concur with the minority coder. In that respect, consensus discussions may represent a distinct advantage over strategies such as averaging or majority rule, where the correct minority opinion would be obscured or overruled.

The consensus approach is not without pitfalls. It may be impractical or impossible to implement for projects of large scope. The consensus meeting may present opportunities for systematic coder bias that would not otherwise arise. One particularly insidious form of such bias occurs if there is a tendency for coders to defer to the most senior, who may be the most likely to have a conscious or unconscious agenda for the analysis. Nevertheless, coder consensus can be a highly effective approach. Indeed, it is common practice; of the twelve meta-analyses published in *Psychological Bulletin* in 2006 that reported any measure of rater agreement, half also mentioned that all disagreements were resolved, either by consensus or by appeal to a principal investigator. By the time of the third edition of this volume, reporting practices had changed. In forty-eight meta-analyses published in *Psychological Bulletin* during 2014 and 2015, forty reported double coding for at least some variables, and thirty-three of those reported resolution of conflicts by discussion.

10.4 STRATEGIES TO ASSESS OR CONTROL FOR ERROR

It is always better to improve the accuracy of measurement than to control or correct for measurement error after the fact. For example, methods such as the correction for attenuation from measurement error, though useful, tend to overcorrect for error, because the reliability estimates they use in the denominator often are biased downward. Yet strategies to control for error are necessary because, as documented earlier in this chapter, strategies to reduce error can succeed only to a limited degree. Whether because of deficient reporting quality, the limits of coder judgment, or a combination of the two, there will always be residual error that cannot be eliminated. The question then becomes how to address the error. This may be accomplished either through direct attempts at control (for example, employing reliability assessment or confidence ratings as covariates), or through sensitivity analyses that can account for the possible impact of coding error.

10.4.1 Reliability Assessment

10.4.1.1 Rationale As in primary research applications, the amount of observer error in research synthesis can be at least partially estimated via one or more forms of interrater reliability (IRR). The reanalysis of the Smith, Glass, and Miller psychotherapy data suggests that failing

to assess IRR and to consider it in subsequent analyses can yield misleading results (Orwin and Cordray 1985).

In the original synthesis of psychotherapy outcomes, a sixteen-variable simultaneous multiple regression analysis was used to predict outcomes. Identical regressions were run within three treatment classes and six subclasses, as defined through multidimensional scaling techniques with the assistance of substantive experts (see Smith, Glass, and Miller 1980). Orwin and Cordray recoded a sample of studies from the original synthesis and computed IRRs (1985). They then reran the original regression analyses with corrections for attenuation. Approximately twice per class and twice per subclass, on average, corrections based on one or more reliability estimates caused the sign of an uncorrected predictor's coefficient to reverse. The implication of a sign reversal (when significant) is that interpretation of the uncorrected coefficient would have led to an incorrect conclusion regarding the direction of the predictor's influence on the effect size. In every run in every class and subclass, the reliability correction altered the ranking of the predictors in their capacity to account for variance in effect size.

At least two of the major conclusions of the original study were brought into question by these findings. The first was that the study disconfirmed allegations by critics of psychotherapy that poor-quality research methods have accounted for observed positive outcomes.

That conclusion was based on the trivial amount of observed correlation ($r = 0.03$) between internal validity and effect size, which was taken as evidence that design quality had no effect. The reanalyses suggested that unreliability may have so seriously attenuated both the bivariate correlation between the two variables and the contribution of internal validity to the regression equations that only an unrealistically large relationship could have been detected. The reliability of internal validity may have been as low as 0.36 (Orwin and Cordray 1985).[9]

The second Smith, Glass, and Miller conclusion—bearing another look in light of these reanalyses—was that the outcomes of psychotherapy treatments cannot be very accurately predicted from characteristics of studies. Although the reliability-corrected regressions do not account for enough additional variance in effect size to claim "very accurate" prediction, they do improve the situation. It would be likely to improve still more with better-specified models, at least to the extent that reporting quality and coder judgment would permit them. Although unreliability alone cannot explain the poor performance of the Smith, Glass, and Miller model, it could be part of a larger process that does.

Unreliability in the primary study's dependent measure attenuates not only relationships with effect size, but the effect-size estimate itself. Under classical test theory assumptions, measurement error has no effect on the means of the treatment and comparison groups, but increases the within-group variance. In 1985, Larry Hedges and Ingram Olkin showed that under that model, the true effect size for a given study is equal to the observed effect size over the square root of the reliability of the dependent measure. If the dependent measure had a reliability of 0.70, for example, the estimated true effect size would equal the observed effect size times $1/0.70^{1/2}$, so the observed effect size would underestimate the true effect size by 16 percent. Error in coding effect-size estimates, then, exacerbates the problem by increasing the variance of the effect-size distribution at the aggregate level.

Despite wide recognition by writers on research synthesis of the need to assess IRR, in practice addressing coder reliability has always been problematic but is becoming less so. For example, only 29 percent of the meta-analyses published in *Psychological Bulletin* from 1986 through 1988 reported a measure of coder reliability (Yeaton and Wortman 1993). By 2006, a survey of the nineteen meta-analytic papers that appeared in *Psychological Bulletin* that year suggests that although the situation had changed for the better, double coding was still far from standard practice. Seven of the nineteen papers presented no information on coding reliability, and only eight reported on the reliability of coding for all variables. Interestingly, only three papers reported that effect-size estimates were double coded. By 2014–2015, forty of forty-eight published papers reported double coding for at least some of the variables and thirty-four reported the use of reliability statistics. Given the high methodological standards of *Psychological Bulletin* relative to many other social research journals that publish meta-analyses, the field-wide percentage may be significantly lower; nevertheless, the trend appears to be toward more frequent double coding.

10.4.1.2 Across-the-Board Versus Per-Variable Agreement Smith, Glass, and Miller's reliability assessment consisted of the computation of a simple agreement rate across all variables in an abbreviated coding form, which came to 92 percent (1980). According to William Stock and his colleagues, who examined the practice of synthesists regarding interrater reliabilities, subsequent synthesists did much the same thing, if that much (1982). That is, a single agreement rate, falling somewhere between 0.7 and 1.0, is the extent of what typically gets

reported. That picture has improved in the interval since 1982. Thirty-four meta-analyses published in *Psychological Bulletin* in 2014 and 2015 reported some form of interrater reliability statistics. Of those, only seven papers reported a single overall value, as opposed to fourteen that reported at least some coding reliability results for individual variables. Ten articles reported a range of reliability coefficients. The remaining three reported perfect reliability after differences were resolved, with no information about initial coding reliability.

There are at least two major problems with this practice. First, it makes little psychometric sense. The coding form is simply a list of items; it is not a multi-item measure of a trait, such that a total-scale reliability would be meaningful. Some items, such as publication date, will have very high interrater agreement, whereas others, such as internal validity, may not. Particularly if reliabilities are to be meaningfully incorporated into subsequent analyses (for example, by correcting correlation matrices for attenuation), it is this variation that needs to be recognized. In their 1985 reanalysis of the 1980 Smith, Glass, and Miller psychotherapy data, Orwin and Cordray replicated an equivalently high overall agreement rate. However, agreement across individual variables ranged from 0.24 to 1.00 (other indices of IRR showed similar variability).

Second, an across-the-board reliability fails to inform the synthesist of specific variables needing refinement or replacement. Thus, an opportunity to improve the process is lost. In sum, the synthesist should assess IRR on a per-variable basis. In addition, the reliability of scales constructed from several variables by the meta-analyst should be assessed. We discourage the common practice of reporting a single summary reliability across a number of items.

10.4.1.3 Specific Indices of Interrater Reliability Laurel Oliver notes that authorities do not agree on the best index of IRR to use in coding syntheses (1987). This presentation does not attempt to resolve all the controversies, but will—it is hoped—provide enough of a foundation for the synthesist who is not a statistician to make informed choices. Six indices are presented: agreement rate, kappa and weighted kappa, delta, Krippendorff's alpha, intercoder correlation, and intraclass correlation. The discussion of each will include a description (including formulas when possible), a computational illustration, and a discussion of strengths and limitations in the context of research synthesis. The following section covers the selection, interpretation, and reporting of IRR indices.

10.4.1.3.1 Agreement Rate. Percentage agreement, alternately called agreement rate (*AR*), has been the most widely used index of IRR in research synthesis. The formula for *AR* is as follows:

$$AR = \frac{\text{number of observations agreed upon}}{\text{total number of observations}}. \quad (10.1)$$

Table 10.1 presents a hypothetical data set that might have been created had three coders independently rated twenty-five studies on a three-point study characteristic (two of which we use for the following examples). For example, if this variable was the internal validity scale from the 1980 Smith, Glass, and Miller study, a rating of 1 = low, 2 = medium, and 3 = high. As shown, the first two coders agreed in fifteen cases out of twenty-five, so $AR = 0.60$.

AR is computationally simple and intuitively interpretable, being basically a batting average. Yet numerous writers on observational measurement have discussed the

Table 10.1 Illustrative Data: Ratings of Studies

	Coder		
Study	1	2	3
1	3	2	3
2	3	1	1
3	2	2	2
4	3	2	3
5	1	1	1
6	3	1	3
7	2	2	1
8	1	1	1
9	2	2	1
10	2	1	3
11	2	2	2
12	3	3	3
13	3	1	2
14	2	1	1
15	1	1	1
16	1	1	2
17	3	3	1
18	2	2	2
19	2	2	2
20	3	1	1
21	2	1	2
22	1	1	3
23	3	2	2
24	3	3	3
25	2	2	3

SOURCE: Authors' compilation.

pitfalls of using it (Cohen 1960; Hartmann 1977; Light 1971; Scott 1955). When variables are categorical (the usual application), the main problem is chance agreement, particularly when response marginal totals are extreme. For example, the expected agreement between two raters on a yes-no item in which each rater's marginal response rate is 10 to 90 percent would be 82 percent by chance alone (Hartmann 1977). In other words, these raters could post a respectable (by most standards) interrater reliability simply by guessing without ever having observed an actual case. Extreme marginal response rates are commonplace in many contexts (for example, psychiatric diagnosis of low-prevalence disorders), including research synthesis (see, for example, historical effects in Kulik, Kulik, and Cohen 1979; design validity in Smith 1980). Additional problems arise when marginal response rates differ across raters (Cohen 1960).

When applied to ordinal (as opposed to nominal) categorical variables, *AR* has an additional drawback: the inability to discriminate between degrees of disagreement. In Smith, Glass, and Miller's three-point internal validity scale (low, medium, high), for example, a low-high interrater pattern indicates greater disagreement than a low-medium or medium-high pattern does, yet a simple *AR* registers identical disagreement for all three patterns (1980). The situation is taken to the extreme with quantitative variables.

10.4.1.3.2 Cohen's Kappa and Weighted Kappa. Various statistics have been proposed for categorical data to improve on *AR*, particularly with regard to removing chance agreement (see Light 1971). Of these, Cohen's kappa (κ) has frequently received high marks (Cohen 1960; Fleiss, Cohen, and Everett 1969; Hartmann 1977; Light 1971; Shrout, Spitzer, and Fleiss 1987). The parameter κ is defined as the proportion of the best possible improvement over chance that is actually obtained by the raters (Shrout, Spitzer, and Fleiss 1987). The formula for the estimate *K* of kappa computed from a sample is as follows:

$$K = \frac{P_o - P_e}{1 - P_e}, \tag{10.2}$$

where P_o and P_e are the observed and expected agreement rates, respectively. The observed agreement rate is the proportion of the total count for which there is perfect agreement between raters (that is, the sum of the diagonals in the contingency table divided by the total count). The expected agreement rate is the sum of the expected agreement cell probabilities, which are computed exactly as in a chi-square test of association. That is,

$$P_e = \frac{1}{n^2}\sum_{i=1}^{C} n_{i\bullet}\, n_{\bullet i}, \tag{10.3}$$

where n is the number of observations, C is the number of response categories, and $n_{i\bullet}$ and $n_{\bullet i}$ are the observed row and column marginal totals for response i for raters 1 and 2, respectively. The formula for the estimate of weighted kappa (K_w) is as follows:

$$K_w = 1 - \frac{\Sigma_{ij} w_{ij} P_{oij}}{\Sigma_{ij} w_{ij} P_{oij}}, \tag{10.4}$$

where w_{ij} is the disagreement level assigned to the cell at the intersection of row i, column j, and P_{oij} and P_{eij} are the observed and expected proportions, respectively, in row i and column j. For weighted kappa, proportions are calculated for every cell combination instead of only combinations on the diagonal. In this case, the formulas for the proportions change slightly to $P_{oij} = n_{ij}/n$ for each observed proportion and $P_{eij} = \frac{1}{n^2}\sum_{ij} n_{i\bullet} n_{\bullet j}$ for each expected proportion. If regular (unweighted) K is re-expressed as $1 - D_o/D_e$, where D_o and D_e are observed and expected proportions of disagreement, it can be seen that K is a special case of K_w in which all disagreement weights equal 1.

Panel A of table 10.2 shows the cell counts and marginal totals from the illustrative data shown in table 10.1. Panel B shows the observed and expected proportions in all cells. For simple kappa, the sum of the diagonal expected proportions in the table is P_e = [(12)(5) + (10)(10) + (3)(10)] / $(25)^2$ = 0.304. The proportion of observed agreement, P_o, is the sum of the diagonal of observed proportions (5 + 7 + 3) / 25 = 0.6, so that kappa is K = (0.6 – 0.304) / (1 – 0.304) = 0.43. Thus, chance-corrected agreement in this example is slightly less than half of what it could have been.

To compute K_w, the weights (w_{ij}) must be assigned first. Agreement cells (the diagonals) are assigned weights of 0. With a three-point scale such as this one, some logical weights would be 2 for the low-high interrater pattern and 1 for the low-medium and medium-high patterns, although other weights are possible of course. Those weights are shown in panel C of table 10.2. Once the weights are determined, the observed and

Table 10.2 Illustrative Data: Cell Counts and Marginal Totals

		Coder 1			
	Value	1	2	3	Sum
A: Observed cell counts					
	1	5	3	4	12
Coder 2	2	0	7	3	10
	3	0	0	3	3
	Sum	5	10	10	25
B: Observed (expected) cell proportions					
	1	.200 (.096)	.120 (.192)	.160 (.192)	
Coder 2	2	.000 (.080)	.280 (.160)	.120 (.160)	
	3	.000 (.024)	.000 (.048)	.120 (.048)	
C: Weights					
	1	0	1	2	
Coder 2	2	1	0	1	
	3	2	1	0	

SOURCE: Authors' compilation.

expected proportions are calculated for each cell. For the cell at row 1 and column 2, $P_{o12} = \frac{3}{25} = 0.12$ and $P_{e12} = \frac{1}{25^2} 12 * 10 = 0.192$. Those values and all other proportions are shown in panel B. Summing across the elements of $w_{ij}P_{oij}$ products for each cell yields $\Sigma\, w_{ij}P_{oij} = 0.56$. Similarly, summing across the $w_{ij}P_{eij}$ produces $\Sigma\, w_{ij}P_{eij} = 0.912$. Then, $K_w = 1 - 0.56 / 0.912 = 0.39$. In this case, a relatively modest percentage of disagreements (four of ten) were by two scale points, as one would hope to be the case with a 3-point scale. Consequently, K_w is only slightly lower than K. With more scale points, the difference will generally be larger.

As is evident from the formula for K, chance agreement is directly removed from both numerator and denominator. Kappa has other desirable properties:

It is a true reliability statistic, which in large samples is equivalent to the intraclass correlation coefficient (discussed later in this section; see also Fleiss and Cohen 1973).

Because P_e is computed from observed marginal totals, no assumption of identical marginal totals across raters (required with certain earlier statistics) is needed.

It can be generalized to multiple (that is, more than two) raters (Fleiss 1971; Light 1971).

It can be weighted to reflect varying degrees of disagreement (Cohen 1968), for example, for the Smith, Glass, and Miller trichotomous internal validity variable mentioned earlier.

Large-sample standard errors have been derived for both K and K_w (Fleiss, Cohen, and Everett 1969), thus permitting the use of significance tests and confidence intervals.

It takes on negative values when agreement is less than chance (range is –1 to 1), thus indicating the presence of systematic disagreement as well as agreement.

It can be adapted to stratified reliability designs (Shrout, Spitzer, and Fleiss 1987).

Thus, K is not a single index, but rather a family of indices that can be adapted to various circumstances.

One issue that the analyst must keep in mind is that indices such as K may not be well estimated with the sample sizes available for some meta-analyses. Asymptotic standard errors for K and weighted K are available (Fleiss, Cohen, and Everett 1969); these may be relatively large unless the number of studies coded is high. It

is impossible to provide a simple rule of thumb for a minimum sample size because the standard errors depend partly on the cell proportions. However, it is noteworthy that for the data from table 10.1 (which produced an estimate of $K = 0.43$), the standard error is 0.12, leading to a 95 percent confidence interval from 0.19 to 0.67. For the example provided in the original paper on asymptotic standard errors (Fleiss, Cohen, and Everett 1969), a sample size of 60 would be required for the standard error of unweighted K to be as low as 0.10. Hence, for the numbers of studies present in many meta-analyses, there may be very little information about coding reliability.

Although K and K_w resolve most of the problems of *AR*, potential problems remain if observations are concentrated in only a few cells (Jones et al. 1983). This circumstance increases the probability that a high proportion of scores will be assigned to one or two rating categories and that other categories receive small or nonexistent proportions. Allan Jones and his colleagues and others, such as Nancy Burton (1981), argue that this distribution violates the assumption on which K and K_w are based, reducing the usefulness of the information. In a similar vein, Grover Whitehurst notes that K is sensitive to the degree of disagreement between raters, which is desirable, but also remarkably sensitive to the distribution of ratings, which he argued was not desirable (1984). A highly skewed distribution of ratings will yield high estimates of chance agreement in just those cells in which obtained agreement is also high. This makes for low estimates of true agreement in the face of high levels of obtained agreement and, consequently, low estimates of IRR. As noted earlier, the phenomenon is common in psychiatric diagnosis when prevalence is low and has been termed the base-rate problem with K (Carey and Gottesman 1978).

However, others view this as a case of shooting the messenger (see for example, Shrout, Spitzer, and Fleiss 1987). As they see it, the observation that low base rates decrease K is not an indictment of K, but rather represents the real problem of making distinctions in increasingly homogeneous populations. Similarly, Rebecca Zwick questions whether the sensitivity of K to the shape of the marginal distributions is necessarily undesirable: if cases are concentrated into a small number of categories, it is less demonstrable that coders can reliably discriminate among *all C* categories, and the IRR coefficient should reflect this (1988). Finally, there is no mathematical necessity for small K values with low base rates (the maximum value of K remains at 1.0 even when the base rate is low) and high Ks have in fact been demonstrated empirically with base rates as low as 2 percent (American Psychiatric Association 1980).

10.4.1.3.3 Andrés and Marzo's Delta. One measure of agreement that addresses the base-rate problem of kappa (if, indeed, that is a problem) adapts theory from guessing corrections in multiple-choice testing (Andrés and Marzo 2004). The proposed index ($\hat{\Delta}$) has one somewhat more relaxed assumption than K—nonconcordance rather than independence of raters. In addition to allowing an assessment of the concordance of multiple raters, the index can be computed specifying one rater as a gold standard. It depends, however, on an assumed probability model that describes how raters choose among the options. Moreover, the index lacks a closed-form solution, relying instead on iterative maximum likelihood estimation (for which software is readily available). Nonetheless, the approach does have the advantage of tending to agree with K in the absence of the base-rate problem, and tending to produce higher values that may be more representative of actual coder agreement when the base-rate problem is present. For the example of table 10.2 (for which unweighted K was 0.43), the value of delta-hat is 0.41.

10.4.1.3.4 Krippendorff's Alpha. From the domain of content analysis, Klaus Krippendorff offers a generalizable reliability coefficient that can apply to any number of raters and any type of data, including missing values (2011). In general, the statistic takes the form of

$$\alpha = 1 - \frac{D_O}{D_E}, \tag{10.5}$$

where D_O and D_E are observed and expected disagreement rates, respectively. This notation echoes that used to describe unweighted kappa. In the case of Krippendorff's alpha, the functional definitions of D_O and D_E differ from those for kappa. The question of exactly what form these disagreement rates take is a complicated function of the level of measurement of the data. However, Krippendorff argues that the statistic provides a general one-size-fits-all coefficient that simultaneously corrects for chance agreement, data type, and sample size (2011). Although the statistic is used most frequently in purely qualitative studies, it is seeing increasing use in meta-analyses in the social sciences (see, for example, Mares and Pan 2013; Vukasović and Bratko 2015). For the data in table 10.1, the value of Krippendorff's alpha is 0.403 if the data are treated as nominal, 0.278 if ordinal, and 0.311 if ratio.

10.4.1.3.5 Intercoder Correlation. *K* and its alternatives were designed for categorical data and are not appropriate for continuous variables, to which we now turn. Numerous statistics have been suggested to assess IRR on continuous variables as well. One of the more popular is the common Pearson correlation coefficient (*r*), sometimes called the intercoder correlation in this context. Pearson *r* is as follows:

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{s_X}\right)\left(\frac{Y_i - \overline{Y}}{s_Y}\right), \qquad (10.6)$$

where (X_i,Y_i) are the n pairs of values, and s_X and s_Y are the standard deviations of the two variables computed using n in the denominator. With more than two coders, this index is generally obtained by calculating *r* across all pairs of coders rating the phenomenon (Jones et al. 1983). The resulting correlations can then be averaged to determine an overall value.

For the continuous examples, we can think of our ratings in table 10.1 as continuous instead of categorical. Then, $\overline{X}$ and $\overline{Y}$ are 2.2 and 1.64, and s_X and s_Y are 0.75 and 0.69, respectively, and n is 25. Plugging the values into equation 10.5, $r = 5.8 / 12.94 = 0.45$. In practice, the synthesist will only rarely need to calculate *r* by hand, as all standard statistical packages and many hand calculators compute it.

The use of Pearson *r* with observational data to estimate IRRs is analogous to its use in education to estimate test reliabilities when parallel test forms are available (see Stanley 1971). Thus it bears the same relationship to formal reliability theory.[10] The Pearson *r* also has some drawbacks. First, although it describes the degree to which the scores produced by each coder covary, it says nothing about the degree to which the scores themselves are identical. In principle, this means that coders can produce high *r*s without actually agreeing on any scores. Conversely, an increase in absolute agreement could actually reduce *r* if the disagreement had been part of a consistent disagreement pattern (for example, one rater rating targets consistently lower than the other).[11] If the IRR estimate will be used only to adjust subsequent analyses, this may not be particularly important. If it will be used to diagnose coder consistency in interpreting instructions (for example, instructions for extracting effect sizes), it can be quite important. Second, the between-coders variance is always removed in computing the product-moment formula. Robert Ebel notes that this is especially problematic when comparisons are made among single raw scores assigned to different subjects by different coders—in research synthesis, of course, the subject is the individual study (1951).

10.4.1.3.6 Intraclass Correlation. Analogous to the concepts of true score and error in classical reliability theory, the intraclass correlation (r_I) is computed as a ratio of the variance of interest over the sum of the variance of interest plus error. Ebel suggests that r_I was preferable to Pearson *r* as an index of IRR with continuous variables because it permits the researcher to choose whether to include the between-raters variance in the error term (1951).

Like *K*, r_I is not a single index but a family of indices that permit great flexibility in matching the form of the reliability index to the reliability design used by the synthesist; here we mean design in the sense used in Cronbach's generalizability (*G*) theory (see Cronbach et al. 1972). Generalizability theory enables the analyst to isolate different sources of variation in the measurement, for example, forms or occasions, and to estimate their magnitude using the analysis of variance (Shavelson and Webb 1991). The reliability design as discussed here is a special case of the one-facet *G* study, in which coders are the facet (Shrout and Fleiss 1979).

So far, the running example has been discussed in terms of twenty-five studies being coded by two coders, but in fact at least three reliability designs share this description.

Design 1: Each study is rated by a different pair of coders, randomly selected from a larger population of coders (one-way random-effects model).

Design 2: A random pair of coders is selected from a larger population, and each coder rates all twenty-five studies (two-way random-effects model).

Design 3: All twenty-five studies are rated by each of the same pair of coders, who are the only coders of interest (two-way mixed-effects model).

Designs 1 and 2 are common in research synthesis. Although actual random selection of coders is rare, the synthesist's usual intent is that those selected represent, at least in principle, a larger population of coders who might have been selected. Whether explicitly stated or not, synthesists typically present their substantive findings as generalizable across that population. The findings will hold little scientific interest if, say, only a particular subset of graduate students at a particular university can reproduce them. Design 3 will hold in rare instances in

which the specific coders selected are the population of interest and can therefore be modeled as a fixed effect. This could happen if, for example, the coders are a specially convened panel of all known experts in the content area (for example, the Wortman and Bryant school desegregation synthesis described earlier might have met this criterion).

Each design requires a different form of r_I, based on a different analysis of variance (ANOVA) structure. These forms for the running example are estimated as follows:

$$r_I\,(\text{design 1}) = \frac{BMS - WMS}{BMS + WMS}, \tag{10.7}$$

$$r_I\,(\text{design 2}) = \frac{BMS - EMS}{BMS + EMS + 2(CMS - EMS)/25} \tag{10.8}$$

$$r_I\,(\text{design 3}) = \frac{BMS - EMS}{BMS + EMS}, \tag{10.9}$$

where *BMS* and *WMS* are the between-studies and within-study mean squares, respectively, and *CMS* and *EMS* are the between-coders and residual (error) mean square components, resulting from the partitioning of the within-study sum of squares (for a full discussion of the statistical models underlying the estimates, see Shrout and Fleiss 1979).[12]

The next step is the computation of sums of squares and mean squares from the coder data (for computational details, see any standard statistics text, for example, Hays 1994). Table 10.3 shows the various mean squares and their associated degrees of freedom for the data from table 10.1. That approach to the problem can handle only balanced designs (that is, ones in which the same raters rate the same studies). Plugging the values into equations 10.7, 10.8, and 10.9 yields the following results:

$$r_I\,(\text{design 1}) = 0.28,$$
$$r_I\,(\text{design 2}) = 0.35,$$
$$r_I\,(\text{design 3}) = 0.44.$$

On average, r_I (design 1) will give smaller values than r_I (design 2) or r_I (design 3) for the same set of data (Shrout and Fleiss 1979).

More modern mixed-model approaches circumvent the need for a balanced design and allow use with alternative, non-normal distributions (Marcoulides 1990). Using maximum likelihood estimation instead of ANOVA loosens the restrictions of our assumptions, often with few drawbacks (Donner and Koval 1980). Instead of mean squares, the mixed-model intraclass correlation can be calculated as a ratio of variance components:

$$r_I = \frac{\sigma^2_{Study}}{\sigma^2_{Study} + \sigma^2_{Error}}, \tag{10.10}$$

where σ^2_{Study} is the variance component estimated from study-related variance and σ^2_{Error} is the estimated residual error. The variance component for design 1 is taken from a model in which only the studies are treated as random effects. For design 2, the regression model treats both studies and raters as random effects. For design 3, studies are treated as random effects and raters are treated as fixed effects. Each model results in different values for the variance components, but regardless of the model intended, all three designs employ the same equation for the intraclass correlation (but for specific components of the estimated residual error between designs, see McGraw and Wong 1996, table 4).

When its assumptions are met, the r_I family has been mathematically shown to provide appropriate estimates of classical reliability for the IRR case (Lord and Novick 1968). Its flexibility and linkage into *G* theory also argue in its favor.[13] Perhaps most important, its different variants reinforce the idea that a good IRR assessment requires the synthesist to think through the appropriate IRR design, not just calculate reliability statistics. Like *K*, it requires substantial between-items variance to show a significant indication of agreement. As with *K*, some writers consider r_I to be less useful as an index of IRR when the distributions are concentrated in a small range (for example,

Table 10.3 Analysis of Variance for Illustrative Ratings

Source of Variance	Degrees of Freedom	Mean Squares
Between-studies (BMS)	24	.78
Within-study (EMS)	25	.44
Between-coders (CMS)	1	3.92
Residual (EMS)	24	.30

SOURCE: Authors' compilation.

Table 10.4 Estimates of Interrater Agreement for Different Types of Data Distributions

Distributional Conditions	Kappa	Weighted Kappa	Agreement Rate	Average Correlation	Intraclass Correlation
Variations in ratings across jobs and high agreement among raters	.43	.45	.88	.79	.74
Variations in ratings across jobs and low agreement among raters	.01	.04	.16	.13	.05
Little variation in ratings across jobs and high agreement among raters	.04	.04	.77	–.01	–.03

SOURCE: Jones et al. 1983.

Jones et al. 1983). The arguments and counterarguments on this issue are essentially as described for *K*.

10.4.1.4 Selecting, Interpreting, and Reporting Interrater Reliability Indices With the exception of *AR*, the indices presented in the previous section have appealing psychometric properties as reliability measures and, when properly applied, can be recommended for use in evaluating coding decisions. The *AR* was included because of its simplicity and widespread use by research synthesis practitioners, but for the reasons discussed earlier, it cannot be recommended as highly. These six indices by no means exhaust all those proposed in the literature for assessing IRR. Most of these, such as Finn's *r* (Whitehurst 1984), Yule's *Y* (Spitznagel and Helzer 1985), Maxwell's *RE* (Janes 1979), and Scott's π (Zwick 1988), were proposed to improve on perceived shortcomings of the indices presented here, in particular the r_I, and *K* family. Interested readers can explore that literature and draw their own conclusions. They should keep in mind, however, that the prevailing view among statisticians is that, to date, alternatives to r_I and *K* are *not* improvements, but in fact propose misleading solutions to misunderstood problems (see, for example, Cicchetti 1985 on Whitehurst 1984; Shrout, Spitzer, and Fleiss 1987 on Spitznagel and Helzer 1985). The synthesist who sticks to the appropriate form of *K* for categorical variables and the appropriate model of r_I for continuous variables will be on solid ground. Either *AR* or *r* can be used to supplement *K* and r_I for the purpose of computing multiple indices for sensitivity analysis. The synthesist who chooses to rely on one of the less common alternatives (such as delta) would be wise to first become intimately familiar with its strengths and weaknesses and should probably anticipate defending the choice. The defense is likely to be viewed more favorably if the alternative approach is used to supplement one of the more usual indexes.

Once an index is selected and computed, the obvious question is how large it should be. The answer is less obvious. For *r*, 0.80 is considered adequate by many psychometricians (for example, Nunnally 1978), and at that level correlations between variables are attenuated very little by measurement error. Rules of thumb have also been suggested and used for *K* and r_I (Fleiss 1981; Cicchetti and Sparrow 1981).

These benchmarks, however, were suggested for evaluating IRR of psychiatric diagnoses, and their appropriateness for synthesis coding decisions is not clear. Indeed, general caution is usually appropriate about statistical rules of thumb that are not tied to a specific context; for that reason, we believe it inappropriate to provide them here (for a similar take on the inappropriateness of context-independent rules of thumb, see chapter 7).

The issue of how large is large enough is further complicated by distributional variation. The running example showed how the different indices varied for a particular data set, but not how they vary over different conditions. Allan Jones and his colleagues systematically compare the results obtained when these indices are applied to a common set of ratings under various distributional assumptions (1983). Four raters rated job descriptions using a questionnaire designed and widely used (and validated) for that purpose. Most ratings were on six-point Likert-type scales with specific anchors for each point; a few were dichotomous. The *K*, K_w, *AR*, *r* (in the form of average pairwise correlations), and r_I were computed on each item. Aggregated across items, *K* and K_w yielded the lowest estimates of IRR, producing median values of 0.19 and 0.22, respectively. The *AR* yielded the highest, with a

median value of 0.63; r and r_I occupied an intermediate position, with median values of 0.51 and 0.39, respectively. For the reasons noted earlier, an overall aggregate agreement index is not particularly meaningful for evaluating coding decisions, but is still useful for illustrating the wide variation among indices.

To examine the effect of the distributions of individual items on the variation across indices, sample items representing three conditions were analyzed: high variation across ratings and high agreement across raters, high variation across ratings and low agreement across raters, and low variation across ratings and high agreement across raters. As shown in table 10.4, the effects can be substantial, in particular under the third condition, where *AR* registered 77 percent agreement, but the other indices suggested no agreement beyond chance. With moderate to high variance, it makes far less difference which index is used. Therefore, Jones and his colleagues argue, a proper interpretation of the different indices—including whether they are large enough—requires an understanding of the actual distributions of the data (1983).

It should be evident by now that the indices are not directly comparable; indeed the range of their metrics is not the same across all measures. For example, correlations range from –1 to 1 where negative values indicate extreme disagreement. On the other hand, kappa and Krippendorff's alpha range from –1 to 1, with negative values meaning that raters are agreeing beyond expectation. Other metrics such as AR and delta range from 0 to 1. It is therefore essential that synthesists report not only their IRR values, but the indices used to compute them. To give the reader the full picture, it would also be wise to include information about the raters' base rates, as William Grove and his colleagues suggest, particularly when they are very low or very high (1981).

10.4.1.5 Assessing Coder Drift Whatever index of IRR is chosen, the synthesist cannot assume that IRR will remain stable throughout the coding period, particularly when the number of studies to code is large. As Gregg Jackson notes, coder instability arises because many hours are required, often over a period of months, to code a set of studies (1980). When the coding is lengthy, IRR may change over time, and a single assessment may not be adequate.

Orwin and Cordray assess coder drift in their 1985 reanalysis of the 1980 Smith, Glass, and Miller psychotherapy synthesis. Before coding, the twenty-five reports in the main sample were randomly ordered, with coders instructed to adhere to the resulting sequence. It was assumed, given the random ordering, that any trend in IRRs over reports would be attributable to changes in coders over time (for example, practice or boredom).[14] Equating for order permits this trend to be detected and, if desired, removed.

Following the completion of coding, a per-case agreement rate was computed, consisting of the number of variables agreed on divided by the number of variables coded.[15] In itself, per-case agreement rate is not particularly meaningful because it weights observations on all variables equally and is sensitive to the particular subset of variables (considered here as sampled from a larger domain) constituting the coding form. Change in this indicator over time is meaningful, however, because it speaks to the stability of agreement. Per-case agreement rate was plotted against case sequence, and little if any drift was observed. The absence of significant correlation between the two variables substantiated this conclusion. In regression terms, increasing the case sequence by one resulted in a decrease of the agreement rate by less than one-tenth of 1 percent ($b = -0.09$). From this exercise, it was concluded that further consideration of coder drift was unnecessary in the synthesis. However, there is no reason to presume that this finding would generalize to other syntheses. Coder drift will be more of a concern in some syntheses than in others and needs to be assessed on a case-by-case basis. In cases where coder drift is evident, some combination of revisiting the coding scheme, retraining coders, and attempting to reach consensus can be useful.

The drift assessment could also be integrated into the main IRR assessment with a *G* theory approach. The one-facet *G* study with coders as the facet could be expanded to a two-facet *G* study with time (if measured continuously) or occasions (if measured discretely) as the second facet. This would permit the computation of a single IRR coefficient that captured the variance contributed by each facet, as well as by their interaction.

10.4.2 Confidence Ratings

10.4.2.1 Rationale As a strategy for evaluating coding decisions, IRR is indirect. Directly, it assesses only coder disagreement, which in turn is an indicator of coder uncertainty, but a flawed one. It is desirable, therefore, to seek more direct methods of assessing the accuracy of the numbers gleaned from reports. Tagging each number with a confidence rating is one such method. It is based on the premise that questionable information should not be

discarded, and should not be allowed to freely mingle with less questionable information, as is usually done. The former procedure wastes information, which though flawed may be the best available on a given variable, whereas the latter injects noise at best and bias at worst into a system already beset by both problems. With confidence ratings, questionable information can be described as such, and both data point and descriptor entered into the database.

In other contexts, confidence ratings and facsimiles (for example, adequacy ratings, certainty ratings) have been around for some time. In his international comparative political parties project, Kenneth Janda constructed an adequacy-confidence scale to evaluate data quality (1970). Similarly, it is not unusual to find interviewer confidence ratings embedded in questionnaires. In the *Addiction Severity Index*, for example, interviewers provided confidence ratings of the accuracy of patient responses (McLellan et al. 1988). Specifically, they coded whether in their view the information was significantly distorted by the patient's misrepresentation or the patient's inability to understand.

Two early meta-analyses are used to illustrate how some of the preceding issues can be assessed. One integrates the literature on the effects of psychotherapy (Smith and Glass 1977; Smith, Glass, and Miller 1980). Another is devoted to the influence of class size on achievement (Glass and Smith 1979). For both, the data files were obtained and subjected to reanalysis (Cordray and Orwin 1981; Orwin and Cordray 1985). In examining the primary studies, these authors were careful to note (for some variables) how the data were obtained. For example, in the class size study, the actual number of students enrolled in each class was not always reported directly, so the values recorded in the synthesis were not always based on equally accurate information. In response, Glass and Smith included a second variable that scored the perceived accuracy of the numbers recorded. Unfortunately, the accuracy scales were never used in the analysis (Gene Glass, personal communication, August 1981). Similarly, in the psychotherapy study, source of IQ and confidence of treatment classifications were also coded and not used.

The synthesis of class size and achievement compared smaller and larger classes on achievement (Glass and Smith 1979). After assigning each comparison an effect size, the authors regressed effect size on three variables: the size of the smaller class (S), the size of the smaller class squared (S^2), and the difference between the larger class and smaller class (L—S). This procedure was done for the total database and for several subdivisions of the data, one of which was well controlled versus poorly controlled studies. Cordray and Orwin reran the regressions as originally reported, except that a dummy variable for overall accuracy was added to the equations (1981). The results are presented in table 10.5. Most remarkable is that accuracy made its only nontrivial contribution in well controlled studies. A possible explanation is that differential accuracy in the reporting of class sizes is only one of many method factors operating in poorly controlled studies (and in the entire sample, as well-controlled studies are a minority) and that these mask the influence of accuracy by interacting with it in unknown ways or by inflating error variance. In any event, the considerable influence of accuracy in well-controlled studies suggested that it does matter, at least under some circumstances.

In the synthesis of psychotherapy outcomes described earlier, an identical simultaneous multiple regression analysis was run within each treatment class and subclass, but with a different orientation. Rather than specifying the treatment in the model and crosscutting by nontreatment characteristics, as with class size, the nontreatment characteristics (diagnosis, method of client solicitation, and so on) were included in the model. Orwin and Cordray recoded twenty variables and found that across items, high confidence proportions ranged from 1.00 for comparison type and location down to 0.09 for therapist experience (1985; for selection criteria, see Orwin 1983). Across studies, the proportion of variables coded with certainty or near certainty ranged from 52 to 83 percent. At the opposite pole, the proportion of guesses ranged from 0 to 25 percent.

The effect of confidence on reliabilities was remarkable; for example, the mean AR for high confidence

Table 10.5 Comparison of R^2 for Original Glass and Smith (1979) Class Size Regression (R_1^2) and Original with Accuracy Added (R_2^2)

	R_1^2	R_2^2	$R_2^2 - R_1^2$
Total sample ($n = 699$)	.1799	.1845	.0046, $F_{(1, 694)} = 3.94$*
Well-controlled studies ($n = 110$)	.3797	.4273	.0476, $F_{(1, 105)} = 8.73$*
Poorly controlled studies ($n = 338$)	.0363	.0369	.0006, $F_{(1, 333)} = 0.65$

SOURCE: Cordray and Orwin 1981.

*$p < .05$

observations more than doubled that for low confidence observations (0.92 versus 0.44). As to the effect of confidence on relationships between variables, 82 percent of the correlations increased in absolute value when observations rated with low and medium confidence were removed. Moreover, all correlations with effect size increased. The special importance of these should be self-evident; the relationships between study characteristics and effect size were a major focus of the original study (and are in research synthesis generally). The absolute values of correlations increased by an average of 0.15 (0.21–0.36), or in relative terms, 71 percent.

Questions posed by these findings include how credible the lack of observed relationship between therapist experience and effect size can be when therapist experience is extracted from only one in ten reports with confidence. Even for variables averaging considerably higher confidence ratings, the depression of IRRs and consequent attenuation of observed relationships between variables is evident. There are plausible alternative explanations for individual variables. For example, the outcome measures that are easiest to classify could also be the most reliable. In such circumstance, the outcome type effect-size relationship should be stronger for high confidence observations simply because the less reliable outcome types have been weeded out. An explanation like this can be plausible for individual variables, but cannot account for the generally consistent and nontrivial strengthening of observed relationships throughout the system. Each study in the sample has its own pattern of well-reported and poorly reported variables; there was not a particular subset of studies, presumably different from the rest, that chronically fell out when high confidence observations were isolated. Consequently, any alternative explanation must be individually tailored to a particular relationship. Postulating a separate ad hoc explanation for each relationship would be unparsimonious to say the least.

10.4.2.2 Empirical Distinctness from Reliability It was argued previously that confidence judgments provide a more direct method of evaluating coding decisions than does IRR. It is therefore useful to ask whether this conceptual distinctness is supported by an empirical distinctness.

Table 10.6 presents agreement rates by level of confidence for the complete set of variables ($K = 25$) to which confidence judgments were applied. The table indicates that interrater agreement is neither guaranteed by high confidence nor precluded by low confidence. Yet it also shows that confidence and agreement are associated. Whereas table 10.6 shows the nonduplicativeness of confidence and agreement on a per-variable basis, table 10.7 shows this nonduplicativeness across variables. To calculate the values in table 10.7, variables were first rank ordered by the proportion of observations in which confidence was judged as high. As is clear in the table, the correlation between the two sets of rankings was modest, regardless of the reliability estimate selected. Although some of the low correlations could be attributed to attenuation or poor validity, it is probable that each measure has something unique to contribute. Even if they represented exactly the same construct, it would be valuable to include both in the analysis.

10.4.2.3 Methods for Assessing Coder Confidence There is no set method for assessing confidence. Orwin and Cordray used a single confidence item for each variable (1985). This is not to imply that confidence is unidimensional; no doubt it is a complex composite of numerous factors. These factors can interact in various ways to affect the coder's ultimate confidence judgment, but Orwin and Cordray did not attempt to spell out rules for handling the many contingencies that arise. Confidence judgments reflected the overall pattern of information as deemed appropriate by the coder.

Alternative schemes are of course possible. One might use two confidence judgments per data point—one rating confidence in the accuracy or completeness of the information as reported and the other rating confidence in the coding interpretation applied to that information. The use of two or more such ratings explicitly recognizes multiple sources of error in the coding process (as described earlier) and makes some attempt to isolate them. More involved schemes, such as Janda's, might also be attempted, particularly if information is being sought from multiple sources. As noted, Janda constructed an adequacy-confidence scale to evaluate data quality (1970). The scale was designed to reflect four factors considered important in determining the researchers' belief in the accuracy of the coded variable values: the number of sources providing relevant information for the coding decision, the proportion of agreement to disagreement in the information reported by different sources, the degree of discrepancy among sources when disagreement exists, and the credibility attached to the various sources of information. An adequacy-confidence value was then assigned to each recorded variable value.

Along with the question of what specific indicators of confidence to use is the question of how to scale them. Orwin and Cordray pilot tested a 5-point scale (1 = low, . . . , 5 = high) for each confidence rating, fashioned

Table 10.6 Agreement Rate by Level of Confidence

	Low	Medium	High
Experimenter affiliation	1.00 (14)	1.00 (37)	.95 (75)
Blinding	.83 (6)	.91 (66)	.93 (44)
Diagnosis	—	1.00 (12)	.99 (114)
Client IQ	1.00 (11)	.18 (74)	1.00 (41)
Client age	1.00 (1)	.68 (50)	.83 (75)
Client source	—	1.00 (10)	.89 (116)
Client assessment	—	.53 (17)	.98 (103)
Therapist assessment	.00 (15)	.67 (18)	.75 (93)
Internal validity	—	.52 (27)	.88 (93)
Treatment mortality	.00 (4)	1.00 (1)	.94 (121)
Comparison mortality	.00 (4)	1.00 (1)	.93 (121)
Comparison type	—	—	1.00 (126)
Control group type	—	.00 (9)	.69 (114)
Experimenter allegiance	.10 (10)	.64 (36)	1.00 (80)
Modality	—	.33 (4)	1.00 (122)
Location	—	—	1.00 (126)
Therapist experience	.55 (55)	.56 (57)	1.00 (11)
Outcome type	.00 (1)	.00 (11)	.92 (114)
Follow-up	.00 (2)	.92 (47)	.83 (75)
Reactivity	.83 (6)	.28 (65)	.94 (47)
Client participation	1.00 (1)	.67 (3)	1.00 (122)
Setting type	—	.40 (15)	.99 (111)
Treatment integrity	.60 (10)	.83 (60)	1.00 (56)
Comparison group contamination	.30 (37)	.32 (50)	1.00 (39)
Outcome Rxx	.13 (150)	.56 (70)	.92 (38)

SOURCE: Orwin 1983.

NOTE: Selected variables from Smith, Glass, and Miller (1980) ($n = 126$). Sample sizes in parentheses.

Table 10.7 Spearman Rank Order Correlations

	r_{RHO}
All variables ($K = 25$)	
Agreement rate	.71
Variables for which kappa was computed ($K = 20$)	
Agreement rate	.71
Kappa	.62
Variables for which intercoder correlation was computed ($K = 15$)	
Agreement rate	.81
Intercoder rate	.67
Variables for which all three estimates were computed ($K = 10$)	
Agreement rate	.73
Kappa	.79
Intercoder correlation	.66

SOURCE: Authors' compilation.

NOTE: Between confidence and interrater agreement for selected variables from Smith, Glass, and Miller (1980).

after the Smith, Glass, and Miller confidence of treatment classification variable (1980). Analysis of the pilot results revealed that five levels of confidence were not being discriminated. Specifically, the choices of 2 versus 3 and 3 versus 4 seemed to be made arbitrarily. The five categories were then collapsed into three for the main study. In addition, each category was labeled with a verbal descriptor to concretize it and minimize coder drift: 3 = certain or almost certain, 2 = more likely than not, 1 = guess (see Kazdin 1977). Discrepancies were resolved through discussion. The 3-point confidence scale was simple yet adequate for its intended purpose—to establish a mechanism for discerning high-quality from lesser-quality information in the conduct of subsequent analyses.

There is a larger question about how confidence ratings should be used in a meta-analysis. One may use such ratings to adjust the score on a scale through a weighting mechanism, or one may include the confidence ratings as moderators. Both approaches have advantages and disadvantages. For the first, the weighting mechanism is arbitrary to some degree. In contrast, including confidence ratings as covariates in the analysis is straightforward, but can easily result in too many variables, which will not be the case for weighting (for further arguments against weighting, specifically of study effects, see chapter 7). Further, a confidence rating system cannot apply to discrete variables. For continuous variables, any weighting scheme will have weights confounded with individual raters' use of the confidence scale. Because of those issues, we recommend that any use of confidence ratings be confined to sensitivity analyses. If changes in findings are substantial as a result, one should still report original analysis, but qualify it by reference to the sensitivity analysis.

10.4.3 Sensitivity Analysis

10.4.3.1 Rationale Sensitivity analysis can assess robustness and bound uncertainty. It has been a part of research synthesis at least since the original Smith and Glass psychotherapy study (1977). Glass's position of including methodologically flawed studies was attacked by his critics as implicitly advocating the abandonment of critical judgment (for example, Eyesenck 1978). He rebutted these and related charges on multiple grounds, but the most enduring was the argument that meta-analysis does not ignore methodological quality, but instead presents a way of determining empirically whether particular methodological threats systematically influence outcomes. Glass's indicators of quality (for example, the 3-point internal validity scale) were crude then and appear even cruder in retrospect, and may not have been at all successful in what they were attempting to do.[16] The principle, however—that of empirically assessing covariance of quality and research findings rather than assuming it a priori—was perfectly sensible and consistent with norms of scientific inquiry. In essence, the question was whether research findings were sensitive to variations in methodological quality. If not, lesser-quality studies can be analyzed along with high-quality studies, with consequent increase in power, generalizability, and so on. The worst-case scenario—that lesser-quality studies produce systematically different results and cannot be used—is no worse than had the synthesist followed the advice of the critics and excluded those studies a priori.

The general issue of sensitivity analysis in research synthesis is discussed in chapter 13. This chapter focuses on applying the logic of sensitivity analysis to the evaluation of coding decisions.

10.4.3.2 Multiple Ratings of Ambiguous Items The sources of error identified earlier (for example, deficient reporting, ambiguities of judgment) are not randomly dispersed across all variables. The synthesist frequently knows at the outset which variables will be problematic. If not, a well-designed pilot test will identify them. For those variables, multiple ratings should be considered. Like multiple measures in other contexts, multiple ratings help guard against biases stemming from a particular way of assessing the phenomenon (monomethod bias) and can set up sensitivity analyses to determine if the choice of rating makes any difference. Mary Smith's synthesis of sex bias is an example (1980). When an investigator reported only their significant effect, Smith entered an effect size of 0 for the remaining effects. Aware of the pitfalls of this approach, she alternately used a different procedure and deleted these cases. Neither changed the overall mean effect size by more than a small fraction. In their synthesis of sex differences, Alice Eagly and Linda Carli took the problem of unreported nonsignificant effects a step further (1981). Noting that the majority of these (fifteen of sixteen) tended in the positive (female) direction, they reasoned that setting effect size at 0 would lead to an underestimation of the true mean effect size, whereas deleting them would lead to overestimation. They therefore did both to be confident of at least bracketing the true value. They also used two indicators of researcher's sex (an important variable in this area), overall percentage of male authors, and sex of first author. These were found to differ only slightly in their correlations with outcome.

Each of these examples illustrates sensitivity analysis directed at specific rival hypotheses. It should be evident that many opportunities exist for thoughtful sensitivity analyses in the evaluation of coding decisions and that these do not require specialized technical expertise; conscientiousness and common sense will frequently suffice.

10.4.3.3 Multiple Measures of Interrater Agreement As described earlier, the choice of IRR index can have a significant effect on the reliability estimate obtained. Although the guidelines presented earlier should narrow the range of choices, they may not uniquely identify the best one. Computing multiple indices is therefore warranted.

In their 1985 reanalysis of the 1980 Smith, Glass, and Miller psychotherapy data, Orwin and Cordray computed multiple estimates of IRR for each variable. For continuous variables, *AR* and *r* were computed (r_I was not computed, but could have been). For ordinal categorical variables, *AR*, *r*, and K_w were computed. For nominal categorical variables, *AR* and *K* were computed; when nominal variables were dichotomous, *r* (in the form of phi) was also computed. Four regressions were then run, each using a different set of reliability estimates. The first used the highest estimate available for each variable. Its purpose was to provide a lower bound on the amount of change produced by disattenuation. The second and third runs were successively more liberal (for details, see Orwin and Cordray 1985). A final run, intended as the most liberal credible analysis, disattenuated the criterion variable (effect size) as well as the predictors. The reliability estimates for the four runs are shown in table 10.8.

10.4.3.4 Isolating Questionable Variables Both IRRs and confidence ratings are useful tools for flagging variables that may be inappropriate for use as meta-analytic predictors of effect size. In the Orwin and Cordray study, for example, the therapist experience variable was coded with high confidence in only 9 percent of the studies, was "guessed" in 45 percent, and had an IRR (using Pearson *r*) of 0.56 (1985). Such numbers would clearly suggest the exercise of caution in further use of that variable. Conducting analyses with and without it, and comparing the results, would be a logical first step.

In regard to both cases and variables, the finding of significant differences between inclusion and exclusion does

Table 10.8 Reliability Estimates

Variable	Run 1	Run 2	Run 3	Run 4
Diagnosis: neurotic, phobic, or depressive	.98	.98	.89	.89
Diagnosis: delinquent, felon, or habituée	1.00	1.00	1.00	1.00
Diagnosis: psychotic	1.00	1.00	1.00	1.00
Clients self-presented	.97	.57	.71	.71
Clients solicited	.93	.86	.81	.81
Individual therapy	1.00	1.00	.85	.85
Group therapy	.98	.96	.94	.94
Client IQ	.69	.69	.60	.60
Client age[a]	.99	.99	.91	.91
Therapist experience × neurotic diagnosis	.76	.75	.70	.70
Therapist experience × delinquent diagnosis	1.00	1.00	1.00	1.00
Internal validity	.76	.71	.42	.42
Follow-up time[b]	.99	.99	.95	.95
Outcome type[c]	.87	.70	.76	.76
Reactivity[d]	.57	.56	.57	.57
ES	1.00	1.00	1.00	.78

SOURCE: Orwin and Cordray 1985.

NOTES: Reliability-corrected regression runs on the Smith, Glass, and Miller (1980) psychotherapy data.

[a]Transformed age = $(\text{age} - 25)(|\text{age} - 25|)^{1/2}$.
[b]Transformed follow-up = $(\text{follow-up})^{1/2}$.
[c]"Other" category removed for purpose of dichotomization.
[d]Transformed reactivity = $(\text{reactivity})^{2.25}$.

not automatically argue for outright exclusion; rather, it alerts the analyst that mindless inclusion is not warranted. As in primary research, the common practices of dropping cases, dropping variables, and working from a missing-data correlation matrix result in loss of information, loss of statistical power and precision, and biased estimates when, as is frequently the case, the occurrence of missing data is nonrandom (Enders 2010). Therefore, more sophisticated approaches are preferable when feasible (for more on the treatment of missing data in research synthesis, see chapter 17).

10.5 EXAMPLES OF VARIOUS ANALYSES, WITH CODE

This section illustrates how to conduct various reliability calculations using, where possible, the statistical software R (R Core Team 2017). R is an extendable statistics program for which various user-supplied packages have been implemented. Many of the interrater reliability statistics are available in the package "irr" by Matthias Gamer and his colleagues (2012). The two exceptions are Andrés and Marzo's Delta, which requires special software, and the mixed-model computation of the intraclass correlation, which uses the R package "lme4" package (Bates et al. 2015). The R Core Team's website supplies an introduction to the basics, and this section assumes basic familiarity with the program such as can be obtained from that source. In order to run the examples, the user will need to install the irr package, which is easily accomplished using a pull-down menu in R. The illustrations continue to use the data from table 10.1 with either two or three coders depending on the example.

10.5.1 Categorical Variables

Although various methods exist in R for reading data from files, for the sake of simplicity the data here are entered from the keyboard by creating a vector of data for each rater and binding the columns of the table together. We create objects that we named TwoCategoricalRaters and ThreeCategoricalRaters, for which TwoCategoricalRaters is the first two columns of the table, and ThreeCategorical Raters is the complete table.

```
> Rater1 <- c(3,3,2,3,1,3,2,1,2,2,2,3,3,2,
1,1,3,2,2,3,2,1,3,3,2)
> Rater2 <- c(2,1,2,2,1,1,2,1,2,1,2,3,1,1,
1,1,3,2,2,1,1,1,2,3,2)
> Rater3 <- c(3,1,2,3,1,3,1,1,1,3,2,3,2,1,
1,2,1,2,2,1,2,3,2,3,3)
> TwoCategoricalRaters
<- cbind(Rater1,Rater1)
> TwoCategoricalRaters
      Rater1 Rater2
[1,]       3      2
[2,]       3      1
[3,]       2      2
 ⋮         ⋮      ⋮
[23,]      3      2
[24,]      3      3
[25,]      2      2
>
> ThreeCategoricalRaters
<- cbind(Rater1,Rater2,Rater3)
> ThreeCategoricalRaters
      Rater1 Rater2 Rater3
[1,]       3      2      3
[2,]       3      1      1
[3,]       2      2      2
⋮          ⋮      ⋮      ⋮
[23,]      3      2      2
[24,]      3      3      3
[25,]      2      2      3
```

10.5.1.1 Agreement Rate The "agree" function in the irr package calculates percentage of agreement. It can be applied to any number of raters. First, we observe the agreement rate for the two real raters.

```
> library(irr)
Loading required package: lpSolve
> agree(TwoCategoricalRaters)
Percentage agreement (Tolerance=0)

 Subjects = 25
  Raters = 2
%-agree = 60
```

Note that the agreement rate is 60 percent. The agreement rate for the three-rater table is somewhat lower, in part because it represents the percentage of cases for which all three raters agree, a more stringent hurdle.

```
> agree(ThreeCategoricalRaters)
Percentage agreement (Tolerance=0)

 Subjects = 25
  Raters = 3
%-agree = 36
```

10.5.1.2 Cohen's Kappa and Weighted Kappa Cohen's kappa for two raters is available in the irr package through the kappa2 function. This calculates the unweighted form of kappa.

```
> library(irr)
Loading required package: lpSolve
> kappa2(TwoCategoricalRaters)
Cohen's Kappa for 2 Raters (Weights:
unweighted)

 Subjects = 25
   Raters = 2
    Kappa = 0.425

        z = 3.51
  p-value = 0.000455
```

Note that kappa is 0.425. If one attempts to apply kappa2 to more than two raters, the function returns an informative error message.

```
> kappa2(ThreeCategoricalRaters)
Error in kappa2(ThreeCategoricalRaters) :
  Number of raters exceeds 2. Try kappam.
fleiss or kappam.light.
```

As the message suggests, Fleiss's kappa for more than two raters is available:

```
> kappam.fleiss(ThreeCategoricalRaters)
Fleiss' Kappa for m Raters

 Subjects = 25
   Raters = 3
    Kappa = 0.315

        z = 3.85
  p-value = 0.00012
```

Similar to agreement rate, coding reliability is lower for the three-rater data. It is also possible to calculate Light's kappa, which is the average of all possible two-rater kappas:

```
> kappam.light(ThreeCategoricalRaters)
Light's Kappa for m Raters

Subjects = 25
Raters = 3
Kappa = 0.328

z = 5.86
p-value = 4.74e-09
```

The kappa2 function can also calculate weighted kappa by specifying a weight argument. The most likely value for weight is *equal* but weights based on squared distance are also available with the value *squared*.

```
> kappa2(TwoCategoricalRaters,"equal")
Cohen's Kappa for 2 Raters (Weights:
equal)

 Subjects = 25
   Raters = 2
    Kappa = 0.386

        z = 3.22
  p-value = 0.00126
```

10.5.1.3 Andrés and Marzo's Delta Andrés and Marzo's (2004) delta is not implemented in any available R package. A program executable on Windows computers is available online (see http://www.ugr.es/~bioest/Delta.exe). The graphical interface involves a number of screens and is not amenable to presentation in text. For the two-rater data set, delta is 0.410 with a standard error of 0.131.

10.5.1.4 Krippendorff's Alpha The irr package implements Krippendorff's alpha. However, the package expects a matrix in which rows represent raters, not columns, so that the data matrix must be transposed using R's t function. The calculation of Krippendorff's alpha depends on the level of measurement of the variable being coded. The default is nominal, which is not appropriate here, but we calculate it to illustrate the syntax before we change to ordinal:

```
> library(irr)
Loading required package: lpSolve
> kripp.alpha(t(TwoCategoricalRaters))
 Krippendorff's alpha

 Subjects = 25
   Raters = 2
    alpha = 0.403

> kripp.alpha(t(TwoCategoricalRaters),
method="ordinal")
Krippendorff's alpha

 Subjects = 25
   Raters = 2
    alpha = 0.278
```

Note that the function can be applied seamlessly to multiple-rater problems:

```
> kripp.alpha(t(ThreeCategoricalRaters))
Krippendorff's alpha

 Subjects = 25
   Raters = 3
    alpha = 0.32

> kripp.alpha(t(ThreeCategoricalRaters),
method="ordinal")
Krippendorff's alpha

 Subjects = 25
   Raters = 3
    alpha = 0.256
```

10.5.2 Continuous Variables

As previously, we use the same ratings from table 10.1 but treat them as continuous variables in this section. We create objects containing the first two columns and all three columns to illustrate analyses with two and more than two raters:

```
> TwoContinuousRaters
      Rater1 Rater2
[1,]       3      2
[2,]       3      1
[3,]       2      2
 ⋮         ⋮      ⋮
[23,]      3      2
[24,]      3      3
[25,]      2      2
> ThreeContinuousRaters <- cbind(Rater1,
Rater2,Rater3)

> ThreeContinuousRaters
      Rater1 Rater2 Rater3
[1,]       3      2      3
[2,]       3      1      1
[3,]       2      2      2
 ⋮         ⋮      ⋮      ⋮
[23,]      3      2      2
[24,]      3      3      3
[25,]      2      2      3
```

10.5.2.1 Intercoder Correlation The R program contains a built-in function, cor, that calculates the Pearson product-moment correlation:

```
> cor(Rater1,Rater2)
[1] 0.4520228

> cor(TwoContinuousRaters)
          Rater1    Rater2
Rater1 1.0000000 0.4520228
Rater2 0.4520228 1.0000000
```

With the irr package, it is possible to calculate the mean correlation for more than two raters:

```
> library(irr)
Loading required package: lpSolve
> meancor(ThreeContinuousRaters)
 Mean of bivariate correlations R

 Subjects = 25
   Raters = 3
        R = 0.331

        z = 1.55
  p-value = 0.121
```

10.5.2.2 Krippendorff's Alpha Krippendorff's alpha, applied to categorical data with the default nominal level of measurement in section 10.5.1.4, may also be applied to continuous data by specifying a higher level of measurement. In the current example, the variable now represents a ratio level measurement. Once again, the function for Krippendorff's alpha expects the transposition of the data matrix.

```
> library(irr)
Loading required package: lpSolve
> kripp.alpha(t(TwoContinuousRaters),
method="ratio")
 Krippendorff's alpha

 Subjects = 25
   Raters = 2
    alpha = 0.311

> kripp.alpha(t(ThreeContinuousRaters),
method="ratio")
 Krippendorff's alpha

 Subjects = 25
   Raters = 3
    alpha = 0.268
```

10.5.2.3 Intraclass Correlation The irr package also provides a function icc to calculate intraclass correlations. The function can take various arguments to specify different modes of icc calculation. For the present example (and for most applications in the area of coding reliability), the default one-way option is correct.

```
> library(irr)
Loading required package: lpSolve
> icc(TwoContinuousRaters)
Single Score Intraclass Correlation

 Model: oneway
 Type : consistency

 Subjects = 25
   Raters = 2
   ICC(1) = 0.278

 F-Test, H0: r0 = 0 ; H1: r0 > 0
  F(24,25) = 1.77 , p = 0.0818

 95%-Confidence Interval for ICC Population
Values:
 -0.118 < ICC < 0.599
> icc(ThreeContinuousRaters)
Single Score Intraclass Correlation

  Model: oneway
  Type : consistency

  Subjects = 25
    Raters = 3
    ICC(1) = 0.265

 F-Test, H0: r0 = 0 ; H1: r0 > 0
  F(24,50) = 2.08 , p = 0.0144

 95%-Confidence Interval for ICC Population
Values:
0.026 < ICC < 0.53
```

An alternative to the ANOVA-based estimation from the irr package is to use the lme4 package, which can calculate variances for an intraclass correlation through a mixed model. This has the advantage of being able to handle codings with missing values or unbalanced designs.

```
> library(lme4)
>
> Rater1NA <- c(3,3,2,3,NA,3,2,1,2,2,NA,3,
3,2,1,1,3,2,2,3,2,1,3,NA,2)
> Rater2NA <- c(2,1,NA,2,1,1,2,1,2,1,2,3,
1,1,NA,1,3,2,2,1,1,1,2,3,2)
> Rater3NA <- c(3,1,2,3,1,3,1,1,NA,3,2,3,
2,1,1,2,1,2,2,1,2,3,2,3,3)
```

Instead of combining the ratings into a matrix, this package expects a data frame in which one column contains all ratings and values in adjacent columns represent the rater and study corresponding to each value.

```
> Ratings <- c(Rater1NA,Rater2NA,Rater3NA)
> Study <- c(1:25,1:25,1:25)
> Rater <- c(rep(1,25),rep(2,25),rep(3,25))
> RaterData <- data.frame(Ratings,Study,
Rater)
> ICCInput <- RaterData[complete.cases
(RaterData),]
> ICCInput
   Ratings Study Rater
1        3     1     1
2        3     2     1
3        2     3     1
4        3     4     1
⋮        ⋮     ⋮     ⋮
71       2    21     3
72       3    22     3
73       2    23     3
74       3    24     3
75       3    25     3
> ICCOutput <- lmer(data = ICCInput,
formula = "Ratings ~ (1 | Study)")
```

Once the package estimates the variance components, we can calculate the ICC manually.

```
> StudyVariance <- data.frame(VarCorr
(ICCOutput))$vcov[1]
> ErrorVariance <- data.frame(VarCorr
(ICCOutput))$vcov[2]
> ICCValue <- StudyVariance/(StudyVariance+
ErrorVariance)
> ICCValue
[1] 0.1794355
```

10.6. SUGGESTIONS FOR FURTHER RESEARCH

10.6.1 Investigating the Antecedents of Coder Reliability

Georg Matt points out that synthesists need to become more aware, and take deliberate account, of the multiple processes associated with uncertainty in reading, understanding, and interpreting research reports (1989). But in keeping with the scientific endeavor principle, they also need to begin to move beyond observation of these phenomena and toward explanation. For example, further research on evaluating coding decisions in meta-analysis should look beyond how to assess IRR toward better classification and understanding of why coders disagree. Evidence from other fields could probably inform some

initial hypotheses. Horwitz and Yu conducted detailed analyses of errors in the extraction of epidemiological data from patient records and suggested a rough taxonomy of six error categories (1984). Of the six, four were errors in the actual data extraction, and two were errors in the interpretation of the extracted data. The two interpretation errors were incorrect interpretation of criteria and correct interpretation of criteria, but inconsistent application. Note the similarity between these types of errors and the types of judgment errors described earlier in regard to coding in research synthesis.

When the coder must make judgment calls, the rationale for each judgment should be documented. In this way, we can begin to formulate a general theory of the coder judgment process. For complex yet critical variables like effect size, interim codes should be recorded. One code might simply indicate what was reported in the original study. A second code might indicate what decision rule was invoked in further interpretation of the data point. A third code—representing the endpoint and the variable of interest to the synthesis (effect size, for example)—would be the number that results from the coder's application of that interpretation. Coder disagreement on the third code would no longer be a black box, but could be readily traced to its source. Coder disagreement could then be broken down to its component parts (that is, the errors could be partitioned by source), facilitating diagnosis much like a failure analysis in engineering or computer programming. The detailed documentation of coding decision rules could be made for examination and reanalysis by other researchers (much as the list of included studies is currently), including any estimation procedures used for missing or incomplete data (for an example, see Bullock and Svyantek 1985). An understanding of the taxonomy of processes associated with coding errors could, in principle, identify variables to be experimentally manipulated. Theoretical models would then be experimentally investigated, and non-experimental data subjected to modern causal models, leading to informed strategies for reducing and controlling error rates.

10.6.2 Assessing Time and Necessity

The coding stage of research synthesis is time consuming and tedious, and in many instances additional efforts to evaluate coding decisions further lengthen and complicate it. The time and resources of practicing synthesists is limited, and they would likely want answers to two questions before selecting a given strategy: How much time will it take? How necessary is it?

The time issue has not been systematically studied. Orwin and Cordray included a small time-of-task side study and concluded that the marginal cost in time of augmenting the task as they did (double codings plus confidence ratings) was quite small given the number of additions involved and the distinct impact they had on the analytic results (1985). In part, this is because though double coding a study approximately doubles the time required to code it, in an actual synthesis only a sample of studies may be double coded. In a synthesis the size of Smith, Glass, and Miller's, the double coding of twenty-five representative studies would increase the total number of studies coded by only 5 percent. When a particular approach to reliability assessment leads to low occurrence rates or restricted range on key variables oversampling techniques can be used. Research that estimated how often oversampling was required would in itself be useful. Other possibilities are easy to envision. In short, this is an area of considerable practical importance that has not been seriously examined to date.

The question of necessity could be studied the same way that methodological artifacts of primary research have been studied. For example, a handful of syntheses have used coder masking techniques, as described earlier. Keeping coders unaware of hypotheses or other information is an additional complication that adds to the cost of the effort and may have unintended side effects. An empirical test of the effect of selective masking on research synthesis coding would be a useful contribution, as such studies have been in primary research, to determine whether the particular artifact that masking is intended to guard against (coder bias) is real or imaginary, large or trivial, and so forth. As was the case with pretest sensitization and Hawthorne effects in the evaluation literature (Shadish, Cook, and Campbell 2002), research may show that some suspected artifacts in synthesis coding are only that—perfectly plausible yet not much in evidence when subjected to empirical scrutiny. In reality, few artifacts will be universally present or absent, but rather will interact with the topic area (for example, coder gender is more likely to matter when the topic is sex differences than when it is acid rain) or with other factors. Research could shed more light on these interactions as well. In sum, additional efforts to assess both the time and need for specific methods of evaluating coding decisions could significantly help the practicing synthesist make cost-effective choices in the synthesis design.

10.7 NOTES

1. Significant reporting deficiencies, and their deleterious effect on research synthesis, are not confined to social research areas (see, for example, synthesis work on the effectiveness of coronary artery bypass surgery in Wortman and Yeaton 1985).
2. For example, Laurel Oliver reported on a colleague who attempted to apply a meta-analytic approach to research on senior leadership, but gave up after finding that only 10 percent of the eligible studies contained enough data to sustain the synthesis (1987).
3. That detailed criteria for decisions fail to eliminate coder disagreement should not be surprising, even if all known contingencies are incorporated (for an analog from coding patient records in epidemiology, see Horwitz and Yu 1984).
4. Low-inference variables are those for which coding is possible based on simple observation with little or no need for subjective judgment.
5. Specifically, this rule aims to exclude effect sizes based on redundant outcome measures. A measure was judged to be redundant if it matched another in outcome type, reactivity, follow-up time, *and* magnitude of effect.
6. The Chalmers review was actually a review of replicate meta-analyses rather than a meta-analysis per se (that is, the individual studies were meta-analyses), but the principle is equally applicable to meta-analysis proper (1987).
7. Here, and elsewhere in the chapter, we rely on meta-analytic work in the most recent complete year of *Psychological Bulletin* as a barometer of current practice.
8. Such improvement is quite plausible. For example, Orwin finds that measures of reporting quality were positively correlated with publication date in the 1980 Smith, Glass, and Miller psychotherapy data (1983).
9. There is nothing anomalous about the present coders' relative lack of agreement on internal validity; lack of consensus on ratings of research quality is commonplace (for an example within meta-analysis, see Stock et al. 1982).
10. This desirable characteristic has led some writers (for example, Hartmann 1977) to advocate extending the use of r, in the form of the phi coefficient (ϕ), to reliability estimation of dichotomous items: $\phi = (BC - AD)/((A + \mathrm{B})(C + D)(A + \mathrm{C})(B + D))^{1/2}$, where A, B, C, and D represent the frequencies in the first through fourth quadrants of the resulting 2×2 table.
11. This phenomenon is easily demonstrated with the illustrative data; the details are left to the reader as an exercise.
12. The *general* forms of the equations are

$$r_I\,(\text{design 1}) = \frac{BMS - WMS}{BMS + (k-1)WMS},$$

$$r_I\,(\text{design 2}) = \frac{BMS - EMS}{BMS + (k-1)EMS + k(CMS - EMS)/n},$$

$$r_I\,(\text{design 3}) = \frac{BMS - EMS}{BMS + (k-1)EMS},$$

    where k is the number of coders rating each study and n is the number of studies.
13. r_1 is closely related to the coefficient of generalizability and index of dependability used in G theory (Shavelson and Webb 1991).
14. Training and piloting were completed before the rater drift assessment began, to avoid picking up obvious training effects.
15. The term *case* is used rather than *study* because a study can have multiple cases (one for each effect size). The term *observation* refers to the value assigned to a particular variable within a case.
16. Chapter 10 shows how the state of the art has evolved (see also U.S. General Accounting Office 1989, appendix II).

10.8 REFERENCES

American Psychiatric Association, Committee on Nomenclature and Statistics. 1980. *Diagnostic and Statistical Manual of Mental Disorders*, 3rd ed. Washington, D.C.: American Psychiatric Association.

American Psychological Association. 2010. *Publication Manual of the American Psychological Association.* Washington, D.C.: American Psychological Association.

Andrés, A. Martin, and P. Fernina Marzo. 2004. "Delta: A New Measure of Agreement Between Two Raters." *British Journal of Mathematical and Statistical Psychology* 57(1): 1–19.

Bates, Douglas, Martin Maechler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67(1): 1–48.

Begg, Colin, Mildred Cho, Susan Eastwood, Richard Horton, David Moher, Ingram Olkin, Roy Pitkin, Drummond Rennie, Kenneth F. Schulz, David Simel, and Donna F. Stroup. 1996. "Improving the Quality of Reporting of Randomised Controlled Trials. The CONSORT Statement." *Journal of the American Medical Association* 276(8): 637–39.

Bullock, R. J., and Daniel J. Svyantek. 1985. "Analyzing Meta-Analysis: Potential Problems, an Unsuccessful Replication, and Evaluation Criteria." *Journal of Applied Psychology* 70(1): 108–15.

Buros, Oscar K. 1978. *Mental Measurements Yearbook.* Highland Park, N.J.: Gryphon Press.

Burton, Nancy W. 1981. "Estimating Scorer Agreement for Nominal Categorization Systems." *Educational and Psychological Measurement* 41(4): 953–61.

Carey, Gregory, and Irving I. Gottesman. 1978. "Reliability and Validity in Binary Ratings: Areas of Common Misunderstanding in Diagnosis and Symptom Ratings." *Archives of General Psychiatry* 35(12): 1454–59.

Chalmers, Thomas C., Jayne Berrier, Henry S. Sacks, Howard Levin, Dinah Reitman, and Raguraman Nagalingham. 1987. "Meta-Analysis of Clinical Trials as a Scientific Discipline, II. Replicate Variability and Comparison of Studies that Agree and Disagree." *Statistics in Medicine* 6(7): 733–44.

Cicchetti, Dominic V. 1985. "A Critique of Whitehurst's 'Interrater Agreement for Journal Manuscript Reviews': De Omnibus Disputandem Est." *American Psychologist* 49(5): 563–68.

Cicchetti, Dominic V., and Sara S. Sparrow. 1981. "Developing Criteria for Establishing the Interrater Reliability of Specific Items in a Given Inventory." *American Journal of Mental Deficiency* 86(2): 127–37.

Cohen, Jacob. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20(1): 37–46.

———. 1968. "Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit." *Psychological Bulletin* 70(4): 213–20.

Cordray, David S., and Robert G. Orwin. 1981. "Technical Evidence Necessary for Quantitative Integration of Research and Evaluation." Paper presented at the joint conference of the International Association of Social Science Information Services and Technology and the International Federation of Data Organizations, Grenoble, France (September 1981).

Cordray, David S., and L. Joseph Sonnefeld. 1985. "Quantitative Synthesis: An Actuarial Base for Planning Impact Evaluations." In *Utilizing Prior Research in Evaluation Planning*, edited by Davis S. Cordray. New Directions for Program Evaluation No. 27. San Francisco: Jossey-Bass.

Cronbach, Lee. J., Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam. 1972. *Dependability of Behavioral Measurements.* New York: John Wiley & Sons.

Donner, Allan, and John J. Koval. 1980. "The Estimation of Intraclass Correlation in the Analysis of Family Data." *Biometrics* 36(1): 19–25.

Eagly, Alice H., and Linda L. Carli. 1981. "Sex of Researchers and Sex-Typed Communications as Determinants of Sex Differences in Influenceability: A Meta-Analysis of Social Influence Studies. *Psychological Bulletin* 90(1): 1–20.

Ebel, Robert L. 1951. "Estimation of the Reliability of Ratings." *Psychometrika* 16(4): 407–24.

Enders, Craig K. 2010. *Applied Missing Data Analysis*. New York: Guilford Press.

Eyesenck, Hans J. 1978. "An Exercise in Mega-Silliness." *American Psychologist* 33(5): 517.

Fleiss, Joseph L. 1971. "Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin* 76(5): 378–82.

———. 1981. *Statistical Methods for Rates and Proportions*, 2nd ed. New York: John Wiley and Sons.

Fleiss, Joseph L., and Jacob Cohen. 1973. "The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability." *Educational and Psychological Measurement* 33(3): 613–19.

Fleiss, Joseph L., Jacob Cohen, J., and B. S. Everett. 1969. "Large Sample Standard Errors of Kappa and Weighted Kappa." *Psychological Bulletin* 72(5): 323–27.

Gamer, Matthias, Jim Lemon, Ian Fellows, and Puspendra Singh. 2012. "irr: Various Coefficients of Interrater Reliability and Agreement. R Package Version 0.84." Accessed December 4, 2018. https://CRAN.R-project.org/package=irr.

Glass, Gene V., and Mary L. Smith. 1979. "Meta-Analysis of Research on the Relationship of Class Size and Achievement." *Educational Evaluation and Policy Analysis* 1(1): 2–16.

Green, Bert F., and Judith A. Hall. 1984. "Quantitative Methods for Literature Reviews." *Annual Review of Psychology* 35(1): 37–53.

Grove, William M., Nancy C. Andreasen, Patricia McDonald-Scott, Martin B. Keller, and Robert W. Shapiro. 1981. "Reliability Studies of Psychiatric Diagnosis: Theory and Practice." *Archives of General Psychiatry* 38(4): 408–13.

Hartmann, Donald P. 1977. "Considerations in the Choice of Interobserver Reliability Estimates." *Journal of Applied Behavior Analysis* 10(1): 103–16.

Hays, William L. 1994. *Statistics,* 5th ed. Fort Worth, Tex.: Harcourt Brace College.

Hedges, Larry V., and Ingram Olkin. 1985. *Statistical Methods for Meta-Analysis.* Orlando, Fl.: Academic Press.

Higgins Julian P. T., and Sally Green, eds. 2011. "Cochrane Handbook for Systematic Reviews of Interventions, Version 5.1.0." The Cochrane Collaboration. Accessed December 4, 2018. http://www.handbook.cochrane.org.

Horwitz, Ralph I., and Eunice C. Yu. 1984. "Assessing the Reliability of Epidemiological Data Obtained from Medical Records. *Journal of Chronic Disease,* 37(11): 825–831.

Hyde, Janet S. 1981. "How Large Are Cognitive Gender Differences? A Meta-Analysis Using W and D." *American Psychologist* 36(8): 892–901.

Jackson, Gregg B. 1980. "Methods for Integrative Reviews." *Review of Educational Research,* 50(3): 438–60.

Janda, Kenneth. 1970. "Data Quality Control and Library Research on Political Parties." In *A Handbook of Method in Cultural Anthropology*, edited by Raoul Narall and Ronald Cohen. New York: Doubleday.

Janes, Cynthia L. 1979. "Agreement Measurement and the Judgment Process." *Journal of Nervous and Mental Disorders* 167(6): 343–47.

Jones, Allan P., Lee A. Johnson, Mark C. Butler, and Deborah S. Main. 1983. "Apples and Oranges: An Empirical Comparison of Commonly Used Indices of Interrater Agreement." *Academy of Management Journal* 26(3): 507–19.

Journal Article Reporting Standards Working Group. 2007. "Reporting Standards for Research in Psychology: Why do WE Need Them? What Might They Be?" Washington, D.C.: American Psychological Association.

Kazdin, Alan E. 1977. "Artifacts, Bias, and Complexity of Assessment: The ABC's of Research." *Journal* of *Applied Behavior Analysis* 10(1): 141–50.

Kerlinger, Fred N., and Elazar E. Pedhazur. 1973. *Multiple Regression in Behavioral Research.* New York: Holt, Rinehart & Winston.

Krippendorff, Klaus. 2011. "Computing Krippendorff's Alpha-Reliability." Working paper, University of Pennsylvania. Accessed December 4, 2018. http://repository.upenn.edu/asc_papers/43.

Kulik, James A., Chen-lin C. Kulik, and Peter A. Cohen. 1979. "Meta-Analysis of Outcome Studies of Keller's Personalized System of Instruction." *American Psychologist* 34(4): 307–18.

Light, Richard J. 1971. "Measures of Response Agreement for Qualitative Data: Some Generalizations and Alternatives." *Psychological Bulletin* 76(5): 365–77.

Light, Richard J., and David B. Pillemer. 1984. *Summing Up: The Science of Reviewing Research.* Cambridge, Mass.: Harvard University Press.

Lord, Frederic M., and Melvin R. Novick. 1968. *Statistical theories of Mental Test Scores.* Reading, Mass.: Addison-Wesley.

Marcoulides, George A. 1990. "An Alternative Method for Estimating Variance Components in Generalizability Theory." *Psychological Reports* 66(2): 379–86.

Mares, Marie-Louise, and Zhongdang Pan. 2013. "Effects of *Sesame Street*: A Meta-Analysis of Children's Learning in 15 Countries." *Journal of Applied Developmental Psychology* 34(1): 140–51.

Matt, Georg E. 1989. "Decision Rules for Selecting Effect Sizes in Meta-Analysis: A Review and Reanalysis of Psychotherapy Outcome Studies." *Psychological Bulletin* 105(1): 106–15.

McGraw, Kenneth, and S. P. Wong. 1996. "Forming Inferences About Some Intraclass Correlation Coefficients." *Psychological Methods* 1(1): 30–46.

McGuire, Joanmarie, Gary W. Bates, Beverly J. Dretzke, Julia E. McGivern, Karen L. Rembold, Daniel R. Seobold, Betty Ann M. Turpin, and Joel R. Levin. 1985. "Methodological Quality as a Component of Meta-Analysis." *Educational Psychologist* 20(1): 1–5.

McLellan, A. Thomas, Lester Luborsky, John Cacciola, Jason Griffith, Peggy McGahan, and Charles O'Brien. 1988. *Guide to the Addiction Severity Index: Background, Administration, and Field Testing Results.* Philadelphia, Pa.: Veterans Administration Medical Center.

Moher, David, Alessandro Liberati, Julie Tetzlaff, Douglas G. Altman, and the PRISMA Group. 2009. "*Preferred Reporting Items for Systematic Reviews and Meta-Analyses:* The PRISMA Statement." *Annals of Internal Medicine* 151(4): 264–69.

Moher, David, Kenneth F. Schulz, Douglas G. Altman, and CONSORT Group 2001. "The CONSORT Statement: Revised Recommendations for Improving the Quality of Reports of Parallel-Group Randomised Trials." *Lancet* 357(9263): 1191–94.

Nunnally, Jum C. 1978. *Psychometric Theory*, 2nd ed. New York: McGraw-Hill.

Oliver, Laurel W. 1987. "Research Integration for Psychologists: An Overview of Approaches." *Journal of Applied Social Psychology.* 17(10): 860–74.

Orwin, Robert G. 1983. "The Influence of Reporting Quality in Primary Studies on Meta-Analytic Outcomes: A Conceptual Framework and Reanalysis." PhD diss., Northwestern University.

———. 1985. "Obstacles to Using Prior Research and Evaluations." In *Utilizing Prior Research in Evaluation Planning*, edited by David S. Cordray. New Directions for Program Evaluation no. 27. San Francisco: Jossey-Bass.

Orwin, Robert G., and David S. Cordray. 1985. "Effects of Deficient Reporting on Meta-Analysis: A Conceptual Framework and Reanalysis." *Psychological Bulletin* 97(1): 134–47.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing.

Sacks, Harold S., Jayne Berrier, Dinah Reitman, Vivette A. Ancona-Berk, and Thomas C. Chalmers. 1987. "Meta-Analyses of Randomized Controlled Trials." *New England Journal of Medicine* 316(8): 450–55.

Schafer, Joseph L., and John W. Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7(2): 147–77.

Scott, William A. 1955. "Reliability of Content Analysis: The Case of Nominal Scale Coding." *Public Opinion Quarterly* 19(3): 321–25.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs.* New York: Houghton Mifflin.

Shapiro, David A., and Diana Shapiro. 1982. Meta-Analysis of Comparative Therapy Outcome Studies: A Replication and Refinement. *Psychological Bulletin* 92(3): 581–604.

Shavelson, Richard J., and Noreen M. Webb. 1991. *Generalizability Theory: A Primer.* Newbury Park, Calif.: Sage Publications.

Shrout, Patrick E., and Joseph L. Fleiss. 1979. "Intraclass Correlations: Uses in Assessing Rater Reliability." *Psychological Bulletin* 86(2): 420–28.

Shrout, Patrick E., Robert L. Spitzer, and Joseph L. Fleiss. 1987. "Quantification of Agreement in Psychiatric Diagnosis Revisited." *Archives of General Psychiatry* 44(2): 172–77.

Smith, Mary L. 1980. "Sex Bias in Counseling and Psychotherapy." *Psychological Bulletin* 87(2): 392–407.

Smith, Mary L., and Gene V. Glass. 1977. "Meta-Analysis of Psychotherapy Outcome Studies." *American Psychologist* 32(9): 752–60.

Smith, Mary L., Gene V. Glass, and Thomas I. Miller. 1980. *The Benefits of Psychotherapy.* Baltimore, Md.: Johns Hopkins University Press.

Spitznagel, Edward L., and John E. Helzer. 1985. "A Proposed Solution to the Base Rate Problem in the Kappa Statistic." *Archives of General Psychiatry* 42(7): 725–28.

Stanley, Julian C. 1971. "Reliability." In *Educational Measurement*, 2nd ed., edited by Robert L. Thorndike. Washington, D.C.: American Council on Education.

Steering Group of the Campbell Collaboration. 2015. *Campbell Collaboration Systematic Reviews: Policies and Guidelines. Supplement 1.* Oslo: Campbell Collaboration. Accessed November 24, 2018. https://campbellcollaboration.org/library/campbell-collaboration-systematic-reviews-policies-and-guidelines.html.

Stock, William A., Morris A. Okun, Marilyn J. Haring, Wendy Miller, Wendy, Clifford Kenney, and Robert C. Ceurvost. 1982. "Rigor in Data Synthesis: A Case Study of Reliability in Meta-Analysis." *Educational Researcher* 11(6): 10–20.

Terpstra, David E. 1981. "Relationship Between Methodological Rigor and Reported Outcomes in Organization Development Evaluation Research." *Journal of Applied Psychology* 66(5): 541–43.

Turner, Lucy, Larissa Shamseer, Douglas G. Altman, Kenneth F. Schulz, and David Moher. 2012. "Does Use of the CONSORT Statement Impact the Completeness of Reporting of Randomised Controlled Trials Published in Medical Journals? A Cochrane Review." *Systematic Reviews* 1 (November): 60. DOI: 10.1186/2046-4053-1-60.

U.S. General Accounting Office. 1989. *Prospective Evaluation Methods: The Prospective Evaluation Synthesis.* GAO/PEMD Transfer Paper no. 10.1.10. Washington: Government Printing Office.

Valentine, Jeffrey C., and Harris Cooper. 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(2): 130–49.

Vukasović, Tena, and Denis Bratko. 2015. "Heritability of Personality: A Meta-Analysis of Behavior Genetic Studies." *Psychological Bulletin*141(4): 769–85.

Whitehurst, Grover J. 1984. "Interrater Agreement for Journal Manuscript Reviews." *American Psychologist* 39(1): 22–28.

Wortman, Paul M., and Fred B. Bryant. 1985. "School Desegregation and Black Achievement: An Integrative Review." *Sociological Methods and Research* 13(3): 289–324.

Wortman, Paul M., and William H. Yeaton. 1985. "Cumulating Quality of Life Results in Controlled Trials of Coronary Artery Bypass Graft Surgery." *Controlled Clinical Trials* 6(4): 289–305.

Yeaton, William H., and Paul M. Wortman. 1993. "On the Reliability of Meta-Analytic Reviews: The Role of Intercoder Agreement." *Evaluation Review* 17(3): 292–309.

Zwick, Rebecca. 1988. "Another Look at Interrater Agreement." *Psychological Bulletin* 103(3): 374–78.

PART V

STATISTICALLY DESCRIBING AND COMBINING STUDY OUTCOMES

11

EFFECT SIZES FOR META-ANALYSIS

MICHAEL BORENSTEIN
Biostat, Inc.

LARRY V. HEDGES
Northwestern University

CONTENTS

11.1 INTRODUCTION

In any meta-analysis, we start with summary data from each study and use this summary data to compute an effect size for that study. An effect size is a number that reflects the magnitude of the relationship between two variables. For example, if a study reports the mean and standard deviation for the treated and control groups, we might compute the standardized mean difference between groups. Or, if a study reports the number of events and non-events in two groups, we might compute an odds ratio. It is these effect sizes that serve as the unit of currency in the meta-analysis.

Consider figure 11.1, the forest plot of a fictional meta-analysis to assess the impact of an intervention. In this plot each study is represented by a square, bounded on either side by a confidence interval. The location of each square on the horizontal axis represents the effect size for that study. The confidence interval represents the precision with which the effect size has been estimated, and the size of each square is proportional to the weight that will be assigned to the study when computing the combined effect. This figure also functions as the outline for this chapter, in which we discuss what these items mean and how they are computed.

11.1.1 Effect Sizes and Treatment Effects

Meta-analyses in medicine often refer to the effect size as a *treatment effect*, a term sometimes assumed to refer to odds ratios, risk ratios, or risk differences, which are common in meta-analyses that deal with medical interventions. Similarly, meta-analyses in the social sciences often refer to the effect size simply as an *effect size*,

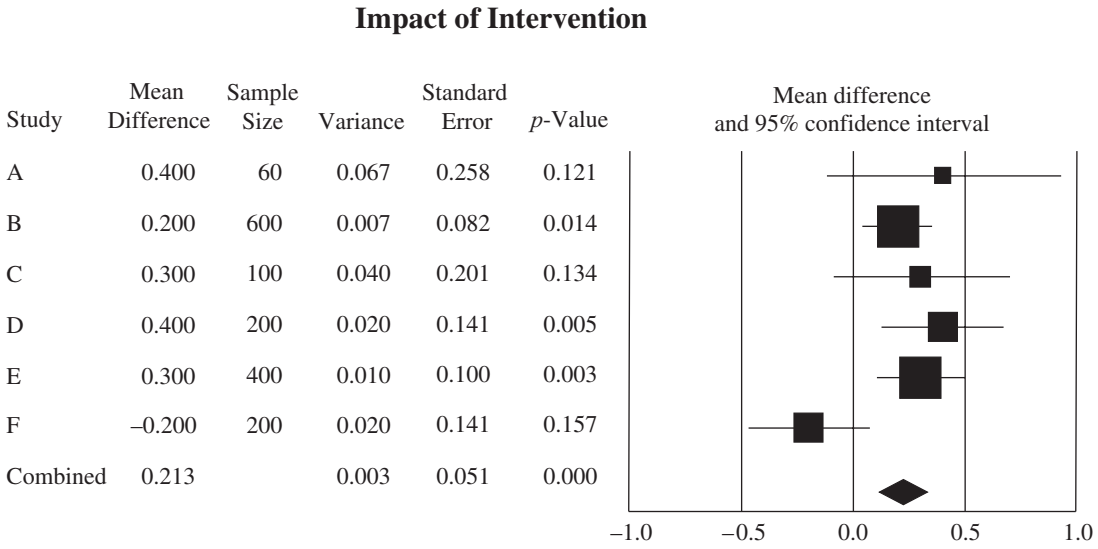

Figure 11.1 Forest Plot for Fictional Studies

SOURCE: Authors' tabulation.
NOTE: The effect size for each study is represented by a square, which is bounded by a confidence interval. The area of the square is proportional to the study's weight in the analysis. The confidence interval is based on the standard error of the estimate for the effect size.

which is sometimes assumed to refer to standardized mean differences or to correlations, which are common in social science meta-analyses.

In fact, though, both terms can refer to any of these indices, and the distinction between these terms lies not in the index itself but rather in the nature of the study. Effect size is appropriate when the index is used to quantify the relationship between any two variables or a difference between any two groups. By contrast, treatment effect is appropriate only for an index used to quantify the impact of a deliberate intervention. Thus, the difference between males and females could be called an effect size only, while the difference between treated and control groups could be called either an effect size or a treatment effect. The classification of an index as an effect size or a treatment effect has no bearing on the computations, however.

Four major considerations should drive the choice of an effect-size index. The first is that the effect sizes from the different studies should be comparable to one another in the sense that they measure (at least approximately) the same thing. That is, the effect size should not depend on aspects of study design that may vary from study to study (such as sample size or whether covariates are used). The second is that the effect size should be substantively interpretable. This means that researchers in the substantive area of the work represented in the synthesis should find the effect size meaningful. The third is that estimates of the effect size should be computable from the information that is likely to be reported in published research reports. That is, it should not require the reanalysis of the raw data. The fourth is that the effect size should have good technical properties. For example, its sampling distribution should be known so that variances and confidence intervals can be computed.

11.1.2 Effect Sizes Rather than *p*-Values

Reports of primary research typically include the *p*-value corresponding to a test of significance. This *p*-value reflects the likelihood that the sample would have yielded the observed effect, or one more extreme, if the null hypothesis was true.

Researchers often use the *p*-value as a surrogate for the effect size, with a significant *p*-value taken to imply a large effect and a nonsignificant *p*-value taken to imply a trivial effect. In fact, however, while the *p*-value is

partly a function of effect size, it is also partly a function of sample size. A *p*-value of 0.01 *could* reflect a large effect but could also reflect a trivial effect in a large sample. Conversely, a *p*-value of 0.20 *could* reflect a trivial effect but could also reflect a large effect in a small sample. In figure 11.1, for example, study A has a *p*-value of 0.12 and study B has a *p*-value of 0.01, but it is study A that has the larger effect size (0.40 versus 0.20) (see, for example Borenstein 1994, 1997, 2000).

In primary studies, we can avoid this kind of confusion by reporting the effect size and the precision separately. The former gives us a pure estimate of the effect in the sample, and the latter gives us a range for the effect in the population. Similarly, in a meta-analysis, we need to work with a pure measure of the effect size from each primary study to determine if the effects are consistent, and to compute a combined estimate of the effect across studies. Here, the precision of each effect is used to assign a weight to that effect in these analyses.

11.1.3 Effect-Size Parameters and Sample Estimates of Effect Sizes

Throughout this chapter, we make the distinction between an underlying effect-size parameter (denoted here by the Greek letter λ) and the sample estimate of that parameter (denoted here by T).

If a study had an infinitely large sample size, it would yield an effect size T that was identical to the population parameter λ. In fact, though, sample sizes are finite and so the effect-size estimate T always differs from λ by some amount. The value of T will vary from sample to sample, and the distribution of these values is the sampling distribution of T. Statistical theory allows us to compute the variance of the effect-size estimates.

11.1.4 Variance of Effect-Size Estimates

A dominant factor in the variance of T is the sample size, larger studies having a smaller variance and yielding a more precise estimate of the effect-size parameter. The use of a matched design or the inclusion of a covariate to reduce the error term will usually lead to a lower variance and more precise estimate. Additionally, the variance of any given effect size is affected by specific factors that vary from one effect size index to the next.

Most effect-size estimates used in meta-analysis are approximately normally distributed with mean λ. Thus the sampling distribution of an effect-size estimate T is determined by the standard error or the variance (which is simply the square of the standard error).

When our focus is on the effect size for a single study, we generally work with the standard error of the effect size, which in turn may be used to compute confidence intervals about the effect size. In figure 11.1, for example, study E has four times the sample size of study C (400 versus 100). Its standard error (the square root of the variance) is therefore half as large (0.10 versus 0.20) and its confidence interval half as wide as that of study C. (In this example, all other factors that could affect the precision were held constant).

By contrast, in a meta-analysis we work primarily with the variance rather than the standard error. In a fixed-effect analysis, for example, we weight by the inverse variance, or $1/V$. In figure 11.1, study E has four times the sample size of study C (400 versus 100) and its variance is therefore one-fourth as large (0.01 versus 0.04). The square representing study E has four times the area as the one for study C, reflecting the fact that it will be assigned four times as much weight in the analysis.

11.1.5 Effect-Size Estimates from Reported Information

When researchers have access to a full set of summary data such as means, standard deviations, and sample size for each group, the computation of the effect size and its variance is relatively straightforward. In practice, however, researchers will often find themselves working with only partial data. For example, a paper may publish only the *p*-value and sample size from a test of significance, leaving it to the meta-analyst to back-compute the effect size and variance. For this reason, each of the following sections includes a table that shows how to compute the effect size and variance from some of the more common reporting formats. For additional information on computing effect sizes from partial information, see the section on resources at the end of this chapter.

11.2 EFFECT SIZES FOR A COMPARISON OF MEANS

Suppose that we want to compare the mean for two populations, or for one population at two points in time. Two options for an effect size are the raw mean difference and the standardized mean difference. We introduce each of these indices and then discuss how to choose between them.

11.2.1 Raw (Unstandardized) Mean Difference *D*

Let μ_1 and μ_2 be the population means of two groups (or of one group at two points in time). The population mean difference is defined as

$$\Delta = \mu_1 - \mu_2. \tag{11.1}$$

Whereas the effect size Δ is the true (population) value (the effect size parameter), any study yields an effect-size estimate D that is based on the observed means and is an estimate of the true effect size. Note that we use uppercase D for the estimate of the raw mean difference, whereas we will use lowercase d for the estimate of the standardized mean difference (below). Here, we show how to compute D from studies that use two independent groups, and from studies that use matched groups or a pre-post design.

11.2.1.1 Computing *D*, Independent Groups Let $\overline{Y}_1$ and $\overline{Y}_2$ be the sample means of the two independent groups. Then D is the sample mean difference, namely

$$D = \overline{Y}_1 - \overline{Y}_2. \tag{11.2}$$

Let S_1 and S_2 be the sample standard deviations of the two groups, and n_1 and n_2 be the sample size in the two groups. If we assume that the two population standard deviations are the same (as is assumed in most parametric data analysis techniques), so that $\sigma_1 = \sigma_2 = \sigma$, then the estimate of the variance of D is

$$V_D = \frac{n_1 + n_2}{n_1 n_2} S^2_{Pooled}, \tag{11.3}$$

where

$$S^2_{Pooled} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}. \tag{11.4}$$

If we don't assume that the two population standard deviations are the same, then the estimate of the variance of D is

$$V_D = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}. \tag{11.5}$$

In either case, the standard error of D is then the square root of V_D,

$$SE_D = \sqrt{V_D}. \tag{11.6}$$

For example, suppose that a study has sample means $Y_1 = 103$, $Y_2 = 100$, sample standard deviations $S_1 = 5.5$, $S_2 = 4.5$, and sample sizes $n_1 = n_2 = 50$. The raw mean difference D is

$$D = 103 - 100 = 3.000.$$

If we assume that $\nu_1^2 = \nu_2^2$, then the pooled standard deviation within groups is

$$S_{Within} = \sqrt{\frac{(50-1)\times 5.5^2 + (50-1)\times 4.5^2}{50+50-2}} = 5.0249,$$

with the variance and standard error of D given by

$$V_D = \frac{50+50}{50\times 50} \times 5.0249^2 = 1.0100$$

and

$$SE_D = \sqrt{1.0100} = 1.0050.$$

If we do not assume that $\nu_1^2 = \nu_2^2$, then the variance and standard error of D are given by

$$V_D = \frac{5.5^2}{50} + \frac{4.5^2}{50} = 1.0100$$

and

$$SE_D = \sqrt{1.0100} = 1.0050.$$

In this example, both formulas yield the same result because $n_1 = n_2$.

11.2.1.2 Computing *D*, Matched Groups or Pre-Post Scores Let $\overline{Y}_1$ and $\overline{Y}_2$ be the sample means of two matched groups. Then D is the sample mean difference, namely,

$$D = \overline{Y}_1 - \overline{Y}_2. \tag{11.7}$$

Let S_1 and S_2 be the sample standard deviations of the two groups, and n be the number of matched pairs. The standard deviation of the paired differences is given by

$$S_{Difference} = \sqrt{S_1^2 + S_2^2 - 2 \times r \times S_1 \times S_2}, \tag{11.8}$$

where r is the correlation between "siblings" in matched pairs. As r moves toward 1.0, the standard error of the difference will decrease.

If $S_1 = S_2$ then (11.8) simplifies to

$$S_{Difference} = \sqrt{2 \times S^2_{Pooled}(1-r)}. \quad (11.9)$$

In either case, the variance of D is computed as

$$V_D = \frac{S^2_{Difference}}{n}, \quad (11.10)$$

and the standard error is then

$$SE_D = \sqrt{V_D}. \quad (11.11)$$

For example, suppose that a study has sample means $Y_1 = 105$, $Y_2 = 100$, sample standard deviations $S_1 = 10$, $S_2 = 10$, and sample sizes = 50 pairs of scores. The correlation between the two sets of scores is 0.50. The raw mean difference D is

$$D = 105 - 100 = 5.000.$$

The variance and standard error of D are given by

$$V_D = \frac{10^2}{50} = 2.000$$

and

$$SE_D = \sqrt{2} = 1.4142.$$

In the calculation of V_D, $S_{Difference}$ is computed using

$$S_{Difference} = \sqrt{10^2 + 10^2 - 2 \times 0.50 \times 10 \times 10} = 10, \quad (11.12)$$

or

$$S_{Difference} = \sqrt{2 \times 10^2 (1 - 0.50)} = 10. \quad (11.13)$$

The formulas for matched designs apply to pre-post designs as well. The pre- and post-means correspond to the group means in the matched groups, n is the number of subjects, and r is the correlation between pre-scores and post-scores.

11.2.2 STANDARDIZED MEAN DIFFERENCE *d* AND *g*

Let μ_1 and σ_1 be the mean and standard deviation of one population, and let μ_2 and σ_2 be the mean and standard deviation of a second population. If the two population standard deviations are the same (as is assumed in most parametric data analysis techniques), so that $\sigma_1 = \sigma_2 = \sigma$, then the population standardized mean difference (the standardized mean difference parameter) is defined as

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}. \quad (11.14)$$

Whereas the effect-size δ is the true or population value, any study yields an effect size that is an estimate of the true effect size. We present two versions of this estimate, d and Hedges' g. Both are intended to estimate δ, but the estimate called d has a bias, and tends to exaggerate the absolute value of δ. The estimate called Hedges' g removes most of this bias. Note that we use lowercase d for the standardized mean difference, whereas we used uppercase D for the raw mean difference.

Here, we show how to compute d and g from studies that use two independent groups, from studies that use matched groups or a pre-post design, and from studies that employed analysis of covariance.

11.2.2.1 Computing *d* and *g*, Independent Groups We can estimate the standardized mean difference (δ) from studies that use two independent groups as

$$d = \frac{\overline{Y}_1 - \overline{Y}_2}{S_{Within}}. \quad (11.15)$$

In the numerator, $\overline{Y}_1$ and $\overline{Y}_2$ are the sample means in the two groups. In the denominator, S_{Within} is the within-groups standard deviation, pooled across groups,

$$S_{Within} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}, \quad (11.16)$$

where n_1, n_2 are the sample size in the two groups, and S_1, S_2 are the standard deviations in the two groups. The reason we pool the two sample estimates of the standard deviation is that even if we assume that the underlying population standard deviations are the same (that is $\sigma_1 = \sigma_2 = \sigma$), it is unlikely that the sample estimates S_1

and S_2 will be identical. By pooling the two estimates of the standard deviation, we obtain a more precise estimate of their common value.

The variance of d is given (to a very good approximation) by

$$V_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}. \qquad (11.17)$$

In this equation, the first term on the right expresses the contribution to the overall uncertainty of d due to the uncertainty in the estimate of the mean difference, the numerator in (11.15), and the second expresses contribution of the uncertainty in the estimate of S_{Within}, the denominator in (11.15).

The standard error of d is the square root of V_d,

$$SE_d = \sqrt{V_d}. \qquad (11.18)$$

It turns out that d has a slight bias, tending to overestimate the absolute value of δ in small samples. If we multiply d by a correction factor, we get an unbiased estimate, called Hedges' g. The correction factor (J) depends on the degrees of freedom. Concretely,

$$J(df) = 1 - \frac{3}{4df - 1}. \qquad (11.19)$$

In this expression, df is the degrees of freedom used to estimate S_{Within}, which for two independent groups is $n_1 + n_2 - 2$. Then,

$$g = J(df)d, \qquad (11.20)$$

$$V_g = [J(df)]^2 V_d \qquad (11.21)$$

and

$$SE_g = \sqrt{V_g}. \qquad (11.22)$$

For example, suppose a study has sample means $Y_1 = 103$, $Y_2 = 100$, sample standard deviations $S_1 = 5.5$, $S_2 = 4.5$, and sample sizes $n_1 = n_2 = 50$. We would estimate the pooled-within-groups standard deviation as

$$S_{Within} = \sqrt{\frac{(50-1)\times 5.5^2 + (50-1)\times 4.5^2}{50 + 50 - 2}} = 5.0249.$$

Then,

$$d = \frac{103 - 100}{5.0249} = 0.5970,$$

$$V_d = \frac{50 + 50}{50 \times 50} + \frac{0.5970^2}{2(50 + 50)} = 0.0418,$$

and

$$SE_d = \sqrt{0.0418} = 0.2045.$$

The correction factor J, Hedges' g, its variance and standard error are given by

$$J(50 + 50 - 2) = \left(1 - \frac{3}{4 \times 98 - 1}\right) = 0.9923,$$

$$g = 0.9923 \times 0.5970 = 0.5924,$$

$$V_g = 0.9923^2 \times 0.0418 = 0.0412,$$

and

$$SE_g = \sqrt{0.0412} = .2030.$$

The correction factor J is always less than 1.0, and so g will always be less than d in absolute value, and the variance of g will always be less than the variance of d. However, J will be very close to 1.0 unless the degrees of freedom are very low (say, less than 10) and so (as in this example) the difference is usually trivial.

Some slightly different expressions for the variance of d (and g) have been given by different authors and even by the same authors at different times. For example, the denominator of the second term of the variance of d is given here as $2(n_1 + n_2)$. This expression is obtained by one method (assuming the n's become large with δ fixed). An alternate derivation (assuming n's become large with $\sqrt{n}\delta$ fixed) leads to a denominator in the second term that is slightly different, namely $2(n_1 + n_2 - 2)$. Unless n_1 and n_2 are very small, these expressions will be almost identical. Similarly, the expression given here for the variance of g is J^2 times the variance of d, but many authors ignore the J^2 term because it is so close to unity in most cases. Again, although it is preferable to include this correction factor, including this factor is likely to make little practical difference.

Table 11.1 provides formulas for computing d and its variance for independent groups, working from information that may be reported in a published paper.

11.2.2.2 Computing *d* and *g*, Pre-Post Scores or Matched Groups We can estimate the standardized mean difference (δ) from studies that used matched groups or pre-post scores in one group. We compute d using

$$d = \frac{\overline{Y}_1 - \overline{Y}_2}{S_{Within}}. \tag{11.23}$$

This is the same formula as for independent groups (11.15). If the study reports S_{Pre} and S_{Post} then we can estimate S_{Within} using

$$S_{Within} = \sqrt{\frac{S_{Pre}^2 + S_{Post}^2}{2}}. \tag{11.24}$$

However, some studies will not report S_{Pre} and S_{Post}. Rather, they will report the standard deviation of the difference, $S_{Difference}$. In this case we can impute S_{Within} using

$$S_{Within} = \frac{S_{Difference}}{\sqrt{2(1-r)}}, \tag{11.25}$$

where r is the correlation between pairs of observations (for example, the correlation between pretest and posttest). Then we can apply (11.15) to compute d. The variance of d is given by

$$V_d = \left(\frac{1}{n} + \frac{d^2}{2n}\right)2(1-r), \tag{11.26}$$

where n is the number of pairs. The standard error of d is just the square root of V_d,

$$SE_d = \sqrt{V_d}. \tag{11.27}$$

Table 11.1 Computing *d*, Independent Groups

Reported	Computation of Needed Quantities
$\overline{Y}_1$, $\overline{Y}_2$ S_{Pooled}, n_1, n_2	$d = \frac{Y_1 - Y_2}{S_{Pooled}}, v = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}$
t, n_1, n_2	$d = t\sqrt{\frac{n_1 + n_2}{n_1 n_2}}, v = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}$
F, n_1, n_2	$d = \pm\sqrt{\frac{F(n_1 + n_2)}{n_1 n_2}}, v = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}$
p(one-tailed), n_1, n_2	$d = \pm t^{-1}(p)\sqrt{\frac{n_1 + n_2}{n_1 n_2}}, v = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}$
p(two-tailed), n_1, n_2	$d = \pm t^{-1}\left(\frac{p}{2}\right)\sqrt{\frac{n_1 + n_2}{n_1 n_2}}, v = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}$

SOURCE: Authors' tabulation.

NOTE: The function $t^{-1}(p)$ is the inverse of the cumulative distribution function of Student's t with $n_1 + n_2 - 2$ degrees of freedom. Many computer programs and spreadsheets provide functions that can be used to compute $t - 1$. Assume $n_1 = n_2 = 10$, so that $df = 18$. Then, in Excel, for example, if the reported p-value is 0.05 (two-tailed) TINV(p,df) = TINV(0.05,18) will return the required value (2.1009). If the reported p-value is 0.05 (one-tailed), TINV($2p$,df) = TINV(0.10,18) will return the required value 1.7341. The F in row 3 is the F-statistic from a one-way analysis of variance. In rows 3 through 5, the sign of d must reflect the direction of the mean difference.

Because the correlation between pretest and posttest scores is required to impute the standard deviation within groups from the standard deviation of the difference, we must assume that this correlation is known or can be estimated with high precision. Often, the researcher will need to use data from other sources to estimate this correlation. If the correlation is not known precisely, one could work with a range of plausible correlations, and use a sensitivity analysis to see how these affect the results.

To compute Hedges' g and associated statistics we would use formulas (11.23) through (11.27). The degrees of freedom for computing J is $n-1$, where n is the number of pairs.

For example, suppose that a study has pretest and posttest sample means $Y_1 = 103$, $Y_2 = 100$, sample standard deviation of the difference $S_{Difference} = 5.5$, and sample size $n = 50$ and a correlation between pretest and posttest of $r = 0.70$.

The standard deviation within groups is imputed from the standard deviation of the difference by

$$S_{Within} = \frac{5.5}{\sqrt{2(1-0.7)}} = 7.1000.$$

Then d, its variance and standard error are computed as

$$d = \frac{103-100}{7.1000} = 0.4225,$$

$$v_d = \left(\frac{1}{50} + \frac{0.4225^2}{2\times 50}\right)(2(1-0.7)) = 0.0131,$$

and

$$SE_d = \sqrt{0.0131} = .1145.$$

The correction factor J, Hedges' g, its variance and standard error are given by

$$J(n-1) = \left(1 - \frac{3}{4\times 49 - 1}\right) = 0.9846,$$

$$g = 0.9846 \times 0.4225 = 0.4160,$$

$$V_g = 0.9846^2 \times 0.0131 = 0.0127,$$

and

$$SE_g = \sqrt{0.0127} = 0.1127.$$

Table 11.2 provides formulas for computing d and its variance for matched groups, working from information that may be reported in a published paper.

11.2.2.3 Computing *d* and *g*, Analysis of Covariance We can estimate the standardized mean difference (δ) from studies that used analysis of covariance. The formulas for the sample estimate of d is

$$d = \frac{\overline{Y}_1^{Adjusted} - \overline{Y}_2^{Adjusted}}{S_{Within}}. \qquad (11.28)$$

This is similar to the formula for independent groups (11.15), but with two differences. The first difference is in the numerator, where we use adjusted means rather than raw means. This has no impact on the expected value of the mean difference but increases precision and possibly removes bias due to pretest differences.

The second is in the mechanism used to compute S_{Within}. When we were working with independent groups the natural unit of deviation was the unadjusted standard deviation within groups. Therefore, this value is typically reported or easily imputed. By contrast, the test statistics used in the analysis of covariance involve the adjusted standard deviation (typically smaller than the unadjusted value since variance explained by covariates has been removed). Therefore, to compute d from this kind of study, we need to impute the unadjusted standard deviation within groups. We can do this using

$$S_{Within} = \frac{S_{Adjusted}}{\sqrt{1-R^2}}, \qquad (11.29)$$

where $S_{Adjusted}$ is the covariate-adjusted standard deviation and R is the covariate outcome correlation (or multiple correlation if there is more than one covariate). Note the similarity to (11.25) which we used to impute S_{Within} from $S_{Difference}$ in matched studies. Equivalently, S_{Within} may be computed as

$$S_{Within} = \sqrt{\frac{MSW_{Adjusted}}{1-R^2}}, \qquad (11.30)$$

where $MSW_{Adjusted}$ is the covariate-adjusted mean square within treatment groups.

Table 11.2 Computing *d*, Matched Groups

Reported	Computation of Needed Quantities
$\overline{Y}_1$, $\overline{Y}_2$ $S_{\text{Difference}}$, r, n (number of pairs)	$d = \left(\frac{\overline{Y}_1 - \overline{Y}_2}{S_{Difference}}\right)\sqrt{2(1-r)}$, $v = \left(\frac{1}{n} + \frac{d^2}{2n}\right)2(1-r)$
t (from paired t-test), r, n	$d = t\sqrt{\frac{2(1-r)}{n}}$, $v = \left(\frac{1}{n} + \frac{d^2}{2n}\right)2(1-r)$
F (from repeated measures ANOVA), r, n	$d = \pm\sqrt{\frac{2F(1-r)}{n}}$, $v = \left(\frac{1}{n} + \frac{d^2}{2n}\right)2(1-r)$
p (one-tailed), r, n	$d = \pm t^{-1}(p)\sqrt{\frac{2(1-r)}{n}}$, $v = \left(\frac{1}{n} + \frac{d^2}{2n}\right)2(1-r)$
p (two-tailed), r, n	$d = \pm t^{-1}\left(\frac{p}{2}\right)\sqrt{\frac{2(1-r)}{n}}$, $v = \left(\frac{1}{n} + \frac{d^2}{2n}\right)2(1-r)$

SOURCE: Authors' tabulation.

NOTE: The function $t^{-1}(p)$ is the inverse of the cumulative distribution function of Student's t with $n - 1$ degrees of freedom. Many computer programs and spreadsheets provide functions that can be used to compute t^{-1}. Assume $n = 19$, so that $df = 18$. Then, in Excel, for example, if the reported p-value is 0.05 (2-tailed), TINV(p,df) = TINV(0.05,18) will return the required value (2.1009). If the reported p-value is 0.05 (1-tailed), TINV(2p,df) = TINV(0.10,18) will return the required value 1.7341. The F in row 3 of the table is the F-statistic from a one-way repeated measures analysis of variance. In rows 3 through 5, the sign of d must reflect the direction of the mean difference.

The variance of d is given by

$$V_d = \frac{(n_1 + n_2)(1 - R^2)}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}, \qquad (11.31)$$

where n_1 and n_2 are the sample size in each group, and R is the multiple correlation between the covariates and the outcome.

To compute Hedges' g and associated statistics we would use formulas (11.19) through (11.22). The degrees of freedom for computing J is $n_1 + n_2 - 2 - q$, where n_1 and n_2 are the number of subjects in each group, 2 is the number of groups, and q is the number of covariates. For example, suppose that a study has sample means $Y_1 = 103$, $Y_2 = 100$, sample standard deviations $S_{Adjusted} = 5.5$, and sample sizes $n_1 = n_2 = 50$. Suppose that we know that the covariate outcome correlation for the single covariate (so that $q = 1$) is $R = 0.70$.

The unadjusted standard deviation within groups is imputed from $S_{Adjusted}$ by

$$S_{Within} = \frac{S_{Adjusted}}{\sqrt{1 - 0.7^2}} = 7.7015.$$

Then d, its variance and standard error are computed as

$$d = \frac{103 - 100}{7.7015} = 0.3895,$$

$$V_d = \frac{(50 + 50)(1 - 0.7^2)}{50 \times 50} + \frac{0.3895^2}{2(50 + 50 - 2 - 1)} = 0.0222,$$

and

$$SE_d = \sqrt{0.0222} = .1490.$$

The correction factor J, Hedges' g, its variance and standard error are given by

$$J(50 + 50 - 2 - 1) = \left(1 - \frac{3}{4 \times 97 - 1}\right) = 0.9922,$$

$$g = 0.9922 \times 0.3895 = 0.3834,$$

$$V_g = 0.9922^2 \times 0.0222 = 0.0219,$$

and

$$S_g = \sqrt{0.0219} = .1480.$$

Table 11.3 provides formulas for computing d and its variance for analysis of covariance, working from information that may be reported in a published paper.

11.2.3 Direction of the Effect

The discussion that follows applies to all the effect-size indices in this section—that is, D, d, and g.

The direction of the effect (for example, group 1 versus group 2 or group 2 versus group 1) is arbitrary, except that it must be consistent from one study to the next. That is, if a mean difference above 0 in study A indicates that group 1 did better, then a mean difference above 0 in studies B, C, . . . must also indicate that group 1 did better. Although the direction is arbitrary, several useful conventions make it easier to interpret the results.

When we are comparing treated versus control, it is conventional to compute the difference as treatment minus control. For a pre- or post-design this would be post minus pre. When we follow this convention, the direction of the effect will be as follows: if the treatment increases the mean score, the mean difference will be positive; if the treatment decreases the mean score, the mean difference will be negative.

As a general rule, it is a good idea to follow this convention because the results will follow the usual pattern. For example, if the treatment is intended to boost scores, we expect to see the mean difference to right of zero. Similarly, if the treatment is intended to lower scores, we expect to see the mean difference to the left of zero.

Although the convention is therefore useful, it is arbitrary. If we chose to reverse the position of the two groups, the effect-size format would change but the substantive meaning would remain the same. Concretely, the

Table 11.3 Computing *d*, Independent Groups Using ANCOVA

Reported	Computation of Needed Quantities
$\bar{Y}_1$, $\bar{Y}_2$ S_{Pooled}, n_1, n_2, R, q	$d = \dfrac{\bar{Y}_1^{Adjusted} - \bar{Y}_2^{Adjusted}}{S_{Pooled}}$, $v = \dfrac{(n_1+n_2)(1-R^2)}{n_1 n_2} + \dfrac{d^2}{2(n_1+n_2)}$
t (from ANCOVA), n_1, n_2, R, q	$d = t\sqrt{\dfrac{n_1+n_2}{n_1 n_2}}\sqrt{1-R^2}$, $v = \dfrac{(n_1+n_2)(1-R^2)}{n_1 n_2} + \dfrac{d^2}{2(n_1+n_2)}$
F (from ANCOVA), n_1, n_2, R, q	$d = \pm\sqrt{\dfrac{F(n_1+n_2)}{n_1 n_2}}\sqrt{1-R^2}$, $v = \dfrac{(n_1+n_2)(1-R^2)}{n_1 n_2} + \dfrac{d^2}{2(n_1+n_2)}$
p (1-tailed, from ANCOVA), n_1, n_2, R, q	$d = \pm t^{-1}(p)\sqrt{\dfrac{n_1+n_2}{n_1 n_2}}\sqrt{1-R^2}$, $v = \dfrac{(n_1+n_2)(1-R^2)}{n_1 n_2} + \dfrac{d^2}{2(n_1+n_2)}$
p (2-tailed, from ANCOVA), n_1, n_2, R, q	$d = \pm t^{-1}\left(\dfrac{p}{2}\right)\sqrt{\dfrac{n_1+n_2}{n_1 n_2}}\sqrt{1-R^2}$, $v = \dfrac{(n_1+n_2)(1-R^2)}{n_1 n_2} + \dfrac{d^2}{2(n_1+n_2)}$

SOURCE: Authors' tabulation.

NOTE: The function $t^{-1}(p)$ is the inverse of the cumulative distribution function of Student's t with $n_1 + n_2 - 2 - q$ degrees of freedom, q is the number of covariates, and R is the covariate outcome correlation or multiple correlation. Many computer programs and spreadsheets provide functions that can be used to compute t^{-1}. Assume $n_1 = n_2 = 11$, and $q = 2$, so that $df = 18$. Then, in Excel, for example, if the reported p-value is 0.05 (2-tailed), TINV(p,df) = TINV(0.05,18) will return the required value (2.1009). If the reported p-value is 0.05 (1-tailed), TINV(2p,df) = TINV(0.10,18) will return the required value 1.7341. The F in row 3 of the table is the F-statistic from a one-way analysis of covariance. In rows 3 through 5, the sign of d must reflect the direction of the mean difference.

three effects sizes would be modified as follows. If we switched the position of the two groups, the new raw mean difference would be

$$D_{New} = -D_{Old}, \tag{11.32}$$

the standardized mean difference would be

$$d_{New} = -d_{Old}, \tag{11.33}$$

and Hedges' g would be

$$g_{New} = -g_{Old}. \tag{11.34}$$

That the decision to put one group or the other first is arbitrary, is especially important when we are comparing two active treatments, rather than a treatment and a control. If some studies compute A minus B and others compute B minus A, we need to ensure that all the results are in the same direction before we enter them into the meta-analysis. We could apply formulas (11.32) to (11.34) to the effect size and confidence interval as needed, to reverse the direction of the effect for any given study.

11.2.4 Choosing an Index

If all studies employed the same scale to measure outcome, we have the option of using either the raw mean difference D or the standardized mean difference d (or g). If some studies used one scale but other studies used an alternate scale, then we must use the standardized mean difference.

Consider the fictional example displayed in table 11.4. These studies compare the level of pain reported by patients who have been randomized to either treatment or control conditions. The first four studies assess pain on a scale that ranges from 0 to 100 with a standard deviation of 20 points. The next four assess pain on a scale that ranges from 0 to 10 with a standard deviation of 2 points. Columns show the raw mean difference (D), the standardized mean difference (d), and the standardized mean difference (g) for each study.

The treatment's impact is identical in all eight studies. For the first set of studies, the mean difference is consistently 10 points; for the second set, it is consistently 1 point.

If we were to run an analysis using only the first four studies, we could use either the raw mean difference or the standardized mean difference. Similarly, if we were to run an analysis using only the last four studies, we could use either the raw mean or the standardized mean difference.

However, if we wanted to run an analysis that includes all eight studies, we could not use the raw mean difference because doing so would make it appear as if the effect size were ten times larger in some studies than others. Rather, we would use the standardized mean difference, which reports the effect size in standard deviations. In the first set of studies, the standard deviation is 20, so a difference of 10 points is 0.50 standard deviations. In the second set, the standard deviation is 2, so a difference of 1 point is 0.50 standard deviations. If we use these values in the

Table 11.4 Fictional Studies: Computing *D*, *d*, and *g*

Study	Mean (Treated)	Mean (Control)	$S_1 = S_2$	$n_1 = n_2$	Mean Difference (D)	Standardized Mean Difference (d)	Standardized Mean Difference (g)
1	40	50	20	20	–10	–0.500	–0.490
2	40	50	20	20	–10	–0.500	–0.490
3	40	50	20	20	–10	–0.500	–0.490
4	40	50	20	20	–10	–0.500	–0.490
5	4	5	2	20	–1	–0.500	–0.490
6	4	5	2	20	–1	–0.500	–0.490
7	4	5	2	20	–1	–0.500	–0.490
8	4	5	2	20	–1	–0.500	–0.490

SOURCE: Authors' tabulation.

analysis, the effect size for all the studies is the same. In any set of real studies, of course, the effect size will vary—but that will be due to real differences in the effects and to sampling error rather than to the use of different scales.

As before, if all studies use the same scale, we have the option of using either the raw mean difference or the standardized mean difference. In that case, we should choose the more intuitive scale. If the scale is based on a physical measurement (such as blood pressure) or is used widely (such as a national test for student achievement), the raw scale will generally be more intuitive. By contrast, if the test is less well known, then the standardized mean difference will generally be more intuitive. In some fields (such as education research) the standardized mean difference is ubiquitous, and will often be the first choice.

To this point, we have focused on choosing between the raw mean difference and the standardized mean difference. If we select the latter, we still need to choose between *d* and *g*. These are both estimating the same value: one (*d*) tends to overestimate that value (to push it further from zero) and the other (*g*) removes most of this bias. Therefore, if we have the option to use either, *g* is generally the better choice. However, we can only compute *g* if we know the sample size in each group, and thus in some cases that is not an option. As a practical matter, unless the sample size is less than 10, the difference between *d* and *g* is usually trivial. In table 11.4, where the sample size is 20 per group, the correct estimate (*g*) is 0.49, but the biased estimate (*d*) is 0.50.

11.2.5 Understanding Raw Mean Difference and Standardized Mean Difference

The effect size index *D* is an intuitive index—it's simply the difference between two means. As such, researchers and consumers tend to be comfortable with this index. We understand how to interpret a difference in means, and if the difference is on a meaningful scale (which it should be) we understand the substantive implication of the difference. If we are told that an intervention increases the mean score on a national math test by 50 points, we understand what that means for the students.

The standardized mean difference is also an intuitive index, once we become familiar with it. It is also the difference between two means—except that this time, it is on a standardized scale (Hedges and Olkin 1985). The example is a case in point. In the first set of studies, the treatment bumped up the score by 0.50 standard deviations. In the second set, the treatment bumped up the score by 0.50 standard deviations. Because the standardized mean difference is based on the standard deviation, we can map the results to any other scale. In our example, the standardized mean difference is 0.50 standard deviations. If we normally use a scale with a standard deviation of 20 points, this corresponds to a difference of 10 points.

A key advantage of the standardized mean difference is the fact that it is widely used, and not tied to any scale. This allows us to view the results of our intervention in the context of other interventions. For example, if we are working in a field where most effects fall in the range of 0.20 to 0.80, we know immediately what it means if our intervention yields an effect size of (for example) 0.50 or 0.90.

Various researchers have suggested other ways to think of the standardized mean difference. Hedges and Olkin explain that this index reflects the extent to which the two groups represent distinct clusters, even if the scales do not measure precisely the same thing (1985). Cohen shows several ways that we can translate the standardized mean difference into the amount of overlap or nonoverlap between the groups (1969, 1977). Various scholars explore similar ideas (see Borenstein 2019; Glass, McGaw, and Smith 1981; Glass 1976; Hedges and Olkin 1985, 2016). Jeffrey Valentine and Ariel Aloe suggest that we can use the standardized mean difference to predict the number needed to treat (2016). Larry Hedges and Ingram Olkin (1985) and Robert Grissom and John Kim (2012) show how to compute *d* for non-normal distributions or from nonparametric data. Elsewhere, we also show how to compute *d* from an array of study designs.

We need now to highlight a few issues that may not be obvious. We define the standardized mean as

$$\delta = \frac{\mu_1 - \mu_2}{\sigma},$$

which is estimated using

$$d = \frac{\overline{Y}_1 - \overline{Y}_2}{S_{Within}}.$$

In these formulas, we standardize by the standard deviation within groups. When a study reports a different standard deviation (such as the standard deviation of the difference, or the standard deviation adjusted for covariates) we do not use that standard deviation in the formula.

Rather, we impute the standard deviation within groups, using either equation (11.25) or equation (11.29), as detailed earlier, and then use that value in the formula.

It is possible to define an index that would standardize by another standard deviation, and later in this chapter, when we cover multilevel studies, we discuss cases where we might want to do so. However, it is critical to understand that the index created in this way would be a different index than the one discussed here. Such an index could not be combined with d in an analysis, and could not be used to put the results in the context of other studies that reported d. To do so would be akin to using feet as the unit of measurement in some studies and meters as the unit of measurement in others.

We said at the outset that we can use the raw mean difference only if all studies use the same scale. However, cases in which different scales are measuring precisely the same thing, except in different metrics, are an exception. For example, suppose that some studies report the outcome in days while others report the outcome in hours. To convert the former to the latter we simply multiply the mean difference and the standard deviation by twenty-four. This works in this example because the two scales are measuring precisely the same thing, and we know the conversion factor. In other cases (such as the example of the pain scales, or scales to measure achievement or depression) the scales are not measuring precisely the same thing, and no conversion factor exists. In those cases, the only option is to use the standardized mean difference.

We noted that we can use the raw mean difference only if all studies used the same scale. Additionally, it is best to use the raw mean difference only if the standard deviation is roughly comparable across studies. If the standard deviation varies dramatically, it is probably better to use the standardized mean difference.

Finally, we note that the standardized mean difference allows us to work with studies that used different scales by converting these scales into a common metric. It does nothing to change the meaning of the scales. We assume that all scales are addressing the same fundamental question.

11.3 CORRELATIONS

11.3.1 Correlation as an Effect Size

Some studies report their results as correlation coefficients. When this is the case, the correlation itself will usually serve as the effect-size index. The correlation coefficient is an intuitive measure that, like δ, has been standardized to take account of different metrics in the original scales. The population parameter is denoted by ρ.

A correlation can take on any value between −1.0 and +1.0 (inclusive). A correlation of zero indicates no relationship between the variables. A correlation less than zero indicates that a high value of one variable is associated with a low value of the other. A correlation greater than zero indicates that a high value of one variable is associated with a high value of the other. The direction of the correlation must have the same meaning in all studies. If some studies report the correlation between education and the number of items correct, but others report the correlation between education and number of mistakes on a test, we would need to reverse the sign on one of these effects to make them comparable.

Although we often think of a correlation as applying to two continuous variables, it is also possible to compute a correlation between a dichotomous variable and a continuous variable. For example, we can code control and treatment as 0 and 1. A positive correlation tells us that the treated group scored higher, and a negative correlation tells us that the treated group scored lower.

Correlations are a key effect size in industrial organizational psychology. The use of correlations in this field is discussed in detail elsewhere in this volume (chapter 15).

11.3.1.1 Computing *r* The estimate of the correlation parameter ρ is simply the sample correlation coefficient, r. The variance of r is approximately

$$v_r = \frac{(1-r^2)^2}{n-1}, \tag{11.35}$$

where n is the sample size.

Most meta-analysts do not perform syntheses on the correlation coefficient itself because the variance depends so strongly on the correlation (see, however, Hunter and Schmidt 2015; chapter 15). Rather, the correlation is converted to the Fisher's z scale (not to be confused with the z-score used with significance tests), and all analyses are performed using the transformed values. The results, such as the combined effect and its confidence interval, would then be converted back to correlations for presentation. This is analogous to the procedure used with odds ratios or risk ratios where all analyses are performed using log transformed values, and then converted back to the original metric.

The transformation from sample correlation r to Fisher's z is given by

$$z = 0.5 \times \ln\left(\frac{1+r}{1-r}\right). \tag{11.36}$$

The variance of Fisher's z is (to an excellent approximation)

$$v_z = \frac{1}{n-3}, \tag{11.37}$$

with standard error

$$SE_z = \sqrt{V_z}. \tag{11.38}$$

The effect size z and its variance would be used in the analysis, which would yield a combined effect, confidence limits, and so on, in the Fisher's z metric. We could then convert each of these values back to correlation units using

$$r = \frac{e^{2z} - 1}{e^{2z} + 1} \tag{11.39}$$

where $e^x = \exp(x)$ is the exponential (anti-log) function. For example, if a study reports a correlation of 0.50 with a sample size of 100, we would compute

$$z = 0.5 \times \ln\left(\frac{1+0.5}{1-0.5}\right) = 0.5493,$$

$$v_z = \frac{1}{100-3} = 0.0103,$$

and

$$SE_z = \sqrt{0.0103} = 0.1015.$$

To convert the Fisher's z value back to a correlation, we would use

$$r = \frac{e^{2(0.5493)} - 1}{e^{2(0.5493)} + 1} = 0.5000.$$

Table 11.5 provides formulas for computing r and its variance, working from information that may be reported in a published paper.

11.3.1.2 Understanding *r* The correlation coefficient is commonly used in primary studies, and as such is an

Table 11.5 Computing *r*

Reported	Computation of Needed Quantities
r, n	$v_r = \frac{(1-r^2)^2}{n-1}, z = 0.5\ln\left(\frac{1+r}{1-r}\right), v_z = \frac{1}{n-3}$
t, n	$r = \pm\sqrt{\frac{t^2}{t^2+n-2}}, v_r = \frac{(1-r^2)^2}{n-1}, z = 0.5\ln\left(\frac{1+r}{1-r}\right), v_z = \frac{1}{n-3}$
t, r	$n = t^2\left(\frac{1-r^2}{r^2}\right) - 2, v_r = \frac{(1-r^2)^2}{n-1}, z = 0.5\ln\left(\frac{1+r}{1-r}\right), v_z = \frac{1}{n-3}$
p(one-tailed), r	$n = \left[t^{-1}(p)\right]^2\left(\frac{1-r^2}{r^2}\right) - 2, v_r = \frac{(1-r^2)^2}{n-1}, z = 0.5\ln\left(\frac{1+r}{1-r}\right), v_z = \frac{1}{n-3}$
p(two-tailed), r	$n = \left[t^{-1}\left(\frac{p}{2}\right)\right]^2\left(\frac{1-r^2}{r^2}\right) - 2, v_r = \frac{(1-r^2)^2}{n-1}, z = 0.5\ln\left(\frac{1+r}{1-r}\right), v_z = \frac{1}{n-3}$

SOURCE: Authors' tabulation.

NOTE: The function $t^{-1}(p)$ is the inverse of the cumulative distribution function of Student's t with $n - 2$ degrees of freedom. Many computer programs and spreadsheets provide functions that can be used to compute t^{-1}. Assume $n = 20$, so that $df = 18$. Then, in Excel, for example, if the reported p-value is 0.05 (2-tailed), TINV(p,df) = TINV(0.05,18) will return the required value (2.1009). If the reported p-value is 0.05 (1-tailed), TINV(2p,df) = TINV(0.10,18) will return the required value 1.7341.

Table 11.6 Cells for a Prospective Study

	Events	Nonevents	Total
Group 1	A	B	n_1
Group 2	C	D	n_2

SOURCE: Authors' tabulation.

intuitive metric for expressing an effect size (for additional insights, see Rosenthal and Rubin 1979, 1982; Rosenthal, Rosnow, and Rubin 2000; Rosnow, Rosenthal, and Rubin 2000; Hunter and Schmidt 2015).

11.4 EFFECT SIZES FOR COMPARING RISKS

This section deals exclusively with prospective studies. Retrospective (case-control studies) are addressed elsewhere.

Consider a study that compares the risk in two groups. We use the term *risk* as shorthand for the presence of an event, but this discussion applies also when the event is not a risk but instead a positive outcome. In either case, three options for the effect size are the risk difference, the risk ratio, and the odds ratio.

Table 11.6 shows the notation that we will use throughout this section. The rows identify the group, while the columns identify the outcome. This yields a 2×2 table with cells A, B, C, D; these labels will be used in the formulas.

If the risk (or probability) of an event in the two populations are π_1 and π_2, natural estimates of these parameters are given by

$$p_1 = \frac{A}{n_1},\ p_2 = \frac{C}{n_2}. \tag{11.40}$$

We will use a fictional study (table 11.7) as the example in this section. In this example, patients are randomly assigned to either Treated or Control conditions, and we record the number of deaths in each group.

Table 11.7 Data from a Fictional Prospective Study

	Dead	Alive	Total
Treated	5	95	100
Control	10	90	100

SOURCE: Authors' tabulation.

There are 100 patients in each group. There are five deaths in the treated group and ten deaths in the control group, so the risks in the two groups are estimated as

$$p_1 = \frac{5}{100} = 0.05,\ p_2 = \frac{10}{100} = 0.10.$$

11.4.1 The Risk Difference

The risk difference (Δ) is defined as the difference in probabilities (or risks) in the two groups. Let π_1 be the risk of an event in group 1, and π_2 be the risk of the event in group 2. Then the risk difference Δ, is defined as

$$\Delta = \pi_1 - \pi_2. \tag{11.41}$$

Whereas the effect-size Δ is the true (population) value, any study yields an effect-size *RD* that is based on the observed risks and is an estimate of the true effect size. To estimate RD we can use

$$RD = p_1 - p_2. \tag{11.42}$$

Using the notation in table 11.6, we estimate *RD* as

$$RD = \frac{A}{n_1} - \frac{C}{n_2}, \tag{11.43}$$

with approximate variance and standard error

$$V_{RD} = \frac{AB}{n_1^3} + \frac{CD}{n_2^3} \tag{11.44}$$

and

$$SE_{RD} = \sqrt{V_{RD}}. \tag{11.45}$$

Using the data in table 11.7, the risk difference *RD* is

$$RD = \frac{5}{100} - \frac{10}{100} = -0.05,$$

with approximate variance and standard error

$$V_{RD} = \frac{5 \times 95}{100^3} + \frac{10 \times 90}{100^3} = 0.0014$$

and

$$SE_{RD} = \sqrt{0.0014} = 0.0374.$$

The risk difference can take any value from –1 to +1. A risk difference of 0 indicates that the two risks are equal. Risk differences above 0 indicate that the risk in group 1 is higher than the risk in group 2, and risk differences below 0 indicate that the risk in group 1 is lower than the risk in group 2.

The direction of effect is arbitrary, except that it must be consistent from one study to the next. That is, if a risk difference above 0 in study A indicates that group 1 did better, then a risk difference above 0 in studies B, C, . . . must also indicate that group 1 did better.

When the event rate in either (or both) groups is zero we can compute the risk difference but cannot compute the variance. Therefore, we compute the risk difference using the original data. Then we add the value 0.5 to each cell, and use these modified values to compute the variance. The same rule applies when the event rate in either (or both) groups is 100 percent.

Table 11.8 provides formulas for computing the risk difference and its variance, working from information that may be reported in a published paper.

11.4.2 Risk Ratio

The risk ratio (λ) (also called the relative risk) is defined as the ratio of the probabilities (or risks) in the two groups. Let π_1 be the risk of an event in group 1, and π_2 be the risk of the event in group 2. Then the risk ratio λ, is defined as

$$\theta = \frac{\pi_1}{\pi_2}. \tag{11.46}$$

Whereas the effect-size λ is the true (population) value, any study yields an effect-size *RR* that is based on the observed risks and is an estimate of the true effect size. To estimate *RR*, we can use

$$RR = \frac{p_1}{p_2}. \tag{11.47}$$

Using the notation in table 11.6, we estimate *RR* as

$$RR = \frac{A/n_1}{C/n_2}. \tag{11.48}$$

We do not use the risk ratio itself in the computations. Instead, we convert the risk ratio to log units, perform all computations using the log units, and then convert the results back into risk ratio units for the report.

The log of the risk ratio is

$$\ln RR = \ln(RR), \tag{11.49}$$

Table 11.8 Computing Risk Difference, Independent Groups in Prospective Study

Reported	Computation of Needed Quantities
A, B, C, D	$RD = \frac{A}{n_1} - \frac{C}{n_2}, V_{RD} = \frac{AB}{n_1^3} + \frac{CD}{n_2^3}$
p_1, p_2, n_1, n_2	$RD = P_1 - P_2, V_{RD} = \frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}$
$RD, UD_{RD}, LL_{RD}, Cl_{Level}$	$RD = \text{Given}, V_{RD} = \left(\frac{UL_{RD} - LL_{RD}}{2Z}\right)^2$ or $\left(\frac{UL_{RD} - RD}{Z}\right)^2$ or $\left(\frac{RD - LL_{RD}}{Z}\right)^2$

SOURCE: Authors' tabulation.

NOTE: The cells (*A*, *B*, *C*, *D*) are defined in table 11.6. In row 1, if any cell (*A*, *B*, *C*, or *D*) has a value of zero, use the original values to compute the risk difference. Then add 0.5 to all cells, and use these modified values to compute the variance. In row 2, if $P_1 = 0$, $P_1 = 1$, $P_2 = 0$, or $P_2 = 1$, use the original values to compute the risk difference. Then, replace P_1 with $\frac{P_1 n_1 + 0.5}{n_1 + 1}$, n_1 with $n_1 + 1$, P_2 with $\frac{P_2 n_2 + 0.5}{n_2 + 1}$, and n_2 with $n_2 + 1$, and use these modified values to compute the variance. In row 3, for the 95 percent confidence interval, the *Z*-value would be 1.96.

with approximate variance and standard error

$$V_{lnRR} = \frac{1}{A} - \frac{1}{n_1} + \frac{1}{C} - \frac{1}{n_2} \tag{11.50}$$

and

$$SE_{\ln RR} = \sqrt{V_{\ln RR}}. \tag{11.51}$$

After we compute the mean effect size in log units we convert that value to risk ratio units using

$$RR = \exp(\ln RR). \tag{11.52}$$

This formula is applied to all the associated statistics (such as the bounds of the confidence interval) as well.

Using the data in table 11.7, the risk ratio *RR* is

$$RR = \frac{5/100}{10/100} = 0.500.$$

In log units, the risk ratio is

$$lnRR = \ln(0.500) = -0.69,$$

with variance and standard error

$$V_{lnRR} = \frac{1}{5} - \frac{1}{100} + \frac{1}{10} - \frac{1}{100} = 0.28$$

and

$$SE_{lnRR} = \sqrt{0.28} = 0.5291.$$

The risk ratio can take any value greater than 0. A risk ratio of 1 indicates that the two risks are equal. Risk ratios above 1 indicate that the risk in group 1 is higher than the risk in group 2, and risk ratios below 1 indicate that the risk in group 1 is lower than the risk in group 2.

The direction of effect is arbitrary, except that it must be consistent from one study to the next. That is, if a risk ratio less than 1 in study A indicates that group 1 did better, then a risk ratio less than 1 in studies B, C, . . . must also indicate that group 1 did better.

The decision about which outcome is the event and which is the nonevent can have major implications for the risk ratio. For example, the treatment may appear to have a major impact on the risk of dying but almost no impact on the risk of surviving. We discuss this later in the chapter.

When one cell has zero events we add 0.5 to all four cells, and use these adjusted numbers to compute both the effect size and the variance (for other options, see Sweeting, Sutton, and Lambert 2004).

When two cells have zero events, the study provides no useful information; we therefore omit it from the analysis. This might seem counterintuitive, given that a study with zero events in both groups might seem to suggest that the risk is comparable in the two groups. In fact, though, this outcome is compatible with a risk ratio from near zero to near infinity, and so the only option is to exclude the study from the analysis. (Other options exist in a Bayesian analysis, but that is beyond the scope of this chapter.)

The same idea applies when all subjects (rather than none) have the event. If 100 percent of subjects in one group have the event, we add 0.5 to all cells. If 100 percent of subjects in both groups have the event, we exclude the study from the analysis.

Table 11.9 provides formulas for computing the risk ratio and its variance, working from information that may be reported in a published paper.

11.4.3 Odds Ratio

The odds ratio (ω) is defined as the ratio of the odds in the two groups, where the odds in any group is the ratio of an event to a nonevent in that group.

Let κ_1 be the risk of the event in group 1. The odds in group 1 are defined as $\kappa_1/(1 - \kappa_1)$. Similarly, let κ_2 be the risk of the event in group 2. The odds in group 2 are defined as $\kappa_2/(1 - \kappa_2)$.

Then, the odds ratio is defined as

$$\omega = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}, \tag{11.53}$$

which is computationally identical to

$$\omega = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}. \tag{11.54}$$

Whereas the effect-size ω is the true (population) value, any study yields an effect-size *OR* that is based on

Table 11.9 Computing Risk Ratio, Independent Groups in Prospective Study

Reported	Computation of Needed Quantities
A, B, C, D	$RR = \frac{A/n_1}{C/n_2}, lnRR = \ln(RR) V_{lnRR} = \frac{1}{A} - \frac{1}{n_1} + \frac{1}{C} - \frac{1}{n_2}$
p_1, p_2, n_1, n_2	$RR = \frac{P_1}{P_2} lnRR = \ln(RR) V_{lnRR} = \frac{1-P_1}{n_1 P_1} + \frac{1-P_2}{n_2 P_2}$
$RR, UD_{RD}, LL_{RD}, Cl_{Level}$	$RR = \text{Given } \ln RR = \ln(RR)\ LL_{lnRR} = \ln(LL_{RR})\ UL_{lnRR} = \ln(\text{UL}_{RR})$
	$V_{lnRR} = \left(\frac{UL_{lnRR} - LL_{lnRR}}{2Z}\right)^2$ or $\left(\frac{UL_{lnRR} - lnRR}{Z}\right)^2$ or $\left(\frac{lnRR - LL_{lnRR}}{Z}\right)^2$

SOURCE: Authors' tabulation.

NOTE: The cells (*A*, *B*, *C*, *D*) are defined in table 11.6. Note that we do not compute a variance for the risk ratio. Rather, all calculations are carried out on the log values. In row 1, if $A = 0$ and $C = 0$, or if $B = 0$ and $D = 0$, then the study carries no information about the risk ratio, and the study would be excluded from the meta-analysis. In row 1, if any cell (*A*, *B*, *C*, or *D*) has a value of zero, add 0.5 to all cells, and use these modified values to compute the risk ratio, log risk ratio, and variance. In row 2, If $P_1 = 0$, $P_1 = 1$, $P_2 = 0$, or $P_2 = 1$, replace P_1 with $\frac{P_1 n_1 + 0.5}{n_1 + 1}$, n_1 with $n_1 + 1$, P_2 with $\frac{P_2 n_2 + 0.5}{n_2 + 1}$, and n_2 with $n_2 + 1$, and use these modified values to compute the risk ratio, log risk ratio, and variance. In row 2, if $P_1 = 0$ and $P_2 = 0$, then the study carries no information about the risk ratio, and the study would be excluded from the meta-analysis. In row 3, for the 95 percent confidence interval, the *Z*-value would be 1.96.

the observed risks and is an estimate of the true effect size. To estimate *OR* we can use

$$OR = \frac{\mathrm{p}_1/(1-\mathrm{p}_1)}{\mathrm{p}_2/(1-\mathrm{p}_2)} \tag{11.55}$$

or

$$OR = \frac{\mathrm{p}_1(1-\mathrm{p}_2)}{\mathrm{p}_2(1-\mathrm{p}_1)}. \tag{11.56}$$

Using the notation in table 11.6, we estimate *OR* as

$$OR = \frac{(A/n_1)(1-C/n_2)}{(C/n_2)(1-A/n_1)} = \frac{A(n_2-C)}{C(n_2-A)} = \frac{AD}{BC}. \tag{11.57}$$

We do not use the odds ratio itself in the computations. Instead, we convert the odds ratio to log units, perform all computations using the log units, and then convert the results back into odds ratio units for the report.

The log of the odds ratio is

$$\ln OR = \ln(OR), \tag{11.58}$$

with approximate variance and standard error

$$V_{lnOR} = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D} \tag{11.59}$$

and

$$SE_{\ln OR} = \sqrt{V_{\ln OR}}. \tag{11.60}$$

After we compute the mean effect size in log units we convert that value to odds ratio units using

$$OR = \exp(\ln OR). \tag{11.61}$$

This formula is applied to the all associated statistics (such as the bounds of the confidence interval) as well.

Using the data in table 11.7, the odds ratio *OR* is

$$OR = \frac{5 \times 90}{10 \times 95} = 0.474.$$

In log units, the odds ratio is

$$lnOR = \ln(0.474) = -0.75,$$

with variance and standard error

$$V_{lnOR} = \frac{1}{5} + \frac{1}{95} + \frac{1}{10} + \frac{1}{90} = 0.32$$

and

$$SE_{lnOR} = \sqrt{0.32} = 0.5657.$$

As was true for the risk ratio, the odds ratio can take any value greater than 0. An odds ratio of 1 indicates that the two odds are equal. Odds ratios above 1 indicate that the odds in group 1 is higher than the odds in group 2, and odds ratios below 1 indicate that the odds in group 1 are lower than the odds in group 2.

The direction of effect is arbitrary, except that it must be consistent from one study to the next. That is, if an odds ratio less than 1 in study A indicates that group 1 did better, then an odds ratio less than 1 in studies B, C, . . . must also indicate that group 1 did better.

When one cell has zero events, we add 0.5 to all four cells, and use these adjusted numbers to compute both the effect size and the variance (for other options, see Sweeting, Sutton, and Lambert 2004).

When two cells have zero events, the study provides no useful information; we therefore omit it from the analysis. This might seem counterintuitive, given that a study with zero events in both groups might seem to suggest that the odds are comparable in the two groups. In fact, though, this outcome is compatible with an odds ratio from near zero to near infinity, and so the only option is to exclude the study from the analysis. (Other options exist in a Bayesian analysis, which is beyond the scope of this chapter.)

The same idea applies when all subjects (rather than none) have the event. If 100 percent of subjects in one group have the event, we add 0.5 to all cells. If 100 percent of subjects in both groups have the event, we exclude the study from the analysis.

When many studies in the analysis have zero events in one group, it may be preferable to use an alternate method for computing the variance of the odds ratio, which is beyond the scope of this chapter (Altman, Deeks, and Sackett 1998).

Table 11.10 provides formulas for computing the odds ratio and its variance, working from information that may be reported in a published paper.

11.4.4 Direction of the Effect

The direction of the effect is arbitrary except that it must be consistent from one study to the next. That is, if a risk difference above 0 in study A indicates that group 1 did

Table 11.10 Computing Odds Ratio, Independent Groups in a Prospective Study

Reported	Computation of Needed Quantities
A, B, C, D	$OR = \frac{AD}{BC}, lnOR = \ln(OR)\, V_{lnOR} = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}$
p_1, p_2, n_1, n_2	$OR = \frac{P_1(1-P_2)}{P_2(1-P_1)}\, lnOR = \ln(OR)\, V_{lnOR} = \frac{1}{n_1P_1} + \frac{1}{n_1(1-P_1)} + \frac{1}{n_2P_2} + \frac{1}{n_2(1-P_2)}$
$OR, UD_{OR}, LL_{OR}, Cl_{Level}$	$OR = \text{Given}\ lnOR = \ln(OR)\ LL_{lnOR} = \ln(LL_{OR})\ UL_{\ln OR} = \ln(\text{UL}_{\text{OR}})$
	$V_{lnOR} = \left(\frac{UL_{lnOR} - LL_{lnOR}}{2Z}\right)^2$ or $\left(\frac{UL_{lnOR} - lnOR}{Z}\right)^2$ or $\left(\frac{lnOR - LL_{lnOR}}{Z}\right)^2$

SOURCE: Authors' tabulation.

NOTE: The cells (*A*, *B*, *C*, *D*) are defined in table 11.6. Note that we do not compute a variance for the odds ratio. Rather, all calculations are carried out on the log values. If any cell (*A*, *B*, *C*, or *D*) has a value of zero, add 0.5 to all cells, and use these modified values to compute the odds ratio, log odds ratio, and variance. If $A = 0$ and $C = 0$, or if $B = 0$ and $D = 0$, then the study carries no information about the odds ratio, and the study would be excluded from the meta-analysis. If $P_1 = 0$, $P_1 = 1$, $P_2 = 0$, or $P_2 = 1$, replace P_1 with $\frac{P_1n_1 + 0.5}{n_1 + 1}$, n_1 with $n_1 + 1$, P_2 with $\frac{P_2n_2 + 0.5}{n_2 + 1}$, and n_2 with $n_2 + 1$, and use these modified values to compute the odds ratio, log odds ratio, and variance. If $P_1 = 0$ and $P_2 = 0$, or if $P_1 = 1$ and $P_2 = 1$, then the study carries no information about the odds ratio, and the study would be excluded from the meta-analysis. In row 3, for the 95 percent confidence interval, the *Z*-value would be 1.96.

better, then a risk difference above 0 in studies B, C, . . . must also indicate that group 1 did better. The same would apply to a risk ratio (above or below 1) and an odds ratio (above or below 1). Although the direction is arbitrary, some useful conventions make it easier to interpret the results.

In table 11.6, we have put treatment before control, and events before nonevents. When we structure the table this way, the value in cell A will be the number of events in the treatment group.

When we follow this convention, the direction of the effect will match the following pattern. If the treatment reduces the probability of the event, the risk difference will be negative, the risk ratio will be less than 1, and the odds ratio will be less than 1. If the treatment increases the probability of the event, the risk difference will be positive, the risk ratio will be greater than 1, and the odds ratio will be greater than 1.

As a general rule, following this convention is advisable because the results will follow the usual pattern. For example, if the treatment is intended to prevent death, we expect to see the risk ratio to the left of 1. Similarly, if the treatment is intended to promote good breathing, we expect to see the risk ratio to the right of 1.

Although the convention is therefore useful, it is arbitrary. If we choose to reverse the position of the two rows (the groups), keeping the columns unchanged, the effect-size format would change but the substantive meaning would remain the same. Concretely, if cell A was events in the control group rather than events in the treatment group, the three effects sizes would be modified as follows.

If we switched the position of the two groups, the new risk difference would be

$$RD_{New} = -RD_{Old}. \tag{11.62}$$

In our example, the risk difference would switch from –0.05 to +0.05. Rather than saying that the risk in one group is 5 points lower, we would be saying that the risk in the other group is 5 points higher—which is essentially the same thing.

If we switched the position of the two groups, the new risk ratio would be

$$RR_{New} = \frac{1}{RR_{Old}}. \tag{11.63}$$

In our example, the risk ratio would switch from 0.50 to 2.0. Rather than saying that risk in one group is half as high, we would be saying that the risk in the other group is twice as high—which is essentially the same thing

If we switched the position of the two groups, the new odds ratio would be

$$OR_{New} = \frac{1}{OR_{Old}}. \tag{11.64}$$

In our example, the odds ratio would switch from 0.47 to 2.11. In round numbers, rather than saying that odds in one group is half as high, we would be saying that the odds in the other group is twice as high—which is essentially the same thing.

That the decision to put one group or the other on top is arbitrary is especially important when we are comparing two active treatments rather than a treatment and a control. If some studies put treatment 1 in the top row for the computations, and others put treatment 2 on the top row, we need to ensure that all the results are in the same direction before we enter them into the meta-analysis. We could apply these formulas to the effect size and confidence interval for any study to reverse the direction of the effect.

Although it is thus a simple matter to reverse the direction of the effect when some studies have reversed the position of the rows, the situation is more complicated when it comes to the columns. We discuss that in the following section.

11.4.5 What Is an Event?

As noted, we speak about the risk of an event, but we are actually talking about the likelihood of an event, and the event can be something that we would prefer to avoid or something that we would prefer to see happen. If we are looking at an intervention to prevent death, the event can be either death or survival. If we are looking at an intervention to keep students in school, the event can be either that the student drops out or graduates. So, we need to give some thought into what we mean by an event.

In table 11.7 we presented results for a fictional study. The outcome was mortality, and we chose to define the event as death. Table 11.11 shows the same data, but this time we chose to define the event as being alive. Note that alive now appears in column 1, and corresponds to cells A and C rather than cells B and D in the table.

Each group includes one hundred patients. Because ninety-five survivors are in the treated group and ninety

Table 11.11 Event Is "Alive" Rather than "Dead"

	Alive	Dead	Total
Treated	95	5	100
Control	90	10	100

SOURCE: Authors' tabulation.

are in the control group, the risks in the two groups are estimated as

$$p_1 = \frac{95}{100} = 0.95,\ p_2 = \frac{90}{100} = 0.90.$$

The risk difference would be

$$RD = \frac{95}{100} - \frac{90}{100} = +0.05.$$

In our example, the risk difference would switch from –0.05 to +0.05. Rather than saying that the risk of death in the treated group is 5 points lower, we would be saying that the probability of survival in the treated group is 5 points higher—which is essentially the same thing. In fact, we can convert between the *RD* in two tables by using the formula

$$RD_{Survival} = -RD_{Death}. \tag{11.65}$$

We will skip over the risk ratio for a moment and proceed to the odds ratio. If we chose to define survival (rather than death) as the event, the new odds ratio would be

$$OR = \frac{95 \times 10}{5 \times 90} = 2.11.$$

This is related to the earlier value by the formula

$$OR_{Survival} = \frac{1}{OR_{Death}}. \tag{11.66}$$

In our example, the odds ratio would switch from 0.47 to 2.11. In round numbers, rather than saying that odds in of dying in the treated group is half as high, we would be saying that the odds of surviving in the treated group is twice as high—which is essentially the same thing.

The situation is different for the risk ratio. Using the data in table 11.7, if we choose death as the event, the risk ratio is

$$RR = \frac{5/100}{10/100} = 0.500.$$

By contrast, if we choose survival as the event (table 11.11), the risk ratio is

$$RR = \frac{95/100}{90/100} = 1.056.$$

The first approach tells us that the treatment reduces the risk of death by 50 percent; the second tells us that the treatment increases the probability of survival by about 5 percent. These are clearly not the same thing. Thus, how we choose to define the event can have a substantial impact on the magnitude of the effect size.

Less obvious but also important for the risk ratio, the definition of event also affects the weight assigned to each study. Note that cells A and C (reflecting the event) appear twice in the formula for the variance (as themselves and also as components of the sample size). By contrast, cells B and D appear only once (as components of the sample size). A study that gets a certain amount of weight when we assess the impact of treatment on death could thus get less (or more) weight if we assess the impact of treatment on survival.

A consensus on whether we should be using one outcome or the other as the event is common. However, when a consensus is lacking, we need to be aware of the potential pitfall and might want to avoid using the risk ratio. If we do use it, we should be clear about the reason for choosing one outcome or the other. We should also be explicit that this has an impact on the results.

In sum, we should generally follow the convention of putting treated before control and events before nonevents, as shown in table 11.6. When we follow this convention, the direction of results will follow an expected pattern and be more intuitive.

If some studies computed the effect size by switching the sequence of rows, we will need to switch the direction of the effect. We can do this using equation (11.62) for the risk difference, (11.63) for the risk ratio, or (11.64) for the odds ratio.

If some studies have switched the polarity of the event (using survival rather than death as the event), we will

also need to switch the direction of the effect. We can do this using formula (11.62) for the risk difference or formula (11.64) for the odds ratio. However, for the risk ratio, we cannot use (11.63). Instead, we would need to have enough information to reconstruct the 2×2 table, and then compute the new risk ratio from that.

11.4.6 Choosing Among Indices

To this point, we have introduced three indices for studies that compare risks in two groups. In any given analysis, we need to choose among them.

We start by considering the risk difference and the risk ratio. The risk difference is an absolute measure and the risk ratio is a relative measure. The risk difference is thus quite sensitive to baseline risk and the risk ratio less so.

Consider a series of studies that assess the utility of a vaccine to prevent a mosquito-borne illness. In this example, we assume that the baseline risk of the disease ranges from 80 percent in countries where the mosquito is prevalent to 2 percent in locations where the mosquito is rare. We also assume that the vaccine consistently reduces the risk of infection by 50 percent. Table 11.12 shows the results for a series of five fictional studies based on these assumptions.

In a high-risk population (first row), the vaccine reduces the risk of infection from 80 percent to 40 percent. The risk ratio is 0.50, and the risk difference is 0.40. In a low-risk population (last row), the vaccine reduces the risk of infection from 2 percent to 1 percent. The risk ratio is 0.50, and the risk difference is 0.01.

In this example, the risk ratio is unaffected by the baseline risk: it is 0.50 in all the populations. In contrast, the risk difference is strongly affected by the baseline risk, varying from –0.40 in the high-risk population to –0.01 in the low-risk population. Which of these numbers is more relevant? The answer depends on our goals. If we want to talk about the utility of the vaccine in general, the relevant number is the risk ratio, precisely because it is not strongly affected by baseline risk. Conversely, if we want to talk about the utility of the vaccine for a specific person, then the relevant number is the risk difference, precisely because it is strongly related to the baseline risk. We might also want to report both, to address both goals.

If we were working with a single study, the choice of an effect size would end there. In contrast, when we are working with all these studies, in a meta-analysis, the process is a little more complicated. In this example, the risk difference shows substantial variation, and the risk ratio shows none. Consider how this affects our options.

Case 1. If we want to report the risk ratio, it is clear that we should run the analysis using the risk ratio. We would get a precise estimate of the effect (precise because there's no variation in the effect size). The summary effect would be especially useful given the lack of variation in the effect (it applies consistently).

Case 2. If we want to report the risk difference, the obvious option would be to run the analysis using the risk difference. This would probably be the preferred option if the baseline risk were consistent across studies. In contrast, for the example in table 11.12, it might make more sense to run the analysis using the risk ratio, and then to use the summary effect size to predict the risk difference for any given baseline risk. Concretely, the risk difference is given by

$$RD = p_c(1 - RR)(-1), \tag{11.67}$$

where p_c is the risk in the control group. If the risk ratio is 0.50, then for a population where the baseline risk is 0.20, the risk difference would be

$$RD = 0.20(1 - 0.50)(-1) = -0.10. \tag{11.68}$$

Table 11.12 Fictional Studies, Baseline Risk Varies and Risk Ratio Constant

Study	Risk (Treated)	Risk (Control)	Risk Difference	Risk Ratio	Odds Ratio	NNT
1	0.40	0.80	–0.40	0.50	0.167	2.5
2	0.20	0.40	–0.20	0.50	0.375	5.0
3	0.10	0.20	–0.10	0.50	0.444	10.0
4	0.05	0.10	–0.05	0.50	0.474	20.0
5	0.01	0.02	–0.01	0.50	0.495	100.0

SOURCE: Authors' tabulation.

This approach makes sense in this example because the risk ratio is relatively constant and the baseline risk varies across studies. In an analysis where either condition does not hold true, it might be preferable to work directly with the risk difference. The key point is that it is important think about the choice of an effect size in this context.

Until now, we have focused on the risk difference and the risk ratio. Researchers and clinicians tend to prefer these two indices because they are relatively intuitive relative to the odds ratio. In contrast, epidemiologists tend to prefer the odds ratio because it has useful statistical properties. Unlike the risk ratio and the risk difference, the odds ratio is not constrained by base rate and (regardless of the base rate) can take on any value greater than 0. It is possible to compute a meaningful odds ratio from a case-control study as well as from a prospective study. This is especially important in a meta-analysis, where we might be working with both types of studies. It may be possible to incorporate odds ratios from logistic regression. Finally, when we are working with the odds ratio, the decision about how to define the event (success or failure) is not critical (for additional insights into effect sizes for binary outcomes, see Grissom and Kim 2012; Borenstein et al. 2009; Rothman 2012; Sanchez-Meca, Marin-Martinez, and Moscoso 2003; Grant 2014; Deeks and Altman 2001; Fleiss, Levin, and Paik 2003; Deeks 2002).

11.4.7 Odds and Risk Ratios in the Same Analysis

When the event rate is low (less than 10 percent), the risk ratio and odds ratio tend to have comparable values. Our example, when the event rate is 5 percent in one group and 10 percent in the other, is a case in point. The risk ratio is 0.50 and the odds ratio is 0.47, and these values are fairly close to each other. For this reason, some researchers assume that it is acceptable to include both indices in the same analysis. In fact, though, including risk ratios and odds ratios in the same analysis is a mistake because the two indices are measuring fundamentally different things. If we are going to report on the magnitude of the effect, we need to have a clear narrative of what that effect means. The summary effect can be the ratio of two risks or of two odds. But the same number cannot represent both because they are not the same thing (Altman, Deeks, and Sackett 1998; Zhang and Yu 1998).

Although it is always a mistake to include the two indices in the same analysis, the ramifications of the mistake will vary. In some cases, such as study 5 in table 11.12, the values may be comparable, and thus the results may be reasonably correct. In other cases, study 1 in table 11.12, the values will differ substantially and the results would be meaningless.

We cannot simply insert an odds ratio into an analysis as though it were a risk ratio. However, we may be able to convert an odds ratio to a risk ratio and then use the risk ratio in the analysis. If we start with the odds ratio and we know the baseline risk, we can compute the risk ratio using

$$RR = \frac{OR}{(1 - p_c) + (OR \times p_c)}. \tag{11.69}$$

We would then apply the same formula to the lower and upper limits of the confidence interval to get the lower and upper limit for the risk ratio confidence interval. Table 11.9 shows how to proceed from there.

11.4.8 Risk and Hazard Ratios in the Same Analysis

Researchers sometimes ask whether it is acceptable to combine risk ratios and hazard ratios in the same analysis. The risk ratio is based on the risk of an event during a given time span. The hazard ratio is based on that risk but also takes into account the timing of the event within the time span. In many cases, it would make sense to say that the two are measuring the same thing, albeit with different levels of nuance. In these cases, it would make sense to include both in the same analysis. The only provision is that then we would compute the variance of each index using the formulas that are appropriate for that index (for details on computing the variance for hazard ratios, see Parmar, Torri, and Stewart 1998; Michiels et al. 2005).

11.4.9 Number Needed to Treat

When we compare the risk of an event in two groups, a useful index for expressing the utility of a treatment is the number needed to treat (NNT). An NNT of 2 tells us that we expect one person to benefit for every two people who are treated. An NNT of 100 tells us that we expect one person to benefit for every hundred people who are treated.

We cannot perform an analysis directly using the NNT. Instead, we could perform an analysis using the risk difference to obtain the mean *RD* along with lower and upper confidence limits. Then we would convert each of these values to an NNT.

The NNT would be estimated using

$$\text{NNT} = \frac{1}{|\text{RD}|}. \tag{11.70}$$

Like the risk difference, the NNT it is strongly related to the baseline risk. In table 11.12, the baseline risk for the first study is 0.80, the risk difference is –0.40, and the NNT is

$$\text{NNT} = \frac{1}{|-0.40|} = 2.5.$$

By contrast, the baseline risk for the last study is 0.02, the risk difference is –0.40, and the NNT is

$$\text{NNT} = \frac{1}{|-0.01|} = 100.$$

For this reason, if we do elect to report the NNT, we would do so in the context of the baseline risk.

It is common to report the NNT when the analysis is based on the risk difference, but we might also report it if the analysis is based on a risk ratio. If the risk ratio is relatively constant across baseline risks, we could use the risk ratio to compute the risk difference for any given baseline risk, using formula (11.67). Then we could use (11.70) to compute the corresponding NNT.

When used properly, the NNT can be an effective tool for communicating the risks and benefits of treatment. However, as with any effect-size index, proper use is critical (for a more complete discussion of the NNT, see Altman and Deeks 2002; Alderson and Deeks 2004; Altman 1998; Cook and Sackett 1995; Chatellier et al. 1996; D'Amico, Deeks, and Altman 1999; Ebrahim 2001; McAlister 2001, 2008; Stang, Poole, and Bender 2010; and Wen, Badgett, and Cornell 2005).

11.5 INDEPENDENT GROUPS FOR A RETROSPECTIVE (CASE-CONTROL) STUDY

When we are working with prospective studies, we can choose to use the risk difference, the risk ratio, or the odds ratio, as discussed. By contrast, when we are working with retrospective (case-control) studies the only appropriate index is the odds ratio, as discussed here.

Table 11.13 Cells for a Case-Control Study

	Cases	Controls
Exposed	A	B
Unexposed	C	D
Total	m_1	m_2

SOURCE: Authors' tabulation.

A retrospective study with two groups (cases and controls) and a binary exposure can be represented as in table 11.13, and data for a fictional study in table 11.14.

The sample sizes in a case-control study are defined as the number of cases and the number of controls rather than the numbers exposed (for example, to treatment) or unexposed. The mathematical computations are identical, but the meaning of some of the numbers is different.

The odds ratio may be computed using

$$OR = \frac{AD}{BC}. \tag{11.71}$$

In log units, the odds ratio is

$$lnOR = \ln(OR), \tag{11.72}$$

with approximate variance and standard error

$$V_{lnOR} = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D} \tag{11.73}$$

Table 11.14 Data from Fictional Case-Control Study

	Cases	Controls
Exposed	25	20
Unexposed	75	80
Total	m_1	m_2

SOURCE: Authors' tabulation.

and

$$SE_{lnOR} = \sqrt{V_{lnOR}}. \tag{11.74}$$

After we compute the mean effect size in log units we convert that value to odds ratio units using

$$OR = \exp(\ln OR) \tag{11.75}$$

This formula is applied to all the associated statistics (such as the bounds of the confidence interval) as well.

Using the data in table 11.7, the odds ratio is

$$OR = \frac{25 \times 80}{20 \times 75} = 1.333.$$

In log units, the odds ratio is

$$lnOR = \ln(1.333) = 0.29,$$

with approximate variance and standard error

$$V_{lnOR} = \frac{1}{25} + \frac{1}{20} + \frac{1}{75} + \frac{1}{80} = 0.12$$

and

$$SE_{lnOR} = \sqrt{0.12} = 0.34.$$

Table 11.15 shows how to compute the odds ratio from some of the statistics that may be reported for a retrospective study.

In the section on odds ratios for prospective studies, we discuss the issues of how to change the direction of the effect, how to define an event, and how to deal with empty cells in the 2×2 table. The same rules apply here.

Although the odds ratio has the same technical meaning when based on a prospective or a retrospective study, the study design still determines what we can learn from the index. A prospective randomized study can be generally be used to test for a causal relationship, but a retrospective study generally cannot.

11.6 CONVERTING EFFECT SIZES

Suppose that some studies report the means for two groups, and we compute a standardized mean difference. Other studies report the risks in two groups, and

Table 11.15 Computing Odds Ratio, Independent Groups in Retrospective Study

Reported	Computation of Needed Quantities
A, B, C, D	$OR = \frac{AD}{BC}\ lnOR = \ln(OR)\ V_{lnOR} = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}$
p_1, p_2, m_1, m_2	$OR = \frac{p_1(1-p_2)}{p_2(1-p_1)}\ lnOR = \ln(OR)\ V_{lnOR} = \frac{1}{m_1 p_1} + \frac{1}{m_2 p_2} + \frac{1}{m_1(1-p_1)} + \frac{1}{m_2(1-p_2)}$
$OR, UD_{OR}, LL_{OR}, Cl_{Level}$	$OR = \text{Given}\ lnOR = \ln(OR)\ LL_{lnOR} = \ln(LL_{OR})\ \text{UL}_{\text{lnOR}} = \ln(\text{UL}_{\text{OR}})$
	$V_{lnOR} = \left(\frac{UL_{lnOR} - LL_{lnOR}}{2Z}\right)^2$ or $\left(\frac{UL_{lnOR} - lnOR}{Z}\right)^2$ or $\left(\frac{lnOR - LL_{lnOR}}{Z}\right)^2$

SOURCE: Authors' tabulation.

NOTE: The cells (A, B, C, D) are defined in table 11.13. Note that we do not compute a variance for the odds ratio. Rather, all calculations are carried out on the log values. If any cell (A, B, C, or D) has a value of zero, add 0.5 to all cells, and use these modified values to compute the odds ratio, log odds ratio, and variance. If $A = 0$ and $C = 0$, or if $B = 0$ and $D = 0$, then the study carries no information about the odds ratio, and the study would be excluded from the meta-analysis. If $p_1 = 0$, $p_1 = 1$, $p_2 = 0$, or $p_2 = 1$, replace p_1 with $\frac{P_1 m_1 + 0.5}{m_1 + 1}$, m_1 with $m_1 + 1$, p_2 with $\frac{P_2 m_2 + 0.5}{m_2 + 1}$, and m_2 with $m_2 + 1$, and use these modified values to compute the odds ratio, log odds ratio, and variance. If $p_1 = 0$ and $p_2 = 0$, or if $p_1 = 1$ and $p_2 = 1$, then the study carries no information about the odds ratio, and the study would be excluded from the meta-analysis. In row 3, for the 95 percent confidence interval, the Z-value would be 1.96.

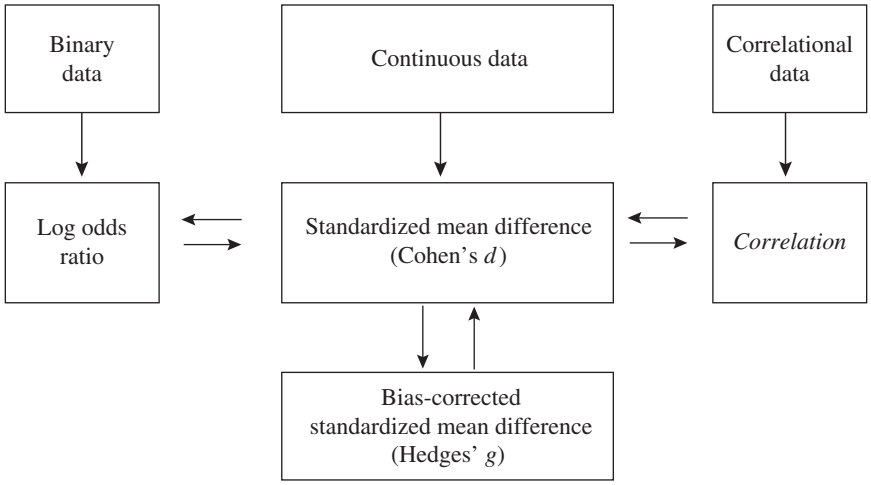

Figure 11.2 Schematic of Effect-Size Index Conversion Process

SOURCE: Authors' tabulation.
NOTE: This schematic outlines the mechanism for incorporating multiple kinds of data in the same meta-analysis. First, each study is used to compute an effect size and variance in its "native" index—log odds ratio for binary data, *d* for continuous data, and *r* for correlational data. Then, we convert all of these indices to a common index, which would be either the log odds ratio, *d*, or *r*. If the final index is *d*, we can move from there to Hedges' *g*. This common index and its variance are then used in the analysis.

we compute an odds ratio. Still others report a correlation. Is it possible to include all of these studies in the same analysis? Technically, it is possible to include all of these studies in the same analysis. In this section, we provide formulas that allow us to convert any of these indices to a common index. We therefore could convert all the effects to a standardized mean difference (or odds ratio, or correlation) and then run the analysis using this index (see figure 11.2).

Although no technical barrier prevents including all the studies in the same analysis, we need to ask whether this is a good idea. Later, we show how to approach that question.

11.6.1 Log Odds Ratio to *d*

We can convert from a log odds ratio [ln(*o*)] to the standardized mean difference *d* using

$$d = \frac{\ln(OR)\sqrt{3}}{\pi}, \tag{11.76}$$

where π is the mathematical constant (approximately 3.14159). The variance of *d* would then be

$$V_d = \frac{3V_{\ln(OR)}}{\pi^2}, \tag{11.77}$$

where $V_{\ln(o)}$ is the variance of the log odds ratio. This method was originally proposed in 1995 but variations have been proposed since then (Hasselblad and Hedges 1995; Sanchez-Meca, Marin-Martinez, and Moscoso 2003; Whitehead 2002; Chinn 2000).

For example, if the log odds ratio were $\ln(o) = 0.9070$ with a variance of $v\ln(o) = 0.0676$, then

$$d = \frac{0.9070\sqrt{3}}{3.1416} = 0.5000,$$

with variance

$$v_d = \frac{3 \times 0.0676}{3.1416^2} = 0.0205.$$

11.6.2 *d* to Log Odds Ratio

We can convert from the standardized mean difference d to the log odds ratio [ln(o)] using

$$\ln(OR) = \frac{\pi d}{\sqrt{3}}, \tag{11.78}$$

where π is the mathematical constant (approximately 3.14159). The variance of ln(o) would then be

$$V_{\ln(OR)} = \frac{\pi^2 v_d}{3}. \tag{11.79}$$

For example, if $d = 0.5000$ and $V_d = 0.0205$ then

$$\ln(OR) = \frac{3.1415 * 0.5000}{\sqrt{3}} = 0.9069$$

and

$$v_{\ln(o)} = \frac{3.1415^2 * 0.0205}{3} = 0.0674.$$

11.6.3 Converting from *r* to *d*

We convert from a correlation (r) to a standardized mean difference (d) using

$$d = \frac{2r}{\sqrt{1-r^2}}, \tag{11.80}$$

and the variance of d computed in this way (converted from r) is

$$v_d = \frac{4v_r}{(1-r^2)^3}. \tag{11.81}$$

For example, if $r = 0.50$ and $V_r = 0.0058$, then

$$d = \frac{2 \times 0.500}{\sqrt{1-0.500^2}} = 1.1547,$$

and the variance of d is

$$v_d = \frac{4 \times 0.0058}{(1-0.50^2)^3} = 0.0550.$$

11.6.4 Converting from *d* to *r*

We can convert from a standardized mean difference (d) to a correlation (r) using

$$r = \frac{d}{\sqrt{d^2 + a}}, \tag{11.82}$$

where a is a correction factor for cases where $n_1 \neq n_2$,

$$a = \frac{(n_1 + n_2)^2}{n_1 n_2}. \tag{11.83}$$

The correction factor a is based on the ratio of n_1 to n_2, rather than the absolute values of these numbers. Therefore, if n_1 and n_2 are not known precisely, use $n_1 = n_2$, which will yield $a = 4$. The variance of r computed in this way (converted from d) is

$$v_r = \frac{a^2 v_d}{(d^2 + a)^3}. \tag{11.84}$$

For example, if $n_1 = n_2$, $d = 1.1547$ and $V_d = 0.0550$, then

$$r = \frac{1.1547}{\sqrt{1.1547^2 + 4}} = 0.5000,$$

and the variance of r converted from d will be

$$v_r = \frac{4^2 \times 0.0550}{(1.1547^2 + 4)^3} = 0.0058.$$

11.7 COMPUTING *d* FROM CLUSTER-RANDOMIZED STUDIES

Studies with nested designs are frequently used to evaluate the effects of social treatments (such as interventions, products, or technologies in education or public health). One common nested design assigns entire sites (often classrooms, schools, clinics, or communities) to the same treatment group, with different sites assigned to different treatments. Experiments with designs of this type are also called *group-randomized* or *cluster-randomized* designs because sites such as schools or communities correspond to statistical clusters. In experimental design terminology, these designs are designs involving clusters as nested

factors. Nested factors are groupings of individuals that occur only within one treatment condition, such as schools or communities in designs that assign whole schools or communities to treatments.

In this section, we deal only with continuous outcomes and the standardized mean difference. Analogous methods exist for use when other effect sizes are used (such as risk differences of odds ratios) are not discussed here. An approximate method for adjusting the variances of effect sizes for the effects of clustering is to multiply the variance of the effect size that would be computed if there is no clustering by a design effect associated with the particular clustered sampling design.

The most widely used designs (or at least the designs most widely acknowledged in analyses) are designs with one level of nesting (so called two-level designs). In such designs, individuals are sampled by first sampling existing groups of individuals (such as classrooms, schools, communities, or hospitals), then individuals are sampled within the groups. In such designs, whole groups are assigned to treatments.

Earlier, we introduced the standardized mean difference, defined as

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

and estimated using

$$d = \frac{\overline{Y}_1 - \overline{Y}_2}{S_{Within}}.$$

In that discussion, we were working with trials that had one level of sampling. We therefore had no need to further define δ—it is the difference in means divided by the standard deviation within groups. By contrast, in a multilevel study there are several possible standard deviations to use in the denominator, and each yields a variant of δ. Therefore, we use a subscript to identify each variant.

We can standardize by the standard deviation within groups *within clusters*, in which case the effect size would be δ_W. We can standardize by the standard deviation between clusters, in which case the effect size would be δ_B. We can also standardize by the total standard deviation (within groups across clusters), in which case the effect size would be δ_T. Each of these is a different index. The one we are calling δ_W is the one that corresponds to the index called (simply) δ earlier in this chapter. If we wanted to include studies that employed simple randomization and cluster randomization in the same analysis, we would need to use δ_W (Hedges 2007, 2011; Spybrook, Hedges, and Borenstein 2014).

11.7.1 Model and Notation

The data structure of cluster-randomized trials is more complex than individually randomized trials. To adequately define effect sizes, we need to first describe the notation we will use in this section. Let Y_{ij}^T ($i = 1, \ldots, m^T$; $j = 1, \ldots, n$) and Y_{ij}^C ($i = 1, \ldots, m^C$; $j = 1, \ldots, n$) be the jth observation in the ith cluster in the treatment and control groups respectively. We will have m^T clusters in the treatment group, m^C clusters in the control group, and a total of $M = m^T + m^C$ clusters with n observations each. Thus the sample size is $N^T = m^T n$ in the treatment group, $N^C = m^C n$ in the control group, and the total sample size is $N = N^T + N^C$.

Let $\overline{Y}_{i\bullet}^T$ ($i = 1, \ldots, m^T$) and $\overline{Y}_{i\bullet}^C$ ($i = 1, \ldots, m^C$) be the means of the ith cluster in the treatment and control groups, respectively. In addition, let $\overline{Y}_{\bullet\bullet}^T$ and $\overline{Y}_{\bullet\bullet}^C$ be the overall (grand) means in the treatment and control groups, respectively. Define the (pooled) within-cluster sample variance S_W^2 via

$$S_w^2 = \frac{\sum_{i=1}^{m^T}\sum_{j=1}^{n}\left(Y_{ij}^T - \overline{Y}_{i\bullet}^T\right)^2 + \sum_{i=1}^{m^C}\sum_{j=1}^{n}\left(Y_{ij}^C - \overline{Y}_{i\bullet}^C\right)^2}{N - M}$$

and the total pooled within-treatment group variance S_T^2 via

$$S_T^2 = \frac{\sum_{i=1}^{m^T}\sum_{j=1}^{n}\left(Y_{ij}^T - \overline{Y}_{\bullet\bullet}^T\right)^2 + \sum_{i=1}^{m^C}\sum_{j=1}^{n}\left(Y_{ij}^C - \overline{Y}_{\bullet\bullet}^C\right)^2}{N - 2}.$$

Suppose that observations within the treatment and control group clusters are normally distributed about cluster means μ_i^T and μ_i^C with a common within-cluster variance ν_W^2. That is

$$Y_{ij}^T \sim N\left(\mu_i^T, \sigma_w^2\right), i = 1, \ldots, m^T; j = 1, \ldots, n$$

and

$$Y_{ij}^C \sim N\left(\mu_i^C, \sigma_w^2\right) i = 1, \ldots, m^C; j = 1, \ldots, n.$$

Suppose further that the clusters have random effects (for example, are considered a sample from a population of clusters) so that the cluster means themselves have a normal sampling distribution with means $\mu_{\cdot}^{T}$ and $\mu_{\cdot}^{C}$ and common variance σ_B^2. That is,

$$\mu_i^T \sim N\left(\mu_{\cdot}^T, \sigma_B^2\right), i = 1, \ldots, m^T$$

and

$$\mu_i^C \sim N\left(\mu_{\cdot}^C, \sigma_B^2\right), i = 1, \ldots, m^C.$$

Note that in this formulation, σ_B^2 represents true variation of the population means of clusters over and above the variation in sample means that would be expected from variation in the sampling of observations into clusters.

These assumptions correspond to the usual assumptions that would be made in the analysis of a cluster-randomized trial by a hierarchical linear models analysis, an analysis of variance (with treatment as a fixed effect and cluster as a nested random effect), or a *t*-test using the cluster means in treatment and control group as the unit of analysis.

Note on the assumption of equal sample sizes. Most studies are planned with the intention of equal sample sizes in each cluster, which is simpler and maximizes statistical power. Although the eventual data collected may not have exactly equal sample sizes, sizes are often nearly equal. As a practical matter, exact cluster sizes are frequently not reported, so the research reviewer often has access only to the average sample sizes and thus may have to proceed as if sample sizes are equal. In this chapter, we present results assuming equal sample sizes (for the analogous results when cluster sample sizes are unequal, see Hedges 2007).

In principle, there are three within-treatment group standard deviations, σ_B, σ_W, and σ_T, the latter defined by

$$\sigma_T^2 = \sigma_B^2 + \sigma_W^2.$$

In most educational data when clusters are schools, σ_B^2 is considerably smaller than σ_W^2. Obviously, if the between-cluster variance σ_B^2 is small, then σ_T^2 will be similar to σ_W^2.

11.7.2 Intraclass Correlation with One Level of Nesting

A parameter that summarizes the relationship between the three variances is called the intraclass correlation ρ. It is defined by

$$\rho = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2} = \frac{\sigma_B^2}{\sigma_T^2}. \tag{11.85}$$

The intraclass correlation ρ can be used to obtain one of these variances from any of the others, given that $\sigma_B^2 = \rho\sigma_T^2$ and $\sigma_W^2 = (1 - \rho)\sigma_T^2 = (1 - \rho)\sigma_B^2/\rho$.

11.7.3 Primary Analyses

Many analyses reported in the literature fail to take the effects of clustering into account or take clustering into account by analyzing cluster means as the raw data. In either case, the treatment effect (the mean difference between treatment and control groups), the standard deviations S_W^2 and S_T^2, and the sample sizes may be reported (or deduced from information that is reported). In other cases, the analysis reported may have taken clustering into account by treating clusters as random effects. Such an analysis (such as using the program HLM, SAS Proc Mixed, or the STATA routine XTMixed) usually yields direct estimates of the treatment effect b and the variance components σ_B^2 and σ_W^2 and their standard errors. This information can be used to calculate both the effect size and its approximate standard error. Let $\hat{\mathrm{v}}_B^2$ and $\hat{\mathrm{v}}_W^2$ be the estimates of σ_B^2 and σ_W^2 and let $\mathrm{V}(b)$, $\mathrm{V}(\hat{\mathrm{v}}_B^2)$ and $\mathrm{V}(\hat{\mathrm{v}}_W^2)$ be their variances (the square of their standard errors). Generally, $\mathrm{V}(\hat{\mathrm{v}}_W^2)$ depends primarily on the number of individuals and $\mathrm{V}(\hat{\mathrm{v}}_B^2)$ depends primarily on the number of clusters. Because the number of individuals typically greatly exceeds the number of clusters, $\mathrm{V}(\hat{\mathrm{v}}_W^2)$ will be so much smaller than $\mathrm{V}(\hat{\mathrm{v}}_B^2)$ that $\mathrm{V}(\hat{\mathrm{v}}_W^2)$ can be considered negligible.

11.7.4 Effect Sizes with One Level of Nesting

In designs with one level of nesting (two-level designs), there are three possible standardized mean difference parameters corresponding to the three different standard deviations. The choice of one of these effect sizes should be determined on the basis of the inference of interest to the researcher. If the effect-size measures are to be used in meta-analysis, an important inference goal may be to

estimate parameters that are comparable with those that can be estimated in other studies. In such cases, the standard deviation may be chosen to be the same kind of standard deviation used in the effect sizes of other studies to which this study will be compared. We focus on three effect sizes that seem likely to be the most useful (meaning, the most widely used).

If $\sigma_W \neq 0$ (and hence $\rho \neq 1$), one effect-size parameter has the form

$$\delta_W = \frac{\mu_\bullet^T - \mu_\bullet^C}{\sigma_W}. \tag{11.86}$$

This effect size might be of interest, for example, in a meta-analysis where the other studies with which the current study is compared are typically single-site studies. In such studies δ_W may (implicitly) be the effect size estimated and hence δ_W might be the effect size most comparable with that in other studies.

A second effect-size parameter is of the form

$$\delta_T = \frac{\mu_\bullet^T - \mu_\bullet^C}{\sigma_T}. \tag{11.87}$$

This effect size might be of interest in a meta-analysis where the other studies are multisite studies or studies that sample from a broader population but do not include clusters in the sampling design (this would typically imply that they used an individual rather than a cluster assignment strategy). In such cases, δ_T might be the most comparable with the effect sizes in other studies.

If $\sigma_B \neq 0$ (and hence $\rho \neq 0$), a third possible effect-size parameter would be

$$\delta_B = \frac{\mu_\bullet^T - \mu_\bullet^C}{\sigma_B}. \tag{11.88}$$

but this parameter is seldom used in practice and is not discussed here (for a discussion, see Hedges 2007).

Note that if all of the effect sizes are defined (that is, if $0 < \rho < 1$), and ρ is known, any one of these effect sizes may be obtained from any of the others. In particular, both δ_W and δ_T can be obtained from δ_B and ρ because

$$\delta_W = \delta_B \sqrt{\frac{\rho}{1-\rho}} = \frac{\delta_T}{\sqrt{1-\rho}} \tag{11.89}$$

and

$$\delta_T = \delta_B \sqrt{\rho} = \delta_W \sqrt{1-\rho}. \tag{11.90}$$

11.7.5 Estimation of δ_W

We start with estimation of δ_W, which is the most straightforward. If $\rho \neq 1$ (so that $\sigma_W \neq 0$ and δ_W is defined), the estimate

$$d_W = \frac{\overline{Y}_{\bullet\bullet}^T - \overline{Y}_{\bullet\bullet}^C}{S_W} \tag{11.91}$$

is a consistent estimator of δ_W, which is approximately normally distributed about δ_W. The (estimated) variance of d_W is

$$v_W = \left(\frac{N^T + N^C}{N^T N^C}\right)\left(\frac{1+(n-1)\rho}{1-\rho}\right) + \frac{d_W^2}{2(N-M)}. \tag{11.92}$$

Note that the presence of the factor $(1 - \rho)$ in the denominator of the first term is possible since δ_W is defined only if $\rho \neq 1$.

Note that if $\rho = 0$ and there is no clustering, equation (11.92) reduces to the variance of a mean difference divided by a standard deviation with $(N - M)$ degrees of freedom. The leading term of the variance in equation (11.92) arises from uncertainty in the mean difference. Note that it is $[1 + (n - 1)\rho]/(1 - \rho)]$ as large as would be expected if there were no clustering in the sample (that is if $\rho = 0$). Thus $[1 + (n - 1)\rho]/(1 - \rho)]$ is a kind of variance inflation factor for the variance of the effect-size estimate d_W.

When the analysis properly accounts for clustering by treating the clusters as having random effects, and an estimate b of the treatment effect, its variance $V(b)$, and an estimate $\hat{\mathrm{v}}_W^2$ of the variance component σ_W^2 is available (for example, from an HLM analysis), the estimate of δ_W is

$$d_W = \frac{b}{\hat{\sigma}_W}. \tag{11.93}$$

The approximate variance of the estimate given in (11.93) is

$$v_W = \frac{V(b)}{\hat{\sigma}_W^2}. \tag{11.94}$$

11.7.6 Estimation of δ_T

A consistent estimate of δ_T using the intraclass correlation is

$$d_T = \left(\frac{\overline{Y}^T_{\bullet\bullet} - \overline{Y}^C_{\bullet\bullet}}{S_T}\right)\sqrt{1 - \frac{2(n-1)\rho}{N-2}}, \qquad (11.95)$$

which is normally distributed in large samples with (an estimated) variance of

$$v_T = \left(\frac{N^T + N^C}{N^T N^C}\right)(1+(n-1)\rho) + d_T^2\left(\frac{(N-2)(1-\rho)^2 + n(N-2n)\rho^2 + 2(N-2n)\rho(1-\rho)}{2(N-2)[(N-2)-2(n-1)\rho]}\right). \qquad (11.96)$$

Note that if $\rho = 0$ and there is no clustering, d_T reduces to the conventional standardized mean difference and equation (11.96) reduces to the usual expression for the variance of the standardized mean difference.

The leading term of the variance in equation (11.96) arises from uncertainty in the mean difference. Note that this leading term is $[1 + (n - 1)\rho]$ as large as would be expected if there were no clustering in the sample (that is if $\rho = 0$). The expression $[1 + (n - 1)\rho]$ is the variance inflation factor mentioned by Allan Donner and Neil Klar (2000) and the design effect mentioned by Leslie Kish (1965) for the variance of means in clustered samples and also corresponds to a variance inflation factor for the effect-size estimates like d_T.

11.7.7 Confidence Intervals for δ_W, δ_B, and δ_T

The results in this paper can also be used to compute confidence intervals for effect sizes. If δ is any one of the effect sizes mentioned, d is a corresponding estimate, and v_d is the estimated variance of d, then a $100(1 - \alpha)$ percent confidence interval for δ based on d and v_d is given by

$$d - c_{\alpha/2}v_d \le \delta \le d + c_{\alpha/2}v_d, \qquad (11.97)$$

where $c_{\alpha/2}$ is the $100(1 - \alpha/2)$ percent point of the standard normal distribution (for example, 1.96 for $\alpha/2 = 0.05/2 = 0.025$).

11.7.8 Applications in Meta-Analysis

This section is intended to be useful in deciding what effect sizes might be desirable to use in studies with nested designs. The results in this chapter can be used to produce effect-size estimates and their variances from studies that report (incorrectly analyzed) experiments as if there were no nested factors. The required means, standard deviations, and sample sizes can usually be extracted from what may be reported.

Suppose it is decided that the effect-size δ_T is appropriate because most other studies both assign and sample individually from a clustered population. Suppose that the data are analyzed by ignoring clustering, then the test statistic is likely to be either

$$t = \sqrt{\frac{N^T N^C}{N^T + N^C}}\left(\frac{\overline{Y}^T_{\bullet\bullet} - \overline{Y}^C_{\bullet\bullet}}{S_T}\right)$$

or $F = t^2$. Either can be solved for

$$\left(\frac{\overline{Y}^T_{\bullet\bullet} - \overline{Y}^C_{\bullet\bullet}}{S_T}\right),$$

which can then be inserted into equation (11.95) along with ρ to obtain d_T. This estimate (d_T) of δ_T can then be inserted into equation (11.96) to obtain v_T, an estimate of the variance of d_T.

Alternatively, suppose it is decided that the effect-size δ_W is appropriate because most other studies involve only a single site. We may begin by computing d_T and v_T as before. Because we want an estimate of δ_W, not δ_T, we use the fact given in equation (11.89) that

$$\delta_W = \delta_T / \sqrt{1-\rho}$$

and therefore

$$d_T / \sqrt{1-\rho} \qquad (11.98)$$

is an estimate of δ_W with a variance of

$$v_T/(1-\rho). \qquad (11.99)$$

Example. An evaluation of the connected mathematics curriculum reported by James Ridgway and colleagues (2002) compared the achievement of $m^T = 18$ classrooms

of sixth-grade students who used connected mathematics with that of $m^C = 9$ classrooms in a comparison group that did not use connected mathematics. In this quasi-experimental design the clusters were classrooms. The cluster sizes were not identical but the average cluster size in the treatment groups was $N^T/m^T = 338/18 = 18.8$ and N^C/m^C 162/18 = 18 in the control group. The exact sizes of all the clusters were not reported, but here we treat the cluster sizes as if they were equal and choose $n = 18$ as a slightly conservative sample size. The mean difference between treatment and control groups is $\overline{Y}^T_{\bullet\bullet} - \overline{Y}^C_{\bullet\bullet} = 1.9$, the pooled within-groups standard deviation $S_T = 12.37$. This evaluation involved sites in all regions of the country and it was intended to be nationally representative. Ridgway and colleagues did not give an estimate of the intraclass correlation based on their sample. Larry Hedges and Eric Hedberg (2007) provide an estimate of the grade 6 intraclass correlation in mathematics achievement for the nation as a whole (based on a national probability sample) of 0.264 with a standard error of 0.019. For this example, I assume that the intraclass correlation is identical to that estimate, namely $\rho = 0.264$.

Suppose that the analysis ignored clustering and compared the mean of all the students in the treatment with the mean of all the students in the control group. This leads to a value of the standardized mean difference of

$$\frac{\overline{Y}^T_{\bullet\bullet} - \overline{Y}^C_{\bullet\bullet}}{S_T} = 0.1536,$$

which is not an estimate of any of the three effect sizes considered here. If an estimate of the effect-size δ_T is desired, and we had imputed an intraclass correlation of $\rho = 0.264$, then we use equation (11.95) to obtain

$$d_T = (0.1536)(0.9907) = 0.1522.$$

The effect-size estimate is very close to the original standardized mean difference even though the amount of clustering in this case is not small. However, this amount of clustering has a substantial effect on the variance of the effect-size estimate. The variance of the standardized mean difference ignoring clustering is

$$\frac{324 + 162}{324 * 162} + \frac{0.1531^2}{2(324 + 162 - 2)} = 0.009259.$$

However, computing the variance of d_T using equation (11.96) with $\rho = 0.264$, we obtain a variance estimate of 0.050865, which is 549 percent of the variance ignoring clustering. A 95 percent confidence interval for δ_T is given by

$$-0.2899 = 0.1522 - 1.96\sqrt{0.050865} \le \delta_T \le 0.1522 + 1.96\sqrt{0.050865} = 0.5942.$$

If clustering had been ignored, the confidence interval for the population effect size would have been –0.0350 to 0.3422.

If we wanted to estimate δ_W, then an estimate of δ_W given by expression(11.98) is

$$\frac{0.1522}{\sqrt{1 - 0.264}} = 0.1774,$$

with variance given by expression (11.99) as

$$0.050865/(1 - 0.264) = 0.06911,$$

and a 95 percent confidence interval for δ_W based on (11.97) would be

$$-0.3379 = 0.1774 - 1.96\sqrt{0.06911} \le \delta_W \le 0.1774 + 1.96\sqrt{0.06911} = 0.6926.$$

11.8 COMBINING DATA FROM DIFFERENT TYPES OF STUDIES

We have now showed how to compute the same effect-size index from studies that use different designs. For example, we can compute d from studies that use two independent groups, a pre- or post-design, analysis of covariance, and cluster-randomized studies. We also showed how to convert across effect sizes—such as from an odds ratio to standardized mean difference. Although it is technically possible to make these conversions, we need to consider when it is a good idea to do so. The decision must be made on a case-by-case basis, and our goal here is to provide some context for making this decision.

The goal of a random-effects meta-analysis is not to include a set of identical studies. Instead, we include studies that may differ in a myriad of ways but nevertheless address the same fundamental question. For example,

suppose that we are testing the impact of tutoring on students. Some studies enroll freshmen and some enroll sophomores. Can we include both kinds of studies in the analysis? If we feel that they are addressing the same fundamental question, then the answer is yes. Otherwise, the answer is no. Suppose that some studies tutor the students for sixty minutes a day and some for ninety minutes. Can we include both kinds of studies in the analysis? If we feel that they are addressing the same fundamental question, then the answer is yes Otherwise, it is no.

The same logic applies when we turn to the way we assess outcome. Suppose some studies measure math skill using one test, and others use another. We decide that these tests are (or are not) addressing the same fundamental question. Suppose that we are comparing the risk of an event in two groups. If two studies have different follow-up periods, then the "risk of an event" really has a different meaning in the two studies, and we need to decide whether both are addressing the same fundamental question.

We typically make these kinds of decisions implicitly, as part of the inclusion-exclusion process. We have chosen to make this explicit here because it provides context for thinking about the question at hand. The decision that it is okay (or not) to include different study designs or different types of data in the same analysis is simply another variant of the decision process outlined for populations, protocols, and the like.

Suppose that some studies compared two independent groups, and others compared pre- and post-scores for the same group. In some cases (when the change from pre- to post- is almost certainly the result of the intervention), we might decide that including both kinds of studies in the same analysis is fine. In other cases, when the change could be due to external factors, we might decide not to include them in the same analysis. Our decision thus hinges on the question of whether the two designs are addressing the same fundamental question.

Suppose that some studies compared the means in two groups and that others compared the risk. In some cases, if the risk and the mean are capturing the same fundamental issue, we might decide that including both kinds of studies in the same analysis is fine. In other cases, when the outcome must be seen as a dichotomy, we might decide not to. Again, our decision hinges on the question of whether the two designs are addressing the same fundamental question.

The one rule that cannot be violated is that the effect-size index used in the analysis must be essentially the same for all studies. Thus, we would never include two versions of the standardized mean difference, when one is standardized by the standard deviation within groups and the other by that between groups. Similarly, we would never include odds ratios and risk ratios. (A possible exception for the risk ratio and hazard ratio was discussed earlier).

In this discussion, *different study designs* refers to instances in which, for example, one study uses independent groups and another uses a pre-post design or matched groups. It does not refer to instances when one study was a randomized trial (or quasi-experimental study) and another observational. In general, these two kinds of studies will be addressing fundamentally different questions; it would not be appropriate to include them in the same analysis.

11.9 CONCLUSION

In this chapter, we address the calculation of effect-size estimates and their sampling distributions, a technical exercise that depends on statistical theory. Equally important is the interpretation of effect sizes, a process that requires human judgment and that is not amenable to technical solutions (Cooper 2008).

To understand the substantive implications of an effect size, we need to look at the effect size in context. For example, to judge whether the effect of an intervention is large enough to be important, it may be helpful to compare it with the effects of related interventions that are appropriate for the same population and have similar characteristics (such as cost or complexity). Compendia of effect sizes from various interventions may be helpful in making such comparisons (for example, Lipsey and Wilson 1993).

Alternatively, it may be helpful to compare an effect size to other effect sizes that are well understood conceptually. For example, one might compare an educational effect size to the effect size corresponding to one year of growth. Even here, however, context is crucial. For example, one year's growth in achievement from kindergarten to grade 1 corresponds approximately to a d of 1.0, whereas one year's growth in achievement from grade eleven to grade twelve corresponds to a d of about 0.1 (see, for example, Bloom et al. 2008; Hill et al. 2008; or Lipsey et al. 2012).

Technical methods can help with the interpretation of effect sizes in one limited way. They can permit effect sizes computed in one metric (such as a correla-

tion coefficient) to be expressed in another (such as a standardized mean difference or an odds ratio). Such reexpression can make it easier to compare the effect with other effect sizes that may be known to the analyst and be judged to be relevant (Rosenthal and Rubin 1979, 1982).

11.10 RESOURCES

Most of the formulas discussed in 11.2 to 11.6 have been implemented in the computer program Comprehensive Meta-Analysis. The formulas discussed in 11.7 have been implemented in the computer program Computing Effect-Sizes for Cluster-Randomized Studies. For information on these programs, contact the first author at Biostat100@gmail.com.

Parts of this chapter have been adapted from the book *Computing Effect Sizes for Meta-Analysis* (Borenstein et al. 2019).

11.11 ACKNOWLEDGMENTS

The section on working with cluster-randomized studies describes work that was funded in part by ED-IES-11-C-0037 under the direction of Dr. Edward Metz. Other parts of this chapter were funded in part by the following grants from the National Institutes of Health: "Combining data types in meta-analysis" (AG021360), "Publication bias in meta-analysis" (AG20052), "Software for meta-regression" (AG024771), from the National Institute on Aging, under the direction of Dr. Sidney Stahl; and "Forest plots for meta-analysis" (DA019280) from the National Institute on Drug Abuse, under the direction of Dr. Thomas Hilton.

Parts of this chapter have been adapted from the books *Introduction to Meta-Analysis* and *Computing Effect Sizes for Meta-Analysis* that we co-authored with Julian Higgins and Hannah Rothstein. The section on binary data builds on the chapter written by Jesse Berlin for earlier editions of this handbook. Thanks to Steven Tarlow for checking the accuracy of the formulas.

11.12 REFERENCES

Alderson, Philip, and John J. Deeks. 2004. "Concerning: The Number Needed to Treat Needs an Associated Odds Estimation." *Journal of Public Health* 26(4): 400–401.

Altman, Douglas G. 1998. "Confidence Intervals for the Number Needed to Treat." *British Medical Journal* 317 (November): 1309–12.

Altman, Douglas G., and John J. Deeks. 2002. "Meta-Analysis, Simpson's Paradox, and the Number Needed to Treat." *BMC Medical Research Methodology* 2 (January): 3. DOI: 10.1186/1471-2288-2-3.

Altman, Douglas G., John J. Deeks, and David L. Sackett. 1998. "Odds Ratios Should Be Avoided When Events Are Common." *British Medical Journal* 317(7168): 1318.

Bloom, Howard S., Carolyn J. Hill, A. B. Black, and Mark W. Lipsey. 2008. "Performance Trajectories and Performance Gaps as Achievement Effect-Size Benchmarks for Educational Interventions." *Journal of Research on Educational Effectiveness* 1(4): 289–328. DOI: 10.1080/19345740802400072.

Borenstein, Michael 1994. "The Case for Confidence Intervals in Controlled Clinical Trials." *Control Clinical Trials* 15(5): 411–28.

———. 1997. "Hypothesis Testing and Effect Size Estimation in Clinical Trials." *Annual Allergy Asthma Immunology* 78: 5–11, quiz 12–6.

———. 2000. "The Shift from Significance Testing to Effect Size Estimation." In *Comprehensive Clinical Psychology*, edited by Alan S. Bellack and Michel Hersen. Oxford: Pergamon Elsevier.

———. 2019. *Common Mistakes in Meta-Analysis and How to Avoid Them.* Englewood, N.J.: Biostat, Inc.

Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. 2009. *Introduction to Meta-Analysis.* Chichester, UK: John Wiley & Sons.

———. 2019. *Computing Effect Sizes for Meta-Analysis*. Chichester, UK: John Wiley & Sons.

Chatellier, Gilles, Eric Zapletal, David Lemaitre, Joel Menard, and Patrice Degoulet. 1996. "The Number Needed to Treat: A Clinically Useful Nomogram in Its Proper Context." *British Medical Journal* 312: 426–29.

Chinn, Susan. 2000. "A Simple Method for Converting an Odds Ratio to Effect Size for Use in Meta-Analysis." *Statistics in Medicine* 19: 3127–31.

Cohen, Jacob 1969. *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

———. 1977. *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

Cook, Richard J., and David L. Sackett. 1995. "The Number Needed to Treat: A Clinically Useful Measure of Treatment Effect." *British Medical Journal* 310: 452–54.

Cooper, Harris. 2008. "The Search for Meaningful Ways to Express the Effects of Interventions." *Child Development Perspectives* 2: 181–86.

D'Amico, R., John J. Deeks, and Douglas G. Altman. 1999. "Numbers Needed to Treat Derived from Meta-Analysis.

Length of Follow Up Is Poorly Reported." *British Medical Journal* 319: 1200.

Deeks, John J. 2002. "Issues in the Selection of a Summary Statistic for Meta-Analysis of Clinical Trials with Binary Outcomes." *Statistics in Medicine* 21(11): 1575–600.

Deeks, John J., and Douglas G. Altman. 2001. "Effect Measures for Meta-Analysis of Trials with Binary Outcomes." In *Systematic Reviews in Health Care: Meta-Analysis in Context*, edited by Matthias Egger, G. Davey Smith, and Douglas G. Altman. London: BMJ Publishing Group.

Donner, Allan, and Neil Klar. 2000. *Design and Analysis of Cluster Randomization Trials in Health Research*. New York: John Wiley & Sons.

Ebrahim, Shah. 2001. "Numbers Needed to Treat Derived from Meta-Analyses: Pitfalls and Cautions." In *Systematic Reviews in Health Care: Meta-Analysis in Context*, edited by Matthias Egger, G. Davey Smith, and Douglas G. Altman. London: BMJ Publishing Group.

Fleiss, Joseph L., Bruce Levin, and Myunghee Cho Paik. 2003. *Statistical Methods for Rates and Proportions*. Hoboken, N.J.: John Wiley & Sons.

Glass, Gene V. 1976. "Primary, Secondary, and Meta-Analysis of Research." *Educational Researcher* 5(1): 3–8.

Glass, Gene V., Barry McGaw, and Mary L. Smith. 1981. *Meta-Analysis in Social Research* Thousand Oaks, Calif.: Sage Publications.

Grant, Robert L. 2014. "Converting an Odds Ratio to a Range of Plausible Relative Risks for Better Communication of Research Findings." *British Medical Journal* 348: f7450.

Grissom, Robert J., and John J. Kim. 2012. *Effect Sizes for Research*. New York: Routledge.

Hasselblad, Victor, and Larry V. Hedges. 1995. "Meta-Analysis of Screening and Diagnostic Tests." *Psychological Bulletin* 117(1): 167–78.

Hedges, Larry V. 2007. "Effect Sizes in Cluster Randomized Designs." *Journal of Educational and Behavioral Statistics*, 32: 341–70.

———. 2011. "Effect Sizes in Designs with Two Levels of Nesting." *Journal of Educational and Behavioral Statistics* 36: 346–80.

Hedges, Larry V., and Hedberg, Eric C. 2007. "Intraclass Correlations for Planning Group Randomized Experiments in Education." *Educational Evaluation and Policy Analysis* 29(1): 60–87.

Hedges, Larry V., and Ingram Olkin. 1985. *Statistical Methods for Meta-Analysis*. Orlando, Fl.: Academic Press.

———. 2016. "Overlap Between Treatment and Control Distributions as an Effect Size Measure in Experiments." *Psychological Methods* 21(1): 61–68.

Hunter, John E., and Frank L. Schmidt. 2015. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Thousand Oaks, Calif.: Sage Publications.

Hill, Carolyn J., Howard S. Bloom, A. R. Black, and Mark W. Lipsey. 2008. "Empirical Benchmarks for Interpreting Effect Sizes in Research." *Child Development Perspectives* 2(3): 172–77.

Kish, Leslie. 1965. *Survey Sampling*. New York: John Wiley & Sons.

Lipsey, Mark W., Kelly Puzio, Cathy Yun, Michael A. Hebert, Kasia Steinka-Fry, Mikel W. Cole, Megan Roberts, Karen S. Anthony, and Matthew D. Busick. 2012. "Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms." NCSER 2013–3000. Washington: U.S. Department of Education.

Lipsey, Mark W., and David B. Wilson. 1993. "The Efficacy of Psychological, Educational, and Behavioral Treatment: Confirmation from Meta-Analysis." *American Psychologist* 48(12): 1181–209.

McAlister, Finlay A. 2001. "Applying the Results of Systematic Reviews at the Bedside." In *Systematic Reviews in Health Care: Meta-Analysis in Context*, edited by Matthias Egger, G. Davey Smith, and Douglas G. Altman. London: BMJ Publishing Group.

———. 2008. "The 'Number Needed to Treat' Turns 20 — and Continues to Be Used and Misused." *Canadian Medical Association Journal* 179(6): 549–53.

Michiels, Stefan, Pascal Piedbois, Sarah Burdett, Nathalie Syz, Lesley A. Stewart, and Jean-Pierre Pignon. 2005. "Meta-Analysis When Only the Median Survival Times Are Known: A Comparison with Individual Patient Data Results." *International Journal of Technology Assessment in Health Care* 21(1): 119–25.

Parmar, Mahesh K., Valter Torri, and Lesley A. Stewart. 1998. "Extracting Summary Statistics to Perform Meta-Analyses of the Published Literature for Survival Endpoints." *Statistics in Medicine* 17(24): 2815–34.

Ridgway, James E., Jidth S. Zawojewski, Mark N. Hoover, and Diana V. Lambdin. 2002. "Student Attainment in the Connected Mathematics Curriculum." In *Standards-Based School Mathematics Curricula: What Are They? What Do Students Learn?*, edited by Sharon L. Senk and Denisse R. Thompson. Mahwah, N.J.: Lawrence Erlbaum Associates, Inc.

Rosenthal, Robert, and Donald B. Rubin. 1979. "A Note on Percentage of Variance Explained as a Measure of Importance of Effects." *Journal of Applied Social Psychology* 9(5): 395–96.

———. 1982. "A Simple, General Purpose Display of Magnitude of Experimental Effect." *Journal of Educational Psychology* 74(2): 166–69.

Rosenthal, Robert, Ralph L. Rosnow, and Donald B. Rubin. 2000. *Contrasts and Effect Sizes in Behavioral Research: A Correlational Approach*. New York: Cambridge University Press.

Rosnow, Ralph L., Robert Rosenthal, and Donald B. Rubin. 2000. "Contrasts and Correlations in Effect-Size Estimation." *Psychological Science* 11(6): 446–53.

Rothman, Kenneth J. 2012. *Epidemiology: An Introduction*. New York: Oxford University Press.

Sanchez-Meca, Julio, Fulgencio Marin-Martinez, and Salvador Moscoso. 2003. "Effect-Size Indices for Dichotomized Outcomes in Meta-Analysis." *Psychological Methods* 8(4): 448–67.

Spybrook, Jessaca, Larry V. Hedges, and Michael Borenstein. 2014. "Understanding Statistical Power in Cluster Randomized Trials: Challenges Posed by Differences in Notation and Terminology." *Journal of Research on Educational Effectiveness* 7(4): 384–406.

Stang, Andreas, Charles Poole, and Ralf Bender. 2010. "Common Problems Related to the Use of Number Needed to Treat." *Journal of Clinical Epidemiology* 63(8): 820–25.

Sweeting, Michael J., Alexander J. Sutton, and Paul C. Lambert. 2004. "What to Add to Nothing? Use and Avoidance of Continuity Corrections in Meta-Analysis of Sparse Data." *Statistics in Medicine* 23(9): 1351–75.

Valentine, Jeffrey C., and Ariel M. Aloe. 2016. "How to Communicate Effect Sizes for Continuous Outcomes: A Review of Existing Options and Introducing a New Metric." *Journal of Clinical Epidemiology* 72 (April): 84–89.

Wen, Lonnie, Robert Badgett, and John Cornell. 2005. "Number Needed to Treat: A Descriptor for Weighing Therapeutic Options." *American Journal of Health System Pharmacy* 62(19): 2031–36.

Whitehead, Anne. 2002. *Meta-Analysis of Controlled Clinical Trials*. Chichester, UK: John Wiley & Sons.

Zhang, Jun, and Kai F. Yu. 1998. "What's the Relative Risk? A Method of Correcting the Odds Ratio in Cohort Studies of Common Outcomes." *Journal of the American Medical Association* 280(19): 1690–91.

12

STATISTICALLY ANALYZING EFFECT SIZES: FIXED- AND RANDOM-EFFECTS MODELS

SPYROS KONSTANTOPOULOS
Michigan State University

LARRY V. HEDGES
Northwestern University

CONTENTS

12.1 INTRODUCTION

In this chapter, three general classes of fixed-effects and random-effects models are presented. One class of models is appropriate for estimating the mean effect across studies. Two other classes of models are appropriate when examining the relation between independent (study characteristic) variables and effect size. One of these is appropriate when the independent (study characteristic) variables are categorical. It is analogous to the analysis of variance but is adapted to the special characteristics of effect-size estimates. The other of these is appropriate for either discrete or continuous independent (study characteristic) variables and therefore technically includes the first class as a special case. This class is analogous to multiple regression analysis for effect sizes. In all three cases, we describe the models along with procedures for estimation and hypothesis testing. Although some formulas for hand computation are given, we stress computation via widely available packaged computer programs.

Tests of goodness of fit are given for each fixed-effect model. They test the notion that there is no more variability in the observed effect sizes than would be expected if all (100 percent) of the variation in effect-size *parameters* is "explained" by the data analysis model (the predictor variables). These tests can be conceived as tests of "model specification." That is, if a fixed-effects model explains all of the variation in effect-size parameters, the (fixed-effect) model is appropriate. Models that are well specified can provide a strong basis for inference about

effect sizes in fixed-effects models but are not essential for inference from them. If differences between studies that lead to differences in effects are *not* regarded as random (for example, if they are regarded as consequences of purposeful design decisions) then fixed-effects methods are appropriate for the analysis. Similarly, fixed-effects analyses are appropriate if the inferences desired are regarded as conditional—applying only to observed studies under examination.

12.2 ESTIMATING THE MEAN EFFECT

The simplest meta-analytic analysis is estimating the mean effect from a series of independent studies. Both fixed- and random-effects statistical methods are available for studying the variation in effects. The choice of which to use is sometimes a contentious issue in both meta-analysis as well as primary analysis of data. The choice of statistical procedures should primarily be determined by the kinds of inference the synthesist wishes to make. Two different inference models are available, sometimes called *conditional* and *unconditional* inference (see, for example, Hedges and Vevea 1998). The conditional inference model attempts to make inference about the relation between covariates and the effect-size parameters in the sample of studies that are *observed*. In contrast, the unconditional inference model attempts to make inferences about the relation between covariates and the effect-size parameters in the population of studies from which the observed studies are considered to be a representative sample. Fixed-effects statistical procedures are well suited to drawing conditional inferences, inferences about the *observed* studies (see, for example, Hedges and Vevea 1998). Fixed-effects statistical procedures may also be a reasonable choice when the number of studies is too small to support the effective use of mixed- or random-effects models. Random- or mixed-effects statistical procedures are well suited to drawing unconditional inferences (inferences about the population of studies from which the observed studies are randomly selected). We use the terms *random effects* and *mixed effects* interchangeably.

12.2.1 Fixed-Effects Models

Suppose that the data to be combined arise from a series of k independent studies, in which the ith study reports one observed effect size T_i, with population effect size θ_i, and variance v_i. Thus, the data to be combined consist of k effect size estimates $T_1, \ldots, T_k$ of parameters $\theta_1, \ldots, \theta_k$, and variances $v_1, \ldots, v_k$. Under the fixed effects model, we assume $\theta_1 = \ldots = \theta_k = \theta$, a common effect size. Then a general formula for the weighted average effect size over those studies is

$$\bar{T}_\bullet = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i}, \tag{12.1}$$

where w_i is a weight assigned to the ith study that is defined in equation (12.2).

When all observed effect-size indicators estimate a single population parameter, as is hypothesized under a fixed-effects model, then $\bar{T}_\bullet$ is an unbiased estimate of the population parameter θ. In equation (12.1), T_i may be estimated by any specific effect size statistic.

The weights that minimize the variance of $\bar{T}_\bullet$ are inversely proportional to the conditional variance (the square of the standard error) in each study:

$$w_i = \frac{1}{v_i}. \tag{12.2}$$

Formulas for the conditional variances, v_i, vary for different effect-size indices but they are presented in chapter 11. However, they share in common the fact that conditional variance is inversely proportional to within-study sample size—the larger the sample, the smaller the variance, so the more precise the estimate of effect size should be. Hence, larger weights are assigned to effect sizes from studies that have larger within-study sample sizes. Of course, equation (12.2) defines the optimal weights assuming we know the conditional variances, v_i, for each study. In practice we must estimate those variances ($\hat{v}_i$), so we can only estimate these optimal weights ($\hat{w}_i$).

Given the use of weights as defined in (12.2), the average effect size $\bar{T}_\bullet$ has conditional variance $v_\bullet$, which itself is a function of the conditional variances of each effect size being combined:

$$v_\bullet = \frac{1}{\sum_{i=1}^{k} (1/v_i)}. \tag{12.3}$$

The square root of $v_\bullet$ is the standard error of estimate of the average effect size. Multiplying the standard error by an appropriate critical value C_α and adding and

subtracting the resulting product to $\bar{T}_\bullet$, yields the (95 percent) confidence interval for θ:

$$\bar{T}_\bullet - C_\alpha \le \theta \le \bar{T}_\bullet + C_\alpha \tag{12.4}$$

C_α is often the unit normal $C_\alpha = 1.96$ for a two-tailed test at $\alpha = .05$; but better type I error control and better coverage probability of the confidence interval may occur if C_α is student's t-statistic at $k - 1$ degrees of freedom. In either case, if the confidence interval does not contain zero, we reject the null hypothesis that the population effect size θ is zero. Equivalently, we may test the null hypothesis that $\theta = 0$ with the statistic

$$Z = \frac{|\bar{T}_\bullet|}{\sqrt{v_\bullet}}, \tag{12.5}$$

where $|\bar{T}_\bullet|$ is the absolute value of the weighted average effect size over studies (given by equation 12.1). Under (12.5), $\bar{T}_\bullet$ differs from zero if Z exceeds 1.96, the 95 percent two-tailed critical value of the standard normal distribution.

A test of the assumption of equation (12.1) that studies do, in fact, share a common population effect size uses the following homogeneity test statistic:

$$Q = \sum_{i=1}^{k} \left[(T_i - \bar{T}_\bullet)^2 / v_i \right] = \sum_{i=1}^{k} w_i (T_i - \bar{T}_\bullet)^2. \tag{12.6}$$

A computationally convenient form of (12.6) is

$$Q = \sum_{i=1}^{k} w_i T_i^2 - \frac{\left(\sum_{i=1}^{k} w_i T_i \right)^2}{\sum_{i=1}^{k} w_i}. \tag{12.7}$$

If Q exceeds the upper-tail critical value of chi-square at $k - 1$ degrees of freedom, the observed variance in study effect sizes is significantly greater than we would expect by chance if all studies shared a common population effect size. If homogeneity is rejected, $\bar{T}_\bullet$ should not be interpreted as an estimate of a single effect parameter θ that gave rise to the sample observations, but rather simply as describing a mean of observed effect sizes, or as estimating a (weighted) mean θ, which may, of course, be of practical interest in a particular research synthesis. If Q is rejected, the researcher may wish to disaggregate study effect sizes by grouping studies into appropriate categories or to use regression techniques to account for variance among the θ_i. These latter techniques are discussed in subsequent sections of this chapter, as are methods for conducting sensitivity analyses that help examine the influence of particular studies on combined effect size estimates and on heterogeneity.

Q is a diagnostic tool to help researchers know whether they have, to put it in the vernacular, "accounted for all the variance" in the effect sizes they are studying. Experience has shown that Q is usually rejected in most simple fixed-effects univariate analyses—for example, when the researcher simply lumps all studies into one category or contrasts one category of studies, such as randomized experiments, with another category of studies, such as quasi-experiments. In such simple category systems, rejection of homogeneity makes eminent theoretical sense! Each simple categorical analysis can be thought of as the researcher's theoretical model about what variables account for the variance in effect sizes. We would rarely expect that just one variable, or even just two or three variables, would be sufficient to account for all observed variance. The phenomena we study are usually far more complex than that. Often, extensive multivariate analyses are required to model these phenomena successfully. In essence, then, the variables that a researcher uses to categorize or predict effect sizes can be considered to be the model that the researcher has specified about what variables generated study outcome. The Q statistic tells whether that model specification is statistically adequate. Thus, homogeneity tests can serve a valuable diagnostic function.

A useful supplement to Q is the descriptive statistic:

$$I^2 = 100\% * \left(\frac{Q - (k-1)}{Q} \right), \tag{12.8}$$

which describes the percentage of the total variance in effect-size estimates that is due to variance among the effect-size parameters (Higgins and Thompson 2002; Higgins et al. 2003). Negative values of I^2 are set to zero. Values of I^2 are not affected by the numbers of studies or the effect-size metric. Julian Higgins and Simon Thompson (2002) suggest the following approximate guidelines for interpreting this statistic: $I^2 = 25\%$ (small heterogeneity), $I^2 = 50\%$ (medium heterogeneity), and $I^2 = 75\%$ (large

heterogeneity). We recommend using I^2 as a descriptive statistic without the confidence intervals to supplement Q rather than to replace Q.

Example. Alice Eagly and Linda Carli (1981) reported a synthesis of ten studies of gender differences in conformity using the so-called fictitious norm group paradigm. The effect sizes were standardized mean differences. Table 12.1 presents sums necessary to do the fixed effects computations. (Throughout this chapter, results will vary slightly depending on the number of decimals used in the computations.) The weighted average effect-size estimate is computed using (12.1) as $\bar{T}_\bullet = 34.649/279.161 = 0.124$, and the variance is computed from (12.3) as $v_\bullet = 1/279.161 = 0.00358$, which corresponds to a standard error of $\sqrt{0.00358} = 0.060$. We test the significance of this effect size in either of two equivalent ways: by computing the 95 percent confidence interval using (12.4), which in this case ranges from $0.124 - 1.96\sqrt{0.00358} = 0.007$ to $0.124 + 1.96\sqrt{0.00358} = 0.241$, which does include zero in the confidence interval, and by computing the statistic Z using (12.5), which yields $Z = |0.124|/\sqrt{0.00358} = 2.074$, which exceeds the 1.96 critical value at $\alpha = 0.05$. Hence we conclude that there is a positive average gender difference in conformity. However, homogeneity of effect size is rejected in this data, with the computational version of equation (12.6) yielding $Q = 36.076 - [(34.649)^2/279.161] = 31.775$, which exceeds 21.67, the 99 percent critical value of the chi-square distribution for $10 - 1 = 9$ degrees of freedom, so we reject homogeneity of effect sizes at $p < .01$. Similarly, using equation (12.7) we compute $I^2 = 100\% \times [31.775 - (10 - 1)]/31.775 = 71.6\%$, suggesting nearly three quarters of the variation in effect sizes is due to real heterogeneity of the effect-size parameters. Hence, we might assume that other variables could be necessary to fully explain the variance in these effect sizes; subsequent sections in this chapter will explore such possibilities.

12.2.2 Random-Effects Models

Under a fixed-effects model, the effect-size statistics, T_i, from k studies estimate a population parameter $\theta_1 = \ldots = \theta_k = \theta$. The estimate T_i in any given study differs from the θ due to sampling error, or what we have referred to as conditional variability; that is, because a given study used a *sample* of subjects from the population, the estimate of T_i computed for that sample will differ somewhat from θ for the population.

Under a random effects model, θ_i is not fixed but is itself random and has its own distribution. Hence, total variability of an observed study effect size estimate v_i^* reflects *both* conditional variation v_i of that effect size around each population θ_i *and* random variation τ^2 of the individual θ_i around the mean population effect size:

$$v_i^* = v_i + \tau^2. \tag{12.9}$$

In this equation, we will refer to τ^2 as either the between-studies variance or the variance component (others sometimes refer to this as random-effects variance); to v_i as either within-study variance or the conditional variance of T_i (the variance of an observed effect size conditional on θ being fixed at the value of θ_i; others sometimes call this estimation variance); and to v_i^* as the unconditional variance of an observed effect size T_i (others sometimes call this the variance of estimated effects). If the between-studies variance is zero, then the equations of the random-effects model reduce to those of the fixed-effects model, with unconditional variability of an observed effect size $[v_i^*]$ hypothesized to be due entirely to conditional variability $[v_i]$ (to sampling error).

Once the researcher decides to use a random-effects analysis, a first task is to determine whether or not the variance component differs significantly from zero and, if it does, then to estimate its magnitude. Estimating the variance component can be done in many different ways (Viechtbauer 2005). This unbiased sample estimate of the variance component will sometimes be negative, even though the population variance component must be a positive number. In these cases, it is customary to fix the component to zero.

The most common method for estimating the variance component begins with Q as defined in equation (12.6). The expected value of Q is

$$E\{Q\} = c\tau^2 + (k-1),$$

where

$$c = \sum_{i=1}^{k} w_i - \left[\sum_{i=1}^{k} w_i^2 \Big/ \sum_{i=1}^{k} w_i\right]. \tag{12.10}$$

Solving for τ^2 and substituting Q for its expectation gives an estimator of the variance component:

$$\hat{\tau}^2 = [Q - (k-1)]/c, \tag{12.11}$$

which is set to zero if it is negative.

Table 12.1 Data for the Gender Differences in Conformity Example

Group	Number of Studies in Group	Number of Items	N	T	v	Fixed-Effects Analysis				Random-Effects Analysis			
						w	w^2	wT	wT^2	w	w^2	wT	wT^2
1	1	2	141	–0.330	0.029	34.483	1,189.061	–11.379	3.755	18.519	342.936	–6.111	2.017
1	2	2	119	0.070	0.034	29.412	865.052	2.059	0.144	16.949	287.274	1.186	0.083
		Group 1 sums				63.895	2,054.113	–9.320	3.899	35.468	630.209	–4.925	2.100
2	1	2	191	–0.300	0.022	45.455	2,066.116	–13.636	4.091	21.277	452.694	–6.383	1.915
		Group 2 sums				45.455	2,066.116	–13.636	4.091	21.277	452.694	–6.383	1.915
3	1	38	254	0.350	0.016	62.500	3,906.250	21.875	7.656	24.390	594.884	8.537	2.988
3	2	30	64	0.700	0.066	15.152	229.568	10.606	7.424	10.989	120.758	7.692	5.385
3	3	45	20	0.850	0.218	4.587	21.042	3.899	3.314	4.115	16.935	3.498	2.973
3	4	45	90	0.400	0.045	22.222	493.827	8.889	3.556	14.286	204.082	5.714	2.286
3	5	45	60	0.480	0.069	14.493	210.040	6.957	3.339	10.638	113.173	5.106	2.451
3	6	5	80	0.370	0.051	19.608	384.468	7.255	2.684	13.158	173.130	4.868	1.801
3	7	5	125	–0.060	0.032	31.250	976.563	–1.875	0.113	17.544	307.787	–1.053	0.063
		Group 3 sums				169.811	6,221.757	57.605	28.086	95.120	1,530.750	34.363	17.947
		Overall sums				279.161	10,341.986	34.649	36.076	151.865	2,613.652	23.056	21.962

SOURCE: Authors' tabulation based on data from Eagly and Carli 1981.

Alternative estimators, such as the restricted maximum likelihood estimator described by Wolfgang Viechtbauer (2005) tends to perform better than (12.11), but must be estimated with an iterative procedure which is implemented in several software packages.

The random effects weighted mean of $T_1, \ldots, T_k$, $\overline{T}_*$, is an estimate of μ_θ, the average of the random effects in the population; v_*, the variance of $\overline{T}_*$ (the square root of v_* is the standard error of $\overline{T}_*$); and random-effects confidence limits θ_L and θ_U for μ_θ by multiplying the standard error by an appropriate critical value (often, 1.96 at $\alpha = .05$), and adding and subtracting the resulting product from $\overline{T}_*$. In general, the computations follow (12.1) to (12.5), except that the following unconditional variance estimate is used in (12.2) through (12.5) in place of the conditional variances outlined in the fixed-effects models:

$$v_i^* = \tau^2 + v_i, \tag{12.12}$$

where τ^2 is the variance component estimate yielded by (12.11) and v_i is the conditional variance of T_i. The square root of the variance component describes the standard deviation of the distribution of effect parameters. Multiplying that standard deviation by 1.96 (or an appropriate critical value of *t*), and adding and subtracting the result from the average random-effect size $\overline{T}_*$, yields the limits of an approximate 95 percent confidence interval. All these random-effects analyses assume that the random effects are normally distributed with constant variance, an assumption that is particularly difficult to assess when the number of studies is small.

Example. We apply these procedures to the data from ten studies on the effects of gender differences in conformity. Table 12.1 provides the sums necessary to perform the random-effects calculations. To compute the sample estimate of the variance component we first compute the constant c using (12.10) to obtain $c = 279.161 - 2613.652 = 269.798$. Then we compute the variance component estimate using (12.11) as $\hat{\tau}^2 = [31.775 - (10-1)]/269.795 = 0.084$. Using this value of τ^2 we then compute the random effects weights w_i^*. The estimate of the random effects average effect size is $\overline{T}_\bullet^* = 23.056/151.865 = 0.152$ with a variance of $v_\bullet^* = 1/151.865 = 0.00658$, which corresponds to a standard error of $\sqrt{0.00658} = 0.081$. The limits of the 95 percent confidence interval ranges from $0.152 - 1.96\sqrt{0.00658} = -0.007$ to $0.152 + 1.96\sqrt{0.00658} = 0.311$, which includes zero in the confidence interval, and the statistic $Z = |0.152|/\sqrt{0.00658} = 1.871$, which does not exceed the 1.96, the $\alpha = 0.05$ critical value of the standard normal distribution. Hence, unlike in the fixed-effects analysis, we do not conclude that there is a nonzero average gender difference in conformity.

12.3 ANALYSIS OF VARIANCE FOR EFFECT SIZES

One of the most common situations in research synthesis arises when the effect sizes can be sorted into independent groups according to one or more characteristics of the studies generating them. The analytic questions are whether the groups' (average) population effect sizes vary and whether the groups are internally homogeneous, that is, whether the effect sizes vary within the groups. Alternatively, we could describe the situation as one of exploring the relationship between a categorical independent variable (such as one grouping variable) and the effect-size estimates (the outcome). This is the situation addressed by analysis of variance in experimental data. This section describes an analog to the one-factor analysis of variance for effect sizes. Extensions of these methods to more than one categorical independent variable are available but are usually handled in meta-analysis by using the multiple regression methods. Our numerical examples in this chapter use effect-size estimates that are standardized mean differences (for technical details, see Hedges 1982a; Hedges and Olkin 1985).

12.3.1 Fixed-Effects Analyses

Situations frequently arise in which we wish to determine whether a particular discrete characteristic of studies is related to an outcome (effect-size estimates). For example, we may want to know whether the type of treatment is related to the treatment's effect or whether all variations of the treatment produce essentially the same effect. The effect-sizes analog to one-factor analysis of variance is designed to answer just such questions.

12.3.1.1 Notation In the discussion of the one-factor model, we use a notation emphasizing that the independent effect-size estimates fall into p groups, defined a priori by the independent (grouping) variable. Suppose that there are p distinct groups of effects with m_1 effects in the first group, m_2 effects in the second group, . . . , and m_p effects in the pth group and a total of $k = m_1 + \ldots + m_p$ effect sizes overall. Denote the jth effect parameter in the ith group by θ_{ij} and its estimate by T_{ij} with (conditional) variance v_{ij}. That is, T_{ij} estimates θ_{ij} with standard error $\sqrt{v_{ij}}$. In most cases, v_{ij} will actually be an estimated variance that is a function of the within-study sample size and the effect-size

estimate in study j. However, unless the within-study sample size is exceptionally small we can treat v_{ij} as known. Therefore, in the rest of this chapter we assume that v_{ij} is known for each study. The sample data from the collection of studies can be represented as in table 12.2.

12.3.1.2 Means Making use of the dot notation from the analysis of variance, define the group mean effect estimate for the ith group $\overline{T}_{i\bullet}$ by

$$\overline{T}_{i\bullet} = \frac{\sum_{j=1}^{m_i} w_{ij} T_{ij}}{\sum_{j=1}^{m_i} w_{ij}}, \; i = 1, \ldots, p, \tag{12.13}$$

where the weight w_{ij} is simply the reciprocal of the variance of T_{ij},

$$w_{ij} = 1/v_{ij}. \tag{12.14}$$

Table 12.2 Effect Size Estimates and Sampling Variances for p Groups of Studies

	Effect Size Estimates	Variances
Group 1		
Study 1	T_{11}	v_{11}
Study 2	T_{12}	v_{12}
.	.	.
.	.	.
.	.	.
Study m_1	T_{1m1}	v_{1m1}
Group 2		
Study 1	T_{21}	v_{21}
Study 2	T_{22}	v_{22}
.	.	.
.	.	.
.	.	.
Study m_2	T_{2m2}	v_{2m2}
.		
.		
.		
Group p		
Study 1	T_{p1}	v_{p1}
Study 2	T_{p2}	v_{p2}
.	.	.
.	.	.
.	.	.
Study m_p	T_{pmp}	v_{pmp}

SOURCE: Authors' compilation.

The grand weighted mean $\overline{T}_{\bullet\bullet}$ is

$$\overline{T}_{\bullet\bullet} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{m_i} w_{ij} T_{ij}}{\sum_{i=1}^{p}\sum_{j=1}^{m_i} w_{ij}}. \tag{12.15}$$

The grand mean $\overline{T}_{\bullet\bullet}$ could also be seen as the weighted mean of the group means $\overline{T}_{1\bullet}, \ldots, \overline{T}_{p\bullet}$

$$\overline{T}_{\bullet\bullet} = \frac{\sum_{i=1}^{p} w_{i\bullet} \overline{T}_{i\bullet}}{\sum_{i=1}^{p} w_{i\bullet}}, \tag{12.16}$$

where the weight $w_{i\bullet}$ is just the sum of the weights for the ith group

$$w_{i\bullet} = w_{i1} + \cdots + w_{im_i}.$$

Thus, $\overline{T}_{i\bullet}$ is simply the weighted mean that would be computed by applying formula (12.1) to the studies in group i and $\overline{T}_{\bullet\bullet}$ is the weighted mean that would be obtained by applying formula (12.1) to all of the studies. If all of the studies in group i estimate a common effect-size parameter $\theta_{i\bullet}$, that is, if $\theta_{i1} = \theta_{i2} = \ldots = \theta_{im_i} = \theta_{i\bullet}$, then $\overline{T}_{i\bullet}$ estimates $\theta_{i\bullet}$. If the studies within the ith group do *not* estimate a common effect parameter, then $\overline{T}_{i\bullet}$ estimates the weighted mean of the effect-size parameters θ_{ij} given by

$$\overline{\theta}_{i\bullet} = \frac{\sum_{j=1}^{m_i} w_{ij} \theta_{ij}}{\sum_{j=1}^{m_i} w_{ij}}, \; i = 1, \ldots, p. \tag{12.17}$$

Similarly, if all of the studies in the collection estimate a common parameter $\overline{\theta}_{\bullet\bullet}$, that is if $\theta_{11} = \ldots = \theta_{1m_1} = \theta_{21} = \ldots = \theta_{pm_p} = \overline{\theta}_{\bullet\bullet}$, then $\overline{T}_{\bullet\bullet}$ estimates $\overline{\theta}_{\bullet\bullet}$. If the studies do *not* all estimate the parameter, then $\overline{T}_{\bullet\bullet}$ can be seen as an estimate of a weighted mean $\overline{\theta}_{\bullet\bullet}$ of the effect parameters given by

$$\overline{\theta}_{\bullet\bullet} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{m_i} w_{ij} \theta_{ij}}{\sum_{i=1}^{p}\sum_{j=1}^{m_i} w_{ij}}. \tag{12.18}$$

Alternatively $\bar{\theta}_{\bullet\bullet}$ can be viewed as a weighted mean of the $\bar{\theta}_{i\bullet}$:

$$\bar{\theta}_{\bullet\bullet} = \frac{\sum_{i=1}^{p} w_{i\bullet}\bar{\theta}_{i\bullet}}{\sum_{i=1}^{p} w_{i\bullet}},$$

where $w_{i\bullet}$ is just the sum of the weights w_{ij} for the ith group as in the alternate expression for $\bar{T}_{\bullet\bullet}$ above.

12.3.1.3 Standard Errors The sampling variances $v_{1\bullet}, \ldots, v_{p\bullet}$ of the group mean effect estimates $\bar{T}_{1\bullet}, \ldots, \bar{T}_{p\bullet}$ are given by the reciprocals of the sums of the weights in each group, that is

$$v_{i\bullet} = \frac{1}{\sum_{j=1}^{m_i} w_{ij}}, i = 1, \ldots p. \qquad (12.19)$$

Similarly the sampling variance $v_{\bullet\bullet}$ of the grand weighted mean $\bar{T}_{\bullet\bullet}$ is given by the reciprocal of the sum of all the weights or

$$v_{\bullet\bullet} = \frac{1}{\sum_{i=1}^{p}\sum_{j=1}^{m_i} w_{ij}}. \qquad (12.20)$$

The standard errors of the group mean effect estimates $\bar{T}_{i\bullet}$ and the grand mean $\bar{T}_{\bullet\bullet}$ are just the square roots of their respective sampling variances.

12.3.1.4 Tests and Confidence Intervals The group means $\bar{T}_{1\bullet}, \ldots, \bar{T}_{p\bullet}$ are assumed to be normally distributed about the respective effect-size parameters $\bar{\theta}_{i\bullet}, \ldots, \bar{\theta}_{p\bullet}$ that they estimate. The fact that these means are normally distributed with the variances given in (12.18) leads to rather straightforward procedures for constructing tests and confidence intervals. For example, to test whether the ith group mean effect $\bar{\theta}_{i\bullet}$ differs from a predefined constant θ_0 (for example, to test whether $\bar{\theta}_{i\bullet} - \theta_0 = 0$) by testing the null hypothesis

$$H_{0i}: \bar{\theta}_{i\bullet} = \theta_0,$$

use the statistic

$$Z_i = \frac{\bar{T}_{i\bullet} - \theta_0}{\sqrt{v_{i\bullet}}} \qquad (12.21)$$

and reject H_0 at level α (that is, decide that the effect parameter differs from θ_0) if the absolute value of Z_i exceeds the 100α percent critical value of the standard normal distribution. For example, for a two-sided test that $\bar{\theta}_{i\bullet} = 0$ at $\alpha = .05$ level of significance, reject the null hypothesis if the absolute value of Z exceeds 1.96. When there is only one group of studies, this test is identical to that described earlier in this chapter and given in (12.5).

Confidence intervals for the group mean effect $\bar{\theta}_{i\bullet}$ can be computed by multiplying the standard error $\sqrt{v_{i\bullet}}$ by the appropriate two-tailed critical value of the standard normal distribution ($C_\alpha = 1.96$ for $\alpha = 0.05$ and 95 percent confidence intervals) then adding and subtracting this amount from the weighted mean effect size $\bar{T}i_{\bullet}$. Thus the $100(1 - \alpha)$ percent confidence interval for $\bar{\theta}_{i\bullet}$ is given by

$$\bar{T}_{i\bullet} - C_\alpha\sqrt{v_{i\bullet}} \le \bar{\theta}_{i\bullet} \le \bar{T}_{i\bullet} + C_\alpha\sqrt{v_{i\bullet}}. \qquad (12.22)$$

Example. Return to the ten studies of gender differences in conformity using the so-called fictitious norm group paradigm. The effect sizes were standardized mean differences classified into three groups on the basis of the percentage of male authors of the research report. Group 1 consisted of two studies having 25 percent of male authorship, group 2 consisted of a single study in which 50 percent of the authors were male, and group 3 consisted of seven studies with all male authorship. The data are presented in table 12.3.

The effect-size estimate T_{ij} for each study, its variance v_{ij}, the weight $w_{ij} = 1/v_{ij}$, $w_{ij}T_{ij}$, and $w_{ij}T_{ij}^2$ (which will be used later) are presented in table 12.3. Using the sums for each group from table 12.3, the weighted mean of effect sizes for the three classes $\bar{T}_{1\bullet}$, $\bar{T}_{2\bullet}$, and $\bar{T}_{3\bullet}$ are given by

$$\bar{T}_{1\bullet} = -9.320/63.895 = -0.146,$$

$$\bar{T}_{2\bullet} = -13.636/45.455 = -0.300,$$

$$\bar{T}_{3\bullet} = 57.605/169.811 = 0.339,$$

and the weighted grand mean effect size is

$$\bar{T}_{\bullet\bullet} = 34.649/279.161 = 0.124.$$

Table 12.3 Data for the Male-Authorship Example

Study	% Male Authors	Group	# of Items	T	v	w	wT	wT^2
1	25%	1	2	–0.330	0.029	34.483	–11.379	3.755
2	25	1	2	0.070	0.034	29.412	2.059	0.144
3	50	2	2	–0.300	0.022	45.455	–13.636	4.091
4	100	3	38	0.350	0.016	62.500	21.875	7.656
5	100	3	30	0.700	0.066	15.152	10.606	7.424
6	100	3	45	0.850	0.218	4.587	3.899	3.314
7	100	3	45	0.400	0.045	22.222	8.889	3.556
8	100	3	45	0.480	0.069	14.493	6.957	3.339
9	100	3	5	0.370	0.051	19.608	7.255	2.684
10	100	3	5	–0.060	0.032	31.250	–1.875	0.113

SOURCE: Authors' tabulation based on data from Eagly and Carli 1981.

The variances $v_{1\bullet}$, $v_{2\bullet}$, and $v_{3\bullet}$ of $\bar{T}_{1\bullet}$, $\bar{T}_{2\bullet}$, and $\bar{T}_{3\bullet}$ are given by

$$v_{1\bullet} = 1/63.895 = 0.016,$$
$$v_{2\bullet} = 1/45.455 = 0.022,$$
$$v_{3\bullet} = 1/169.811 = 0.006,$$

and the variance $v_{\bullet\bullet}$ of $\bar{T}_{\bullet\bullet}$ is

$$v_{\bullet\bullet} = 1/279.161 = 0.00358.$$

Using formula (12.22) with $C_{.05} = 1.960$, the limits of the 95 percent confidence interval for the group mean parameter $\bar{\theta}_{1\bullet}$ are given by

$$-0.146 \pm 1.960\sqrt{0.016} = -0.146 \pm 0.245.$$

Thus the 95 percent confidence interval for $\bar{\theta}_{1\bullet}$ is given by

$$-0.391 \le \bar{\theta}_{1\bullet} \le 0.099.$$

Because this confidence interval contains zero, or alternately, because the test statistic

$$Z_1 = |-0.146| / \sqrt{0.016} < 1.960,$$

we cannot reject the hypothesis that $\bar{\theta}_{1\bullet} = 0$ at the $\alpha = 0.05$ level of significance. Similarly 95 percent confidence intervals for the group mean parameters $\bar{\theta}_{2\bullet}$ and $\bar{\theta}_{3\bullet}$ are given by

$$-0.591 = -0.300 - 1.960\sqrt{0.022} \le \bar{\theta}_{2\bullet} \le -0.300 + 1.960\sqrt{0.022} = -0.009$$

and

$$0.189 = 0.339 - 1.960\sqrt{0.006} \le \bar{\theta}_{3\bullet} \le 0.339 + 1.960\sqrt{0.006} = 0.489.$$

Thus we see that the mean effect size for group 2 is significantly less than zero that for group 3 is significantly greater than zero and that for group 1 was not significantly different from zero.

12.3.1.5 Tests of Heterogeneity In the analysis of variance, tests for systematic sources of variance are constructed from sums of squared deviations from means. That the effects due to different sources of variance partition the sums of squares leads to the interpretation that the total variation about the grand mean is partitioned into parts that arise from between-group and within-group sources. The analysis of variance for effect sizes has a similar interpretation. The total heterogeneity statistic

$$Q_T = \sum_{i=1}^{p}\sum_{j=1}^{m_i} w_{ij}(T_{ij} - T_{\bullet\bullet})^2$$

(the weighted total sum of squares about the grand mean) is partitioned into a between-groups-of-studies part Q_B (the weighted sum of squares of group means about the grand mean) and a within-groups-of-studies part Q_W (the total of the weighted sum of squares of the individual effect estimates about the respective group means). These statistics Q_B and Q_W yield direct omnibus tests of variation across groups in mean effects and variation within groups of individual effects.

12.3.1.5.1 An Omnibus Test for Between-Groups Differences. To test the hypothesis that group mean effect sizes do not vary, that is, to test

$$H_0: \overline{\theta}_{1\bullet} = \overline{\theta}_{2\bullet} = \ldots = \overline{\theta}_{p\bullet},$$

we use the between-group heterogeneity statistic Q_B defined by

$$Q_B = \sum_{i=1}^{p} w_{i\bullet}\left(\overline{T}_{i\bullet} - \overline{T}_{\bullet\bullet}\right)^2 \qquad (12.23)$$

where $w_{i\bullet}$ is the reciprocal of the variance of $v_{i\bullet}$. Note that Q_B is just the weighted sum of squares of group mean effect sizes about the grand mean effect size. When the null hypothesis of no variation across group mean effect sizes is true, Q_B has a chi-square distribution with $(p - 1)$ degrees of freedom. Hence we test H_0 by comparing the obtained value of Q_B with the upper tail critical values of the chi-square distribution with $(p - 1)$ degrees of freedom. If Q_B exceeds the $100(1 - \alpha)$ percent point of the chi-square distribution (for example, $C_{.05} = 18.31$ for 10 degrees of freedom and $\alpha = 0.05$), H_0 is rejected at level α and between-group differences are significant.

This test is analogous to the omnibus *F*-test for variation in group means in a one-way analysis of variance in a primary research study. It differs in that Q_B, unlike the *F*-test, incorporates an estimate of unsystematic error in the form of the weights. Thus, no separate error term (such as the mean square within groups as in the typical analysis of variance) is needed and the sum of squares can be used directly as a test statistic.

12.3.1.5.2 An Omnibus Test for Within-Group Variation in Effects. To test the hypothesis that population effect sizes within the groups of studies do not vary, that is to test

$$H_0: \begin{array}{ll} \theta_{11} & = \ldots = \theta_{1m_1} = \theta_{1\bullet} \\ \theta_{21} & = \ldots = \theta_{2m_2} = \theta_{2\bullet} \\ \cdot & \cdot \quad\quad \cdot \\ \cdot & \cdot \quad\quad \cdot \\ \cdot & \cdot \quad\quad \cdot \\ \theta_{p1} & = \ldots = \theta_{pm_p} = \theta_{p\bullet} \end{array}$$

use the within-group homogeneity statistic Q_W given by

$$Q_W = \sum_{i=1}^{p}\sum_{j=1}^{m_i} w_{ij}(T_{ij} - T_{i\bullet})^2, \qquad (12.24)$$

where the w_{ij} are the reciprocals of the v_{ij}, the sampling variances of the T_{ij}. When the null hypothesis of perfect homogeneity of effect-size parameters is true, Q_W has a chi-square distribution with $(k - p)$ degrees of freedom where $k = m_1 + m_2 + \ldots + m_p$ is the total number of studies in the observed sample. Therefore, within-group homogeneity at significance level α is rejected if the computed value of Q_W exceeds the $100(1 - \alpha)$ percent point (the upper tail critical value) of the chi-square distribution with $(k - p)$ degrees of freedom.

Although Q_W provides an overall test of within-group variability in effects, it is actually the sum of p separate (and independent) within-group heterogeneity statistics, one for each of the p groups of effects. Thus

$$Q_W = Q_{W_1} + Q_{W_2} + \cdot \ldots + Q_{W_p}, \qquad (12.25)$$

where each Q_{W_i} is just the heterogeneity statistic Q given in formula (12.6). In the notation used here, Q_{W_i} is given by

$$Q_{W_i} = \sum_{j=1}^{m_i} w_{ij}\left(T_{ij} - \overline{T}_{i\bullet}\right)^2. \qquad (12.26)$$

These individual within-group statistics are often useful in determining which groups are the major sources of within-group heterogeneity and which have relatively homogeneous effects. For example, in analyses of the effects of study quality in treatment-control studies, study-effect estimates might be placed into two groups: those from quasi-experiments and those from randomized experiments. The effect sizes within the two groups might be quite heterogeneous overall, leading to a large

Q_W, but most of that heterogeneity might arise within the group of quasi-experiments so that Q_{W_1} (the statistic for quasi-experiments) would indicate great heterogeneity, but Q_{W_2} (the statistic for randomized experiments) would indicate relative homogeneity (see, for example, Hedges 1983a).

If the effect-size parameters within the ith group of studies are homogeneous, that is, if $\theta_{i1} = \ldots = \theta_{im_i}$, then θ_{W_i} has the chi-square distribution with $m_i - 1$ degrees of freedom. Thus the test for homogeneity of effects within the ith group at significance level α consists of rejecting the hypothesis of homogeneity if Q_{W_i} exceeds the $100(1 - \alpha)$ percent point of the chi-square distribution with $(m_i - 1)$ degrees of freedom.

It is often convenient to summarize the relationships among the heterogeneity statistics via a table analogous to an ANOVA source table (see table 12.4).

12.3.1.5.3 The Partitioning of Heterogeneity. There is a simple relationship among the total homogeneity statistic Q given in formula (12.6) and the between-and within-group fit statistics discussed in this section. This relationship corresponds to the partitioning of the sums of squares in ordinary analysis of variance. That is $Q = Q_B + Q_W$.

One interpretation is that the total heterogeneity about the mean Q is partitioned into between-group heterogeneity Q_B and within-group heterogeneity Q_W. The ideal is to select independent (grouping) variables that explain variation (heterogeneity) so that most of the total heterogeneity is between-groups and relatively little remains within groups of effects. Of course, the grouping variable must, in principle, be chosen a priori (that is, before examination of the effect sizes) to ensure that tests for the significance of group effects do not capitalize on chance.

Table 12.4 Heterogeneity Summary Table

Source	Statistic	Degrees of Freedom
Between groups	Q_{BET}	$p - 1$
Within groups	Q_{W1}	$m_1 - 1$
Group 1	Q_{W2}	$m_2 - 1$
Group 2	.	.
.	.	.
.	.	.
.	Q_{W_p}	$m_p - 1$
Group p		
Total within groups	Q_W	$k - p$
Overall	Q	$k - 1$

SOURCE: Authors' compilation.
NOTE: Here $k = m_1 + m_2 + \ldots + m_p$.

12.3.1.6 Computing the Analysis Although Q_B and Q_W can be computed via a computer program for weighted ANOVA, the weighted cell means and their standard errors cannot generally be obtained this way. Computational formulas can greatly simplify direct calculation of Q_B and Q_W as well as the cell means and their standard errors. These formulas are analogous to computational formulas in the analysis of variance and enable the computation of all of the statistics in one pass through the data (for example, by a packaged computer program). The formulas are expressed in terms of totals (sums) across cases of the weights, of the weights times the effect estimates, and of the weights times the squared effect estimates. Define

$$TW_i = \sum_{j=1}^{m_i} w_{ij}, \qquad TW_\bullet = \sum_{i=1}^{p} TW_i,$$

$$TWD_i = \sum_{j=1}^{m_i} w_{ij} T_{ij}, \qquad TWD_\bullet = \sum_{i=1}^{p} TWD_i,$$

$$TWDS_i = \sum_{j=1}^{m_i} w_{ij} T_{ij}^2, \qquad TWDS_\bullet = \sum_{i=1}^{p} TWDS_i,$$

where $w_{ij} = 1/v_{ij}$ is just the weight for T_{ij}. Then the overall heterogeneity statistic Q is

$$Q = TWDS_\bullet - (TWD_\bullet)^2 / TW_\bullet. \qquad (12.27)$$

Each of the within-group heterogeneity statistics is given by

$$Q_{W_i} = TWDS_i - (TWD_i)^2 / TW_i, i = 1, \ldots, p. \qquad (12.28)$$

The overall within-group homogeneity statistic is obtained as $Q_W = Q_{W_1} + Q_{W_2} + \cdot \ldots + Q_{W_p}$. The between-groups heterogeneity statistic is obtained as $Q_B = Q - Q_W$. The weighted overall mean effects and its variance are

$$\overline{T}_{\bullet\bullet} = TWD_\bullet / TW_\bullet, \qquad v_{\bullet\bullet} = 1/TW_\bullet,$$

and the weighted group means and their variances are

$$v_{i\bullet} = TWD_i / TW_i, i = 1, \ldots, p,$$

and

$$v_{i\bullet} = 1/TW_i, i = 1, \ldots, p.$$

The omnibus test statistics Q, Q_B, and Q_W can also be computed using a weighted analysis of variance program. The grouping variable is used as the factor in the weighted ANOVA, the effect-size estimates are the observations, and the weight given to effect size T_{ij} is just w_{ij}. The weighted between-group or model *sum* of squares is exactly Q_B, the weighted within-group or residual *sum* of squares is exactly Q_W, and the corrected total *sum* of squares is exactly Q.

Example. Return to the data from studies of gender differences in conformity given in table 12.1. Using formula (12.6) and the sums given in table 12.1, the overall heterogeneity statistic $Q = Q_T$ is

$$Q = 36.076 - (34.649)^2/279.161 = 31.776.$$

Using the sums given in table 12.1 in formula (12.28), the within-group heterogeneity statistics Q_{W_1}, Q_{W_2}, and Q_{W_3} are

$$Q_{W_1} = 3.899 - (-9.320)^2/63.895 = 2.540,$$

$$Q_{W_2} = 4.091 - (13.636)^2/45.455 = 0.0004,$$

$$Q_{W_3} = 28.086 - (57.605)^2/169.811 = 8.545.$$

The overall within-group heterogeneity statistic is therefore

$$Q_W = 2.540 + 0.000 + 8.545 = 11.085.$$

Because 11.085 does not exceed 14.067, the 95 percent point of the chi-square distribution with 10 – 3 = 7 degrees of freedom, we do not reject the hypothesis that the effect-size parameters are homogeneous within the groups. In fact, a value this large would occur between 10 and 25 percent of the time due to chance even with perfect homogeneity of effect *parameters*. Thus we conclude that no evidence indicates that effect sizes differ within groups.

The between-group heterogeneity statistic is calculated as

$$Q_B = Q - Q_W = 31.776 - 11.085 = 20.691.$$

Because 20.691 exceeds 5.991, the 95 percent point of the chi-square distribution with 3 – 1 = 2 degrees of freedom, we reject the null hypothesis of no variation in effect size across studies with different proportions of male authors. In other words, there is a statistically significant relationship between the percentage of male authors and effect size.

12.3.1.7 Comparisons or Contrasts Among Mean Effects Omnibus tests for differences among group means can reveal that the mean effect parameters are not all the same, but they are not useful for revealing the specific pattern of mean differences that might be present. For example, the Q_B statistic might reveal that there was variation in mean effects when the effects were grouped according to type of treatment, but the omnibus statistic gives no insight about *which* types of treatment (which groups) were associated with the largest effect size. In other cases, the omnibus test statistic may not be significant, but we may wish to test for a specific a priori difference that the omnibus test may not have been powerful enough to detect. In conventional analysis of variance, contrasts or comparisons are used to explore the differences among group means. Contrasts can be used in precisely the same way to examine patterns among group mean effect sizes in meta-analysis. In fact, all of the strategies used for selecting contrasts in ANOVA (such as orthogonal polynomials to estimate trends, and Helmert contrasts to discover discrepant groups) are also applicable in meta-analysis.

A contrast parameter is just a linear combination of group means

$$\gamma = c_1\bar{\theta}_{1\bullet} + \ldots + c_p\bar{\theta}_{p\bullet}. \qquad (12.29)$$

where the coefficients $c_1, \ldots, c_p$ (called the contrast coefficients) are known constants that satisfy the constraint $c_1 + \ldots + c_p = 0$ and are chosen so that the value of the contrast will reflect a particular comparison or pattern of interest. For example the coefficients $c_1 = 1$, $c_2 = -1$, $c_3 = \ldots = c_p = 0$ might be chosen so that the value of the contrast is the difference between the mean $\bar{\theta}_{1\bullet}$ of group 1 and the mean $\bar{\theta}_{2\bullet}$ of group 2. Sometimes we refer to a contrast among population means as a population contrast or a contrast parameter to emphasize that it is a function of population parameters and to distinguish it from *estimates* of the contrast. The contrast parameter specified by coefficients $c_1, \ldots, c_p$ is usually estimated by a sample contrast

$$g = c_1\bar{T}_{1\bullet} + \ldots + c_p\bar{T}_{p\bullet}. \qquad (12.30)$$

The estimated contrast g has a normal sampling distribution with variance v_g given by

$$v_g = c_1^2 v_{1\bullet} + \ldots + c_p^2 v_{p\bullet}. \qquad (12.31)$$

Although this notation for contrasts suggests that they compare group mean effects, they can be used to compare individual studies (groups consisting of a single study) or to compare a single study with a group mean. All that is required is the appropriate definition of the groups involved.

12.3.1.7.1 Confidence Intervals and Tests of Significance. Because the estimated contrast g has a normal distribution with known variance v_g, confidence intervals and tests of statistical significance are relatively easy to construct. Just as with contrasts in ordinary analysis of variance, however, test procedures differ depending on whether the contrasts were planned or selected using information from the data. Procedures for testing planned comparisons and for constructing nonsimultaneous confidence intervals are given in this section. Procedures for testing post hoc contrasts (contrasts selected using information from the data) follow.

12.3.1.7.2 Planned Comparisons. Confidence intervals for the contrast parameter γ are computed by multiplying the standard error of g, $\sqrt{v_g}$, by the appropriate two-tailed critical value of the standard normal distribution (C_α=1.96 for α=.05 and 95 percent confidence intervals) and adding and subtracting this amount from the estimated contrast g. Thus the $100(1-\alpha)$ percent confidence interval for the contrast parameter γ is

$$g - C_\alpha \sqrt{v_g} \le \gamma \le g + C_\alpha \sqrt{v_g}. \qquad (12.32)$$

Alternatively a (two-sided) test of the null hypothesis that $\gamma = 0$ uses the statistic

$$X^2 = g^2 / v_g. \qquad (12.33)$$

If X^2 exceeds the $100(1-\alpha)$ percent point of the chi-square distribution with one degree of freedom, reject the hypothesis that $\gamma = 0$ and declare the contrast to be significant at the level of significance α.

12.3.1.7.3 Post Hoc Contrasts and Simultaneous Tests. Situations often occur when several contrasts among group means are of interest. If several tests are made at the same nominal significance level α, the chance that at least one of the tests will reject (when all of the relevant null hypotheses are true) is generally greater than α and can be considerably greater if the number of tests is large. Similarly, the probability that tests will reject may also be greater than the nominal significance level for contrasts that are selected because they "appear to stand out" when examining the data. Simultaneous and post hoc testing procedures are designed to address these problems by assuring that the probability of at least one type I error is controlled at a preset significance level α. Many simultaneous test procedures have been developed (see Miller 1981). We now discuss the application of two of these procedures to contrasts in meta-analysis (see also Hedges and Olkin 1985).

The simplest simultaneous test procedure is the Bonferroni method. It exacts a penalty for simultaneous testing by requiring a higher level of significance from each individual contrast for it to be declared significant in the simultaneous test. If a number $L \ge 1$ of contrasts are to be tested simultaneously at level α, the Bonferroni test requires that any contrast be significant at (nonsimultaneous) significance level α/L in order to be declared significant at level α in the simultaneous analysis. For example, if $L = 5$ contrasts were tested at simultaneous significance level $\alpha = 0.05$, any one of the contrasts would have to be individually significant at the 0.01 = 0.05/5 level (that is, X^2 would have to exceed 6.635, the 99 percent point of the chi-square distribution with df = 1) to be declared significant at $\alpha = 0.05$ level by the simultaneous test.

The Bonferroni method can be used as a post hoc test if the number of contrasts L is chosen as the number of contrasts that *could* have been conducted. This procedure works well when all contrasts conducted are chosen from a well-defined class of contrasts. For example, if there are four groups, there are six possible pairwise contrasts, so the Bonferroni method is applied to any pairwise contrast chosen post hoc by treating it as one of six contrasts examined simultaneously. If number L of comparisons (or possible comparisons) is large, the Bonferroni method can be quite conservative given that it rejects only if a contrast has a very low significance value.

An alternative test procedure that is a generalization of the Scheffé method from the analysis of variance can be used for both post hoc and simultaneous testing (see Hedges and Olkin 1985). This procedure for testing contrasts at simultaneous significance level α consists of computing the statistic X^2 given in (12.33) for each contrast and rejecting the null hypothesis whenever X^2 exceeds the $100(1-\alpha)$ percentage point of the chi-square distribution with L' degrees of freedom, where L' is smaller of L (the number of contrasts) or $p - 1$ (the number of groups minus one).

When the number of contrasts (or potential contrasts) is small, simultaneous tests based on the Bonferroni method will usually be more powerful. When the number of contrasts is large, the Scheffé method will be.

Example. Continuing the analysis of standardized mean difference data on gender difference in conformity, recall that there are three groups of effects: those in which 25 percent, 50 percent, and 100 percent (respectively) of the authors are male. To contrast the mean of the effects of group 1 (25 percent male authors) with those of group 3 (100 percent male authors), use the contrast coefficients

$$c_1 = -1.0, c_2 = 0.0, c_3 = 1.0.$$

The value of the contrast estimate g is

$$g = -1.0(-0.146) + 0.0(-0.300) + 1.0(0.339) = 0.485,$$

with an estimated variance of

$$v_g = (-1.0)^2(0.016) + (0.0)^2(0.022) + (1.0)^2(0.006) = 0.022.$$

Hence a 95 percent confidence interval for $\gamma = \bar{\theta}_{3\bullet} - \bar{\theta}_{1\bullet}$ is

$$0.194 = 0.485 - 1.960\sqrt{0.022} \le \gamma \le 0.485 + 1.960\sqrt{0.022} = 0.776$$

Because this confidence interval does not contain zero, or alternatively, because

$$Z = 0.485/\sqrt{0.022} = 3.270$$

exceeds 1.96, we reject the hypothesis that $\gamma = 0$ and declare the contrast statistically significant at the $\alpha = 0.05$ level. Notice that the chi-square test is $X^2 = Z^2 = 10.692$.

12.3.2 Mixed-Models Analyses

The one-factor random-model analysis is analogous to the one-factor fixed-effects analysis. Like the fixed-effects analysis, it is used when we wish to determine whether a particular discrete characteristic of studies is related to effect size. The difference is that in the mixed model we wish to incorporate the effects of between-study but within-class variation of effect sizes as uncertainty in the analysis (Konstantopoulos 2013). Under the mixed model, differences between studies that lead to differences in effects are regarded as random (systematic heterogeneity between studies). The random- or mixed-effects models are appropriate when unconditional inferences need to be drawn. In such cases, the observed studies in the sample are viewed to be randomly selected (representative) from a larger population of studies and the inference drawn is about the population of studies from which the observed studies were randomly selected.

12.3.2.1 Models and Notation Use the same notation for the effect-size parameters, estimate, and variances as in the fixed-effects analysis. Thus there are p disjoint classes of effects with m_1 effects in the first class, m_2 effects in the second class, . . . , and m_p effects in the pth class and a total of $k = m_1 + \ldots + m_p$ effect sizes overall. Denote the jth effect parameter in the ith class by θ_{ij} and its estimate by T_{ij} with (conditional) variance v_{ij}. That is, T_{ij} estimates θ_{ij} with (conditional) standard error $\sqrt{v_{ij}}$. Thus the data from the studies consist of the effect-size estimates and their standard errors (or conditional variances) as shown in table 12.2.

We will assume that T_{ij} is normally distributed about θ_{ij}, that is

$$T_{ij}|\theta_{ij} \sim N(\theta_{ij}, v_{ij}), \qquad j = 1, \ldots, m_i; i = 1, \ldots, p.$$

Unlike the fixed-effects model, the mixed model treats the θ_{ij} as being composed of both fixed and random components. That is,

$$\theta_{ij} = \mu_{i\bullet} + \xi_{ij} \tag{12.34}$$

where $\mu_{i\bullet}$ is the mean of the population of effect parameters in the *ith* class and the ξ_{ij} are independently and identically distributed random effects with mean 0 and variance τ_i^2. Thus the within-class variance of the θ_{ij} or alternatively of the ξ_{ij} is τ_i^2 (between-study within-class variance).

Thus the unknown parameters are the class means $\mu_{1\bullet}$, $\mu_{2\bullet}$, . . ., $\mu_{p\bullet}$ and the within-class variance components τ_1^2, τ_2^2, . . . , τ_p^2. The object of the analysis is to estimate the within-class variance components (or perhaps a pooled variance component across classes) and the means and test various hypotheses about them.

In most cases, v_{ij} will actually be an estimated variance that is a function of the within-study sample size and the effect-size estimate. However, unless the within-study sample size is exceptionally small, we can treat v_{ij} as known. Therefore, in the rest of this chapter we assume that v_{ij} is known for each study.

12.3.2.1.1 Homogeneity of Within-Class Variance Components. Note that this mixed model does not necessarily imply that the within-class variance components

are the same for each class. However, homogeneity of variance components is a convenient assumption that is often consistent with the data. In other cases, the number of effect estimates in each class is so small and hence the information about the actual magnitude of the variance components so poor that the data do not provide a strong basis for deciding that variance components differ across classes. If, for either reason, one can treat the effect-size data as if the within-class variance components are homogeneous, analyses are simplified by the estimation of one pooled within-class variance component instead of one for each class. That case implies adding the restriction that all variance components are equal to a common value τ^2, that is,

$$\tau_i^2 = \tau^2, i = 1, \ldots, p.$$

However, in some cases, homogeneity of within-class variance components is not a reasonable assumption, either because it is inconsistent with the nature of the problem or because of the observed data itself. In such cases, each variance component would need to be estimated separately. With a large enough number of studies in each class, separate variance components can be estimated and incorporated into analyses, but we do not describe these methods here.

12.3.2.2 Between-Studies Variance Components The first step in the random- or mixed-effects analysis is to estimate the between-study, within-class variance component of the effect sizes (or multiple variance components if these are not the same across classes). One estimation method, the method of moments is distribution free in the sense that the estimation method does not depend on the distribution of the random effects (the ξ_{ij}'s). Other methods, such as full or restricted maximum likelihood estimation, can be more efficient, but depend on the assumption that the random effects are normally distributed.

The method of moments estimate of the variance component τ^2 does not depend on assumptions about the form of the distribution of the random effects. However, the sampling distributions of the test statistics do depend on the assumption that the random effects are normally distributed. When the variance of the random effects (the between-study within-class variance component τ^2) is small in respect to the sampling error variances (the v_{ij}), the effect of the random-effects variance on the distribution of the class means will be relatively minor. However, when the random-effects variance is larger (for example, much larger than the typical v_{ij}), the effect on the distribution of the class means and test statistics based on the random-effects model may be substantial.

12.3.2.2.1 Distribution-Free Estimation. The distribution-free estimate of the between-study within-class variance component is analogous to the estimation of the between-study variance component discussed earlier in this chapter for a single group of studies. If there is no reason to assume that between-study variation differs across groups, it is sensible to estimate a single between-study variance component by pooling the within-class estimates from every class.

The pooled estimate of the with-class variance component is given by

$$\hat{\tau}^2 = \left(Q_W - \sum_{i=1}^{p} m_i + p \right) \Big/ \sum_{i=1}^{p} c_i \qquad (12.35)$$

whenever (12.35) is positive and 0 otherwise, where Q_W is the within-classes heterogeneity statistic computed in the fixed-effects analysis and given in (12.26), c_i is given by

$$c_i = \sum_{j=1}^{m_i} w_{ij} - \left(\sum_{j=1}^{m_i} w_{ij}^2 \Big/ \sum_{j=1}^{m_i} w_{ij} \right), i = 1, \ldots, p, \qquad (12.36)$$

and $w_{ij} = 1/v_{ij}$, the same weights used in the fixed-effects analysis.

The pooled with-class variance component is an estimate of the variance of the effect-size parameters with classes. As such, it provides a descriptive statistic to describe the amount of true variation among study results within classes. A more easily interpreted descriptive statistic is actually the square root of this variance component, $\hat{\tau}$.

12.3.2.2.2 Full Maximum Likelihood Estimation. The method of maximum likelihood can also be used to estimate of τ^2 under a model of homogeneous τ^2 across classes. Begin with a preliminary estimate $\hat{\tau}^2_{[0]}$ of τ^2 such as that from (12.35) and define $\overline{T}^*_{1\bullet[0]} = \overline{T}^*_{1\bullet}, \ldots, \overline{T}^*_{p\bullet[0]} = \overline{T}^*_{p\bullet}$. Compute the variance component estimate and the weighted means at the $(s + 1)$st iteration from $\overline{T}^*_{1\bullet[s]}, \ldots, \overline{T}^*_{p\bullet[s]}$ and $\hat{\tau}^2_{[s]}$ via

$$\hat{\tau}^2_{[s+1]} = \frac{\sum_{i=1}^{p} \sum_{j=1}^{m_i} \left(w^*_{ij[s]} \right)^2 \left[\left(T_{ij} - \overline{T}^*_{i\bullet[s]} \right)^2 - v_{ij} \right]}{\sum_{i=1}^{p} \sum_{j=1}^{m_i} \left(w^*_{ij[s]} \right)^2} \qquad (12.37)$$

and

$$\overline{T}^{*}_{i\bullet[s+1]} = \sum_{j=1}^{m_i} w^{*}_{ij[s]} T_{ij} \Big/ \sum_{j=1}^{m_i} w^{*}_{ij[s]}, \; i = 1, \ldots, p, \qquad (12.38)$$

where the weight of the jth study in the ith class at the sth iteration $w^{*}_{ij[s]}$ is

$$w^{*}_{ij[s]} = 1\Big/\left(v_{ij} + \hat{\tau}^{2}_{[s]}\right). \qquad (12.39)$$

The iterative process continues until the change in the estimates between two consecutive iterations is negligible (often only a few iterations). Then, convergence is achieved and the mean and variance estimates are the full maximum likelihood estimates.

12.3.2.2.3 Restricted Maximum Likelihood Estimation. The method of restricted maximum likelihood can also be used to obtain estimates τ^2. It is quite similar to that described for obtaining full maximum likelihood estimates of τ^2. To use this method, start with an initial value of τ^2 as $\hat{\tau}^{2}_{[0]}$, such as the method of moments estimator given in (12.35), and let $\overline{T}^{*}_{1\bullet[0]} = \overline{T}^{*}_{1\bullet}, \ldots, \overline{T}^{*}_{p\bullet[0]} = \overline{T}^{*}_{p\bullet}$. Compute the variance component estimate at the $(s + 1)$st iteration from $\overline{T}^{*}_{1\bullet[s]}, \ldots, \overline{T}^{*}_{p\bullet[s]}$ and $\hat{\tau}^{2}_{[s]}$ via

$$\hat{\tau}^{2}_{[s+1]} = \frac{\displaystyle\sum_{i=1}^{p}\sum_{j=1}^{m_i}\left(w^{*}_{ij[s]}\right)^{2}\left[\left(T_{ij} - \overline{T}^{*}_{i\bullet[s]}\right)^{2} - v_{ij}\right] + \sum_{i=1}^{p}\left[\sum_{j=1}^{m_i}\left(w^{*}_{ij[s]}\right)^{2} \Big/ \sum_{j=1}^{m_i} w^{*}_{ij[s]}\right]}{\displaystyle\sum_{i=1}^{p}\sum_{j=1}^{m_i}\left(w^{*}_{ij[s]}\right)^{2}}. \qquad (12.40)$$

Next compute the weighted means $\overline{T}^{*}_{1\bullet[s+1]}, \ldots, \overline{T}^{*}_{p\bullet[s+1]}$ at the $(s + 1)$st iteration from $\hat{\tau}^{2}_{[s+1]}$ using (12.38). The sequence of estimates through iterations eventually converges to the restricted maximum likelihood estimate of τ^2 (see Goldstein 1989). Notice that the difference between equations (12.37) and (12.40) is the second term of the numerator, which adds a small constant to the estimate of τ^2 at each iteration. The function of this constant is to reduce bias in the squared residual as an estimate of τ^2. As in the case of full maximum likelihood, the iteration proceeds until the estimates of τ^2 converge (that is, change negligibly between two consecutive iterations).

Example. Return to the data from studies of gender differences in conformity given in Dataset I given in table 12.3. Using the sums given in table 12.1, we compute the constant given in (12.36) for each group as

$$c_1 = 63.895 - (2054.113/63.895) = 31.747,$$
$$c_2 = 45.455 - (2066.116/45.455) = 0.0009,$$

and

$$c_3 = 169.811 - (6221.757/169.811) = 133.172.$$

Using the value of Q_W computed earlier, namely $Q_W = 11.085$, we obtain the distribution-free estimate of τ^2 as

$$\hat{\tau}^{2} = (11.085 - 10 + 3)/(31.747 + 0.0009 + 133.172) = 0.025.$$

The full and restricted maximum likelihood estimates of τ^2 are zero.

12.3.2.3 Means We use of the same "dot notation" used in the fixed-effects analysis discussed earlier, except that we denote the quantities computed in the mixed model analysis with an asterisk. Thus the group mean effect estimate for the ith group is denoted by $\overline{T}^{*}_{i\bullet}$ and is given by

$$\overline{T}^{*}_{i\bullet} = \sum_{j=1}^{m_i} w^{*}_{ij} T_{ij} \Big/ \sum_{j=1}^{m_i} w^{*}_{ij}, \; i = 1, \ldots, p, \qquad (12.41)$$

where the weight w^{*}_{ij} is simply the reciprocal of the total (conditional plus unconditional) variance of T_{ij},

$$w^{*}_{ij} = 1\Big/\left(v_{ij} + \hat{\tau}^{2}\right). \qquad (12.42)$$

The grand weighted mean $\overline{T}^{*}_{\bullet\bullet}$ is

$$\overline{T}^{*}_{\bullet\bullet} = \sum_{i=1}^{p}\sum_{j=1}^{m_i} w^{*}_{ij} T_{ij} \Big/ \sum_{i=1}^{p}\sum_{j=1}^{m_i} w^{*}_{ij} = \sum_{i=1}^{p} w^{*}_{i\bullet}\overline{T}^{*}_{i\bullet} \Big/ \sum_{i=1}^{p} w^{*}_{i\bullet}, \qquad (12.43)$$

where the weight $w^{*}_{i\bullet}$ is just the sum of the weights for the ith^{e} group

$$w^{*}_{i\bullet} = w^{*}_{i1} + w^{*}_{i2} + \cdot \ldots \cdot + w^{*}_{im_i}.$$

Thus $\overline{T}^*_{i\bullet}$ is simply the weighted mean that would be computed by applying formula (12.1) with random-effects weights to the studies in group i (except possibly that the variance component estimate was pooled across groups). However, $\overline{T}^*_{\bullet\bullet}$ is not the weighted mean that would be obtained by applying formula (12.1) with random-effects weights to all of the studies ignoring class membership. The reason is that any between-class effects are treated in this mixed model as fixed effects and do not contribute to the overall between-studies variance component that would be calculated if between-class effects were ignored entirely.

The estimates $\overline{T}^*_{1\bullet}, \ldots, \overline{T}^*_{p\bullet}$ are estimates of the within-class means $\mu_{1\bullet}, \ldots, \mu_{p\bullet}$, respectively. If all of the classes in the collection estimate a common parameter $\mu_{\bullet\bullet}$, that is if $\mu_{1\bullet} = \ldots = \mu_{p\bullet} = \mu_{\bullet\bullet}$, then $\overline{T}^*_{\bullet\bullet}$ estimates $\mu_{\bullet\bullet}$. If all of the classes do *not* all estimate the same parameter, then $\overline{T}^*_{\bullet\bullet}$ can be seen as an estimate of a weighted mean $\mu_{\bullet\bullet}$ of the effect parameters given by

$$\overline{\mu}^*_{\bullet\bullet} = \sum_{i=1}^{p} w^*_{i\bullet}\mu_{i\bullet} \Big/ \sum_{i=1}^{p} w^*_{i\bullet}. \qquad (12.44)$$

where $w^*_{i\bullet}$ is just the sum of the weights w^*_{ij} for the *ith* group as in the alternate expression for $\overline{T}^*_{\bullet\bullet}$ above.

12.3.2.4 Standard Errors The sampling variances $v^*_{1\bullet}, \ldots, v^*_{p\bullet}$ of the group mean effect estimates $\overline{T}^*_{1\bullet}, \ldots, \overline{T}^*_{p\bullet}$ are given by the reciprocal of the sum of the weights in each group, that is

$$v^*_{i\bullet} = 1 \Big/ \sum_{j=1}^{m_i} w^*_{ij}, \; i = 1, \ldots, p. \qquad (12.45)$$

Similarly, the sampling variance $v^*_{\bullet\bullet}$ of the grand weighted mean is given by the reciprocal of the sum of all the weights or

$$v^*_{\bullet\bullet} = 1 \Big/ \sum_{i=1}^{p}\sum_{j=1}^{m_i} w^*_{ij} = 1 \Big/ \sum_{i=1}^{p} w^*_{i\bullet} \qquad (12.46)$$

The standard errors of the group mean effect estimates $\overline{T}^*_{i\bullet}$ and the grand mean $\overline{T}^*_{\bullet\bullet}$ are just the square roots of their respective sampling variances. Note that whenever the between studies within-classes variance component (estimate) $\hat{\tau}^2 > 0$, the standard errors $\sqrt{v^*_{1\bullet}}, \ldots, \sqrt{v^*_{1\bullet}}$, of the class means estimated under the mixed model will be larger than $\sqrt{v_{1\bullet}}, \ldots, \sqrt{v_{p\bullet}}$, the standard errors of the corresponding class means estimated under the fixed-effects model. If $\hat{\tau}^2 = 0$, the standard errors of the fixed- and mixed-effects model will be identical.

12.3.2.5 Tests and Confidence Intervals If the random effects are approximately normally distributed, the group means $\overline{T}^*_{1\bullet}, \ldots, \overline{T}^*_{p\bullet}$ are normally distributed about the respective effect-size parameters $\mu_{1\bullet}, \ldots, \mu_{p\bullet}$ that they estimate. As in the fixed-effects case, that these means are normally distributed with the variances given in equation (12.45) leads to rather straightforward procedures for constructing tests and confidence intervals. For example, to test whether the ith group mean effect $\mu_{i\bullet}$ differs from a predefined constant μ_0 (for example to test if $\mu_{i\bullet} - \mu_0 = 0$) by testing the null hypothesis

$$H_0 : \mu_{i\bullet} = \mu_0,$$

use the statistic

$$Z^*_{i\bullet} = \left(\overline{T}^*_{i\bullet} - \mu_0\right) \Big/ \sqrt{v^*_{i\bullet}}, \qquad (12.47)$$

and reject H_0 at level α (that is, decide that the effect parameter differs from μ_0) if the absolute value of $Z^*_{i\bullet}$ exceeds the 100α percent critical value of the standard normal distribution. For example, for a two-sided test that $\mu_{i\bullet} = 0$ at $\alpha = 0.05$ level of significance, reject the null hypothesis if the absolute value of Z^*_iexceeds 1.96. When there is only one group of studies, this test is identical to that using the statistic Z^* given in (12.5) with random-effects weights.

Confidence intervals for the group mean effect and the weighted grand mean effect size can be computed by multiplying the respective standard error $\left(\sqrt{v^*_{1\bullet}}\right)$ or $\left(\sqrt{v^*_{\bullet\bullet}}\right)$ by the appropriate two-tailed critical value of the standard normal distribution ($C_{\alpha/2} = 1.96$ for $\alpha = 0.05$ and 95 percent confidence intervals) then adding and subtracting this amount from the weighted mean effect size $\overline{T}^*_{i\bullet}$ or $\overline{T}^*_{\bullet\bullet}$. For example, the $100(1 - \alpha)$ percent confidence interval for $\mu_{i\bullet}$ is given by

$$\overline{T}^*_{i\bullet} - C_{\alpha/2}\sqrt{v^*_{i\bullet}} \le \mu_{i\bullet} \le \overline{T}^*_{i\bullet} + C_{\alpha/2}\sqrt{v^*_{i\bullet}} \qquad (12.48)$$

The accuracy of tests and confidence intervals can be improved somewhat by substitution $v^{KN}_{i\bullet} = v^*_{i\bullet}Q^*_{wi}$ for $v^*_{i\bullet}$ where

$$Q^*_{wi} = \sum_{j=1}^{m_i} w^*_{ij}\left(T_{ij} - \overline{T}^*_{i\bullet}\right)^2 \qquad (12.49)$$

and $C_{\alpha/2}$ is replaced with the corresponding critical value of student's t-distribution with $m_i - 1$ degrees of freedom (see Hartung and Knapp 2001).

To test whether the grand mean effect $\mu_{\bullet\bullet}$ differs from a predefined constant μ_0 (for example to test if $\mu_{\bullet\bullet} - \mu_0 = 0$) by testing the null hypothesis

$$H_0 : \mu_{\bullet\bullet} = \mu_0,$$

use the statistic

$$Z^*_{\bullet\bullet} = \left(\overline{T}^*_{\bullet\bullet} - \mu_0\right) \Big/ \sqrt{v^*_{\bullet\bullet}},$$

and reject H_0 at level α (that is, decide that the effect parameter differs from μ_0) if the absolute value of $Z^*_{\bullet\bullet}$ exceeds the 100α percent critical value of the standard normal distribution. For example, for a two-sided test that $\mu_{\bullet\bullet} = 0$ at $\alpha = 0.05$ level of significance, reject the null hypothesis if the absolute value of $Z^*_{\bullet\bullet}$ exceeds 1.96. When there is only one group of studies, this test is identical to that described using the statistic Z given in equation (12.5) with random-effects weights.

Confidence intervals for the grand mean effect $\mu_{\bullet\bullet}$ can be computed by multiplying the standard error $\left(\sqrt{v^*_{\bullet\bullet}}\right)$ by the appropriate two-tailed critical value of the standard normal distribution ($C_{\alpha/2} = 1.96$ for $\alpha = 0.05$ and 95 percent confidence intervals) then adding and subtracting this amount from the weighted mean effect size $\overline{T}^*_{\bullet\bullet}$. Thus the $100(1 - \alpha)$ percent confidence interval for $\mu_{\bullet\bullet}$ is given by

$$\overline{T}^*_{\bullet\bullet} - C_{\alpha/2}\sqrt{v^*_{\bullet\bullet}} \le \mu_{\bullet\bullet} \le \overline{T}^*_{\bullet\bullet} + C_{\alpha/2}\sqrt{v^*_{\bullet\bullet}}.$$

12.3.2.6 Tests of Heterogeneity in Mixed Models In the mixed model, tests for systematic sources of variation are constructed much like those in the fixed-effects model. The essential difference is that in the mixed model, the total variance (conditional variance plus variance component) plays the role that the conditional variance did in the fixed-effects model. Thus tests for systematic sources of variance are constructed from sums of squared deviations from means just as in conventional analysis of variance.

12.3.2.6.1 An Omnibus Test for Between-Group Differences. To test the hypothesis that there is no variation in group mean effect sizes, that is to test

$$H_0 : \mu_{1\bullet} = \mu_{2\bullet} = \ldots = \mu_{p\bullet},$$

we use the mixed-model between-group heterogeneity statistic Q^*_B defined by

$$Q^*_B = \sum_{i=1}^{p} w^*_{i\bullet}\left(\overline{T}^*_{i\bullet} - \overline{T}^*_{\bullet\bullet}\right)^2, \qquad (12.50)$$

where $w^*_{i\bullet} = 1/v^*_{i\bullet}$ is the reciprocal of the variance of $\overline{T}^*_{i\bullet}$. Note that Q^*_B is just the weighted sum of squares of mixed-model group mean effect sizes about the mixed-model grand mean effect size. When the null hypothesis of no variation across group mean effect sizes is true, Q^*_B has a chi-square distribution with $p - 1$ degrees of freedom. Hence we test H_0 by comparing the obtained value of Q^*_B with the upper tail critical values of the chi-square distribution with $(p - 1)$ degrees of freedom. If Q^*_B exceeds the $100(1 - \alpha)$ percentage point of the chi-square distribution (for example $C_{0.05} = 18.31$ for 10 degrees of freedom and $\alpha = 0.05$), H_0 is rejected at level α and between-group differences are significant.

This test is analogous to the omnibus F-test for variation in group means in one-way analysis of variance. Like the corresponding fixed-effects test of between-class heterogeneity this test differs in that Q^*_B, unlike the F-test, incorporates an estimate of unsystematic error in the form of the weights. Therefore, no separate error term is needed and the sum of squares can be used directly as a test statistic.

An alternative to the test based on Q^*_B uses the statistic

$$Q^{KN}_B = \frac{(k - p)Q^*_B}{(p - 1)Q^*_W}, \qquad (12.51)$$

where $Q^*_W = Q^*_{W1} + \ldots + Q^*_{wp}$ and rejects the null hypothesis if Q^{KN}_B exceeds the $100(1 - \alpha)$ point of the F-distribution with $(p - 1)$ degrees of freedom in the numerator and $(k - \mathrm{p})$ degrees of freedom in the denominator.

12.3.2.7 Computing the Analysis Although Q^*_B can be computed via a computer program for weighted ANOVA using the random-effects weights given in (12.42), the weighted cell means and their standard errors cannot generally be obtained this way, and an analysis will require at least two passes through the data. It is computationally simplest to start by computing the fixed-effects analysis as described earlier. Then use

$$TW_i = \sum\nolimits_{j=1}^{m_i} w_{ij}, \qquad TW_\bullet = \sum\nolimits_{i=1}^{p} TW_i,$$

from the fixed-effects analysis along with

$$TWS_i = \sum_{j=1}^{m_i} w_{ij}^2 \qquad TWS_\bullet = \sum_{i=1}^{p} TWS_i$$

to compute the method of moments estimate of the variance component estimate $\hat{\tau}^2$ given in (12.35). This value can either be used as the final variance component estimate or as the starting point for computing the full or restricted maximum likelihood estimates of τ^2. Whichever final variance component estimate is chosen it is then used to compute the random-effects weights w_{ij}^* given in equation (12.42). The random-effects weights are then used to compute the weighted group means, their variances, and the omnibus test statistic Q_B^*.

Example. Return to the data from studies of gender differences in conformity given in Dataset I given in table 12.3. Using the sums given in table 12.1, we compute the constant given in equation (12.35) for each group as

$$c_1 = 63.895 - (2054.113/63.895) = 31.747,$$

$$c_2 = 45.455 - (2066.116/45.455) = 0.0009,$$

and

$$c_3 = 169.811 - (6221.757/169.811) = 133.172.$$

The method of moments estimate (the distribution-free estimate) of τ^2 is computed from equation (12.35) as

$$\hat{\tau}^2 = (11.085 - 10 + 3)/(31.747 + 0.0009 + 133.172)$$
$$= 0.025.$$

This value was used to compute the random-effects weights (the w_{ij}^*), their product with the corresponding effect-size estimate ($w_{ij}^* T_{ij}$) and their respective sums for each group, which are presented in table 12.1. Using the sums for each group from table 12.1, the random-effects weighted mean effect sizes for the three classes $\overline{T}_{1\bullet}^*$, $\overline{T}_{2\bullet}^*$, and $\overline{T}_{3\bullet}^*$ are given by

$$\overline{T}_{1\bullet}^* = -4.925/35.468 = -0.139,$$

$$\overline{T}_{2\bullet}^* = -6.383/21.277 = -0.300,$$

and

$$\overline{T}_{3\bullet}^* = 34.363/95.120 = 0.361.$$

and the weighted grand mean effect size is

$$\overline{T}_{\bullet\bullet}^* = 23.056/151.865 = 0.152.$$

Note that the random-effects estimates of the class means and the weighted grand mean differ somewhat from the fixed-effects estimates of the class means and the weighted grand mean. The variances $v_{1\bullet}^*$, $v_{2\bullet}^*$, and $v_{3\bullet}^*$ of $\overline{T}_{1\bullet}^*$, $\overline{T}_{2\bullet}^*$, and $\overline{T}_{3\bullet}^*$ are given by

$$v_{1\bullet}^* = 1/35.468 = 0.028,$$

$$v_{2\bullet}^* = 1/21.277 = 0.047,$$

$$v_{3\bullet}^* = 1/95.120 = 0.0105.$$

and the variance $v_{\bullet\bullet}^*$ of $\overline{T}_{\bullet\bullet}^*$ is

$$v_{\bullet\bullet}^* = 1/151.865 = 0.00658.$$

Using formula (12.48) with $C_{.05} = 1.960$, the limits of the 95 percent confidence interval for the group mean parameter $\mu_{1\bullet}$ are given by

$$-0.139 \pm 1.960\sqrt{0.028} = -0.139 \pm 0.329.$$

Thus the 95 percent confidence interval for $\mu_{1\bullet}$ is given by

$$-0.467 \le \mu_{i\bullet} \le 0.189.$$

Because this confidence interval contains zero, or alternately, because the test statistic $Z_1 = |-0.139| / \sqrt{0.028} < 1.96$, we cannot reject the hypothesis that $\mu_{1\bullet} = 0$ at the $\alpha = 0.05$ level of significance. Similarly, 95 percent confidence intervals for the group mean parameters $\mu_{2\bullet}$ and $\mu_{3\bullet}$ are given by

$$-0.725 = -0.300 - 1.960\sqrt{0.047} \le \mu_{2\bullet} \le -0.300$$
$$+ 1.960\sqrt{0.047} = 0.125$$

and

$$0.160 = 0.361 - 1.960\sqrt{0.0105} \le \mu_{3\bullet} \le 0.361$$
$$+ 1.96\sqrt{0.0105} = 0.562.$$

Thus we see that the mean effect sizes for groups 1 and 2 are not significantly different from zero and that the mean effect size for group 3 is significantly greater than zero at the $\alpha = 0.05$ level.

12.3.2.8 Comparisons Among Mean Effects in Mixed Models Comparisons (contrasts) among class mean effect sizes are computed in a manner analogous to their computation in fixed-effects models. The contrast (parameter) is just a linear combination of group means

$$\gamma^* = c_1\mu_{1\bullet} + \ldots + c_p\mu_{p\bullet}, \tag{12.52}$$

where the coefficients $c_1, \ldots, c_p$ (contrast coefficients) are known constants that satisfy the constraint $c_1 + \ldots + c_p = 0$ and are chosen so that the value of the contrast will reflect a particular comparison or pattern of interest.

The sample contrast G is computed as in equation (12.30) except that the mixed-effects estimates of class means $\overline{T}^*_{i\bullet}$ are substituted for the fixed-effects means $\overline{T}_{i\bullet}$ so that the estimate of the contrast is

$$g^* = c_1\overline{T}^*_{1\bullet} + \ldots + c_p\overline{T}^*_{p\bullet}. \tag{12.53}$$

and the variance v_{g^*} of the contrast is computed as in (12.31) except that the mixed-effects variances of the class means $v^*_{i\bullet}$ are substituted for the fixed-effects variances of the class means $v_{i\bullet}$. Thus the estimated contrast g^* has a normal sampling distribution with variance v^*_g given by

$$v^*_g = c_1^2 v^*_{1\bullet} + \ldots + c_p^2 v^*_{p\bullet}. \tag{12.54}$$

Tests of hypotheses about γ, computations of statistical power of tests about γ, and confidence intervals for γ are computed in exactly the same way from formulas (12.53) and (12.54) as in the fixed-effects model, except that the random- or mixed-effects class means $\overline{T}^*_{i\bullet}$ and the variances of these class means $v^*_{i\bullet}$ are substituted for the fixed-effects class means $\overline{T}_{i\bullet}$ and the fixed-effects variances of the class means $v_{i\bullet}$.

Example. Continuing the analysis of standardized mean difference data on gender difference in conformity, recall that there are three groups of effects: those in which 25 percent, 50 percent, and 100 percent respectively, of the authors are male. To contrast the mean of the effects of group 1 (25 percent male authors) with those of group 3 (100 percent male authors), use the contrast coefficients

$$c_1 = -1.0, c_2 = 0.0, c_3 = 1.0.$$

The value of the contrast estimate g^* is

$$g^* = -1.0(-0.139) + 0.0(-0.300) + 1.0(0.361) = 0.500,$$

with an estimated variance of

$$v^*_g = (-1.0)^2(0.028) + (0.0)^2(0.047) + (1.0)^2(0.0105) = 0.0385.$$

The 95 percent confidence interval for $\gamma^* = \mu_{3\bullet} - \mu_{1\bullet}$ is

$$0.115 = 0.500 - 1.960\sqrt{0.0385} \le \gamma^* \le 0.500 + 1.960\sqrt{0.0385} = 0.885$$

Because this confidence interval does not contain zero, or alternatively, because

$$Z^* = 0.500/\sqrt{0.0385} = 2.548$$

exceeds 1.960, we reject the hypothesis that $\gamma^* = 0$ and declare the contrast statistically significant at the $\alpha = 0.05$ level.

12.4 MULTIPLE REGRESSION ANALYSIS FOR EFFECT SIZES

In many cases, it is desirable to represent the characteristics of research studies by continuously coded variables or by a combination of discrete and continuous variables. In such cases, the reviewer often wants to determine the relationship between these continuous variables and effect size. One very flexible analytic procedure for investigating these relationships is an analog to multiple regression analysis for effect sizes (see Hedges 1982b, 1983b; Hedges and Olkin 1985). These methods share the generality and ease of use of conventional multiple regression analysis, and like their conventional counterparts can be viewed as including ANOVA models as a special case. In the recent medical literature, such methods have been called meta-regression (see Borenstein et al. 2009).

12.4.1 Fixed-Effects Analyses

Suppose that we have k independent effect-size estimates $T_1, \ldots, T_k$ with estimated sampling variances $v_1, \ldots, v_k$. The corresponding effect-size parameters are $\theta_1, \ldots, \theta_k$.

We assume that each T_i is normally distributed about θ_i with known variance v_i, that is

$$T_i - \theta_i = \varepsilon_i \sim N(0, v_i), \qquad i = 1, \ldots, k.$$

Suppose also that there are p known predictor variables $X_1, \ldots, X_p$ that are believed to be related to the effects via a linear model of the form

$$\theta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \tag{12.55}$$

where $x_{i1}, \ldots, x_{ip}$ are the values of the predictor variables $X_1, \ldots, X_p$ for the ith study (that is, x_{ij} is the value of X_j for study i), and $\beta_1, \ldots, \beta_p$ are unknown regression coefficients. Thus the linear model for the T_i could be written as

$$T_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, \ldots, k. \tag{12.56}$$

The $k \times p$ matrix

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{kp} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_k' \end{pmatrix},$$

where each row vector $\mathbf{x}_i' = (x_{i1}, \ldots, x_{ip})$ is used to construct what is called in regression analysis the design matrix $\mathbf{X}$, which is assumed to have no linearly dependent columns; that is, $\mathbf{X}$ has rank p. It is often convenient to assume that the elements of the first column vector are $x_{11} = x_{21} = \ldots = x_{k1} = 1$, so that the first regression coefficient becomes an intercept term, as in ordinary regression.

The model can be written succinctly in matrix notation if we denote the k-dimensional vectors of population and sample effect sizes by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ and $\mathbf{T} = (T_1, \ldots, T_k)'$, respectively. Then equation (12.56) can be written in matrix notation as

$$\mathbf{T} = \boldsymbol{\theta} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the p-dimensional column vector of regression coefficients and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_k)'$ is the k-dimensional column vector of study-specific estimation errors.

12.4.1.1 Estimation and Significance Tests for Individual Coefficients Estimation is usually carried out via weighted least squares algorithms. The formulas for estimators and test statistics can be expressed most succinctly in matrix notation and are given, for example, in Hedges and Olkin (1985). Specifically, the vector of regression coefficients corresponding to $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is $\mathbf{b} = (b_1, \ldots, b_p)'$ and the estimate $\mathbf{b}$ is given (in matrix notation) by

$$\mathbf{b} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{T}, \tag{12.57}$$

where $\mathbf{W}$ is a $k \times k$ diagonal matrix whose ith diagonal element is $1/v_i$. The covariance matrix of $\mathbf{b}$ is

$$\boldsymbol{\Sigma} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}. \tag{12.58}$$

The analysis can be conducted using specialized software such as *Comprehensive Meta-analysis* (Borenstein et al. 2013) or standard computer programs (such as in SAS, SPSS, R or Stata) that compute weighted multiple regression analyses. The regression should be run with the effect estimates as the dependent variable and the predictor variables as independent variables with weights defined by the reciprocal of the sampling variances. That is, the weight for T_i is $w_i = 1/v_i$.

Standard computer programs for weighted regression analysis produce the correct (asymptotically efficient) estimates $b_0, b_1, \ldots, b_p$ of the unstandardized regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$. (Note that unlike the SPSS computer program, we use the symbols $\beta_0, \beta_1, \ldots, \beta_p$ to refer to the population values of the unstandardized regression coefficients *not* to the standardized sample regression coefficients). Although these programs give the correct estimates of the regression coefficients, the standard errors and significance values computed by the programs are based on a slightly different model than those used for fixed-effects meta-analysis and are incorrect for the meta-analysis model. Calculating the correct significance tests for individual regression coefficients requires some straightforward hand computations from information given in the computer output.

The correct standard error S_j of the estimated coefficient estimate b_j is simply

$$S_j = SE_j \Big/ \sqrt{MS_{ERROR}} \tag{12.59}$$

where SE_j is the standard error of b_j *as given by the computer program* and MS_{ERROR} is the error or residual mean square (the error variance) from the analysis of variance for the regression as given by the computer program.

Equation (12.59) corrects the standard error by eliminating from the computation the estimate of the constant error variance, which is not needed because the variances of the effect sizes are known and differ from study to study. Alternatively, the correct standard errors of b_0, $b_1, \ldots, b_p$ are the square roots of the diagonal elements of the inverse of the matrix $(\mathbf{X'WX})$, which is sometimes called the inverse of the weighted sum of squares and cross-products matrix. Many regression analysis programs (such as SAS PROC GLM) will print this matrix as an option. The (correct) standard errors obtained from both methods are, of course, identical. They are simply alternative ways to compute the same thing.

The regression coefficient estimates (the b_js) are normally distributed about their respective parameter (the β_j's) values with standard deviations given by the standard errors (the S_js). Hence a $100(1 - \alpha)$ percent confidence interval for each β_j can be obtained by multiplying S_j by the two-tailed critical value C_α of the standard normal distribution (for $\alpha = 0.05$, $C_\alpha = 1.96$) and then adding and subtracting this product from b_j. Thus the $100(1 - \alpha)$ percent confidence interval for β_j is

$$b_j - C_\alpha S_j \le \beta_j \le b_j + C_\alpha S_j \qquad (12.60)$$

A two-sided test of the null hypothesis that the regression coefficient is zero,

$$H_0: \beta_j = 0,$$

at significance level α consists of rejecting H_0 if

$$Z_j = |b_j|/S_j \qquad (12.61)$$

exceeds the 100α percent two-tailed critical value of the standard normal distribution.

Example. Consider the example of the standardized mean differences for gender differences in conformity given in data set I. In this analysis we fit the linear model suggested by Betsy Becker (1986), who explained variation in effect sizes by a predictor variable that was the natural logarithm of the number of items on the conformity measure (column 4 of table 12.3). This predictor variable is highly correlated with the percentage of male authors used as a predictor in the example given for the categorical model analysis. Using SAS PROC GLM with effect sizes and weights given in table 12.3, we computed a weighted regression analysis. The estimates of the regression coefficients were $b_0 = -0.323$ for the intercept and $b_1 = 0.210$ for the effect of the number of items. The standard errors of b_0 and b_1 could be computed in either of two ways. The $(\mathbf{X'WX})$ inverse matrix computed by SAS was

$$\begin{pmatrix} 0.01224 & -0.00407 \\ -0.00407 & 0.00191 \end{pmatrix}$$

and hence the standard errors can be computed as

$$S_0 = \sqrt{0.01224} = 0.1106,$$
$$S_1 = \sqrt{0.00191} = 0.044.$$

Alternatively, we could have obtained the standard errors by correcting the standard errors printed by the program (which are incorrect for our purposes). The standard errors printed by the SAS program were $SE(b_0) = 0.115$ and $SE(b_1) = 0.046$, and the residual mean square from the analysis of variance for the regression was $MS_{ERROR} = 1.083$. Using formula (12.59) gives

$$S_0 = .115/\sqrt{1.083} = 0.1105,$$
$$S_1 = .046/\sqrt{1.083} = 0.044.$$

A 95 percent confidence interval for the effect β_1 of the number of items using $C_{0.05} = 1.960$, $S_1 = 0.044$, and formula (12.60) is given by $0.210 \pm 1.960(0.044)$

$$0.124 \le \beta_1 \le 0.296.$$

Because the confidence interval does not contain zero, or alternatively, because the statistic

$$Z_1 = 0.210/0.044 = 4.773$$

exceeds 1.96, we reject the hypothesis that there is no relationship between number of items and effect size. Thus the number of items on the response measure has a statistically significant relationship to effect size.

12.4.1.2 Omnibus Tests It is sometimes desirable to test hypotheses about groups or blocks of regression coefficients. For example, stepwise regression strategies may involve entering one block of predictor variables (such as a set reflecting methodological characteristics) and then entering another block (such as a set reflecting

treatment characteristics) to see whether the second block explains any of the variation in effect size not accounted for by the first block. Formally, we need a test of the hypothesis that all of the regression coefficients for predictor variables in the second block are zero.

Suppose that the a predictor variables $X_1, \ldots, X_a$ have already been entered and we wish to test whether the regression coefficients for a block of q additional predictor variables $X_{a+1}, X_{a+2}, \ldots, X_{a+q}$ are simultaneously zero. That is, we wish to test

$$H_0 : \beta_{a+1} = \ldots = \beta_{a+q} = 0.$$

The test statistic is the weighted sum of squares for the addition of this block of variables. It can be obtained directly as the difference in the weighted error sum of squares for the model with a predictors and the weighted error sum of squares of the model with $(a + q)$ predictors. Alternatively, it can be computed from the output of the weighted stepwise regression as

$$Q_{CHANGE} = qF_{CHANGE}MS_{ERROR} \qquad (12.62)$$

where F_{CHANGE} is the value of the F-test statistic for testing the significance of the addition of the block of b predictor variables and MS_{ERROR} is the weighted error or residual mean square (error variance) from the analysis of variance for the regression. The test at significance level α consists of rejecting H_0 if Q_{CHANGE} exceeds the $100(1 - \alpha)$ percent point of the chi-square distribution with q degrees of freedom.

If the number k of effects exceeds p, the number of predictors including the intercept, then a test of goodness of fit or model specification is possible. The test is formally a test of the null hypothesis that the population effect sizes $\theta_1, \ldots, \theta_k$ are exactly determined by the linear model

$$\theta_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, i = 1, \ldots, p,$$

versus the alternative that some of the variation in the θ_is is not fully explained by $X_1, \ldots, X_p$. The test statistic is the weighted residual sum of squares Q_E about the regression line,

$$Q_E = \mathbf{T}'\left(\mathbf{W} - \mathbf{WX}(\mathbf{X}'\mathbf{WX})^{-1}\mathbf{X}'\mathbf{W}\right)\mathbf{T}. \qquad (12.63)$$

The test can be viewed as a test for greater than expected residual variation. This statistic is given in the analysis of variance for the regression and is usually called the error or residual sum of squares on computer printouts. The test at significance level α consists of rejecting the null hypothesis of model fit if Q_E exceeds the $100(1 - \alpha)$ percent point of the chi-square distribution with $(k - p)$ degrees of freedom.

The tests of homogeneity of effect given in chapter 13 and tests of homogeneity of effects within groups of independent effects described in connection with the analysis of variance for effect sizes are special cases of the test of model fit given here. That is, the statistic Q_E reduces to the statistic Q given in formula (12.6) when there are no predictor variables, and Q_E reduces to the statistic Q_W given in formula (12.24) when the predictor variables are dummy coded to represent group membership.

Example. Continue the example of the regression analysis of the standardized mean differences for gender differences in conformity, using SAS PROC GLM to compute a weighted regression of effect size on the logarithm of the number of items on the conformity measure. Although we can illustrate the test for the significance of blocks of predictors, there is only one predictor. We start with $a = 0$ predictors and add $q = 1$ predictor variables. The weighted sum of squares for the regression in the analysis of variance for the regression gives $Q_{CHANGE} = 23.112$. We could also have computed Q_{CHANGE} from the F-test statistic F_{CHANGE} for the R-squared change and the MS_{ERROR} for the analysis of variance for the regression. Here $F_{CHANGE} = 21.341$ and $MS_{ERROR} = 1.083$, so using formula (12.62)

$$Q_{CHANGE} = 1(21.341)(1.083) = 23.112,$$

identical to the result obtained directly. Comparing 23.112 with 3.841, the 95 percent point of the chi-square distribution with 1 degree of freedom, we reject the hypothesis that the (single) predictor is unrelated to effect size. This is, of course, the same result obtained by a test for the significance of the regression coefficient.

We also test the goodness of fit of the regression model. The weighted residual sum of squares was computed by SAS PROC GLM as $Q_E = 8.664$. Comparing this value to 15.507, the 95 percent critical value of the chi-square distribution with $10 - 2 = 8$ degrees of freedom, we see that we cannot reject the fit of the linear model. In fact, chi-square values as large as 8.664 would occur between 25 and 50 percent of the time *due to chance* if the model fit exactly.

12.4.2 Random-Effects Analyses

In many cases, systematic variation among effect-size parameters is non-negligible even after controlling for the factors of interest in the analysis. That is, residual variation is greater than would be expected from sampling error alone. If the researcher believes that this variation should be included in the estimation of the uncertainty of the estimates of the regression coefficients and the estimates of the regression coefficients themselves, fixed-effects models are not appropriate because such excess residual variation has no effect on the computation of the estimates or their uncertainty of the estimates in fixed-effects models. The random- or mixed-effects model is a generalization of the fixed-effects model that incorporates a component of between-study variation into the uncertainty of effect-size parameters and their estimates that increases residual variation.

Random- or mixed-effects models are appropriate under the same kinds of circumstances discussed earlier in connection with mixed-effects categorical models. In fact, the categorical random-effects models and the analyses discussed are special cases of the random-effects models discussed here.

12.4.2.1 Model and Notation Suppose that we have k independent effect-size estimates $T_1, \ldots, T_k$ with (estimated) sampling variances $v_1, \ldots, v_k$. The corresponding effect-size parameters are $\theta_1, \ldots, \theta_k$. As in the fixed-effects model, we assume that each T_i is normally distributed about θ_i with known variance v_i, that is

$$T_i - \theta_i = \varepsilon_i \sim N(0, v_i), \qquad i = 1, \ldots, k.$$

Suppose also that there are p known predictor variables $X_1, \ldots, X_p$ that are believed to be related to the effects via a linear model of the form

$$\theta_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \xi_i,$$

where $x_{i1}, \ldots, x_{ip}$ are the values of the predictor variables $X_1, \ldots, X_p$ for the *ith* study (that is x_{ij} is the value of X_j for study i), and $\beta_1, \ldots, \beta_p$ are unknown regression coefficients, and ξ_i is a random effect with variance τ^2. Thus the linear model for the T_i could be written as

$$T_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \xi_i + \varepsilon_i, \qquad i = 1, \ldots, k. \quad (12.64)$$

As in the fixed-effects model, the $k \times p$ matrix

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{kp} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_k' \end{pmatrix},$$

where each row vector $\mathbf{x}_i' = (x_{i1}, \ldots, x_{ip})$ is used to construct what is called in regression analysis the design matrix $\mathbf{X}$, which is assumed to have no linearly dependent columns; that is, $\mathbf{X}$ has rank p. It is often convenient to assume that the elements of the first column vector are $x_{11} = x_{21} = \ldots = x_{k1} = 1$, so that the first regression coefficient becomes an intercept term, as in ordinary least squares regression.

We denote the k-dimensional vectors of population and sample effect sizes by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ and $\mathbf{T} = (T_1, \ldots, T_k)'$, respectively. Equation (12.64) can be written succinctly in matrix notation as

$$\mathbf{T} = \boldsymbol{\theta} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi} + \boldsymbol{\varepsilon}, \quad (12.65)$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the p-dimensional vector of regression coefficients, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_k)'$ is the k-dimensional vector of study-specific random effects, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_k)' = \mathbf{T} - \boldsymbol{\theta}$ is a k-dimensional vector of residuals of $\mathbf{T}$ about $\boldsymbol{\theta}$.

12.4.2.1.1 Terminology of Mixed- or Random-Effects Models. The models described in this section are usually called random-effects models in meta-analysis because the effect-size parameters are considered random. Similar models in other applications are often called mixed-effects models or just mixed models because the regression coefficients of the model (but not the residual $\boldsymbol{\xi}$) are themselves considered fixed, but unknown constants, so that the model includes both fixed effects (the regression coefficients) and random effects (the residual $\boldsymbol{\xi}$).

12.4.2.1.2 Homogeneity of Variance of Random Effects. In the model, the random effects ξ_i are taken to have the same variance. This is not necessary in principle, but it makes both computations and conceptual models much simpler. More general formulations including models with heterogeneous variance components are possible but are not discussed here.

12.4.2.2 Relation to Classical Hierarchical Linear Models Progress has been considerable in developing

software to estimate and to test the statistical significance of parameters in mixed general linear models. To illustrate the connection, it is most convenient to use the representation of the general mixed model as a hierarchical linear model, one that is widely used in the social sciences (for example, Bryk and Raudenbush 1992; Goldstein 1987; Konstantopoulos 2011; Longford 1987). In this representation, the data is regarded as hierarchically structured and the structural model is defined for each level of the hierarchy. In meta-analysis, level I is that of the study and the model for level I is

$$T_i = \theta_i + \varepsilon_i, i = 1, \ldots, k \qquad (12.66)$$

where ε_i is a sampling error of T_i as an estimate of θ_i. Level II of the model describes between-study variation in the study-specific effect-size parameters (the θ_i). In this chapter the linear model in equation (12.66) would imply a level II model like

$$\theta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \xi_i, \qquad (12.67)$$

where ξ_i is a study-specific random effect. Most of the attention in estimating these models has focused on the case where both the ε_i and the ξ_i are independently and normally distributed with zero mean and unknown variances, that is

$$\varepsilon_i \sim N(0, v), i = 1, \ldots, k$$

and

$$\xi_i \sim N(0, \tau^2), i = 1, \ldots, k.$$

There are two important differences in the hierarchical linear models (or general mixed models) usually studied and the model in equation (12.65) used in meta-analysis. The first is that in meta-analysis models, such as in (12.65), the variances of the sampling errors $\varepsilon_1, \ldots, \varepsilon_k$ are *not* identical across studies. The sampling error variances usually depend on various aspects of study design (particularly sample size) that cannot be expected to be constant across studies. The second is that the sampling error variances in meta-analysis are generally assumed to be known. Therefore the model in (12.65) used in meta-analysis can be considered a special case of the general hierarchical linear model where the level I variances are unequal, but known. Consequently software for the analysis of hierarchical linear models can be used for mixed model meta-analysis if it permits (as do the programs HLM and SAS PROC MIXED) the specification of first level variances that are unequal but known.

12.4.2.3 Estimation of the Residual Variance Component τ^2 The first step in the mixed model analysis is the estimation of the residual variance component τ^2. As in the case of analysis of variance style models for effect sizes, the variance component can be estimated using methods that do not depend on assumptions about the distribution of the random effects (method of moments estimators) or using methods that assume that the random effects are normally distributed (full or restricted maximum likelihood estimation). The methods discussed earlier in section 12.3.2.3.1 are distribution-free in the sense that the derivations of the estimates and their standard errors do not depend on the sampling distribution of the random effects (the ξ_is). However, the sampling distributions of the test statistics and probability statements (such as about confidence intervals) do depend on the distribution of the random effects. Some evidence from simulation studies of estimates based on linear models suggests effect sizes with miss-specified distributions for the random effects (Hedges and Vevea 1998). These in turn suggest that, as long as the variance τ^2 of the random effects is not large relative to the typical conditional variance of T_i given θ_i, the confidence intervals for effects are not substantially affected by even if the distribution of the random effects deviates substantially from normality.

12.4.2.3.1 Distribution-Free Analyses. The distribution-free method of estimation involves computing an estimate of the residual variance component by the method of moments and then computing a weighted least squares analysis conditional on this estimate. Whereas the estimates and their standard errors are distribution free in the sense that they do not depend on the form of the distribution of the random effects, the tests and confidence statements associated with these methods are only strictly true if the random effects are normally distributed.

Two alternative methods of moments estimators are most frequently used. The most frequently used estimator is based on the statistic used to test the significance of the residual variance component (the inverse conditional-variance-weighted residual sum of squares). It is the natural generalization of the homogeneity test described in equation (12.6). An alternative estimator is based on the residual sum of squares from the unweighted regression.

The usual estimator of the residual variance component is given by

$$\hat{\tau}^2 = (Q_E - k + p)/c \tag{12.68}$$

where Q_E is the residual sum of squares from the fixed-effects weighted regression given in (12.63) and c is a constant given by

$$c = \sum_{i=1}^{k} w_i - \text{tr}\left[\left(\sum_{i=1}^{k} w_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^{k} w_i^2 \mathbf{x}_i \mathbf{x}_i'\right)\right], \tag{12.69}$$

where the $w_i = 1/v_i$ are the fixed-effects weights and tr($\mathbf{A}$) is the trace of the square matrix $\mathbf{A}$.

When the random effects are normally distributed, the variance of $\hat{\tau}^2$ is given by

$$\left[SE(\hat{\tau}^2)\right]^2 = \frac{\sum_{i=1}^{k} w_i^2\left(v_i^*\right)^2 - 2\text{tr}\left[\left(\sum_{i=1}^{k} w_i^3\left(v_i^*\right)^2 \mathbf{x}_i \mathbf{x}_i'\right)\right] + \text{tr}\left[\mathbf{B}^2\right]}{c^2} \tag{12.70}$$

where the $p \times p$ matrix $\mathbf{B}$ is given by

$$\mathbf{B} = \left(\sum_{i=1}^{k} w_i^2 v_i^* \mathbf{x}_i \mathbf{x}_i'\right)\left(\sum_{i=1}^{k} w_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1}.$$

The distribution of the residual variance component estimate is not close to normal unless $k - p$ is large. Consequently, probability statements based on SE($\hat{\tau}^2$) and the assumption of normality should be viewed as rough unless $k - p$ is large.

12.4.2.3.2 Estimators of τ^2 Assuming Normally Distributed Random Effects. If the random effects are normally distributed (or if we just assume that this is approximately true), the method of maximum likelihood can be used to obtain estimates of the variance component and in turn, estimates of the regression coefficients and their standard errors. The likelihood of the mixed model in (12.65) can be separated into parts involving the fixed effects (β) and the variance components (τ^2) so that computing the maximum likelihood estimate of β depends on first computing the maximum likelihood estimate of τ^2. There are two different estimation strategies. One is the unrestricted maximum likelihood, sometimes called full maximum likelihood and the other is the restricted maximum likelihood. Full maximum likelihood estimates are obtained by maximizing the likelihood of the observations, which involves both β and τ^2. Restricted estimates are obtained by maximizing the likelihood regarded as a function of τ^2 alone. Restricted estimates of τ^2 have the potential advantage that they take into account the uncertainty in estimating β, while the full maximum likelihood estimates do not (see, for example, Raudenbush and Bryk 2002). When the restricted maximum likelihood estimates of variance components can be computed analytically, they are often unbiased (whereas the full maximum likelihood estimates are biased) and the restricted maximum likelihood estimates often appear to be less biased in other situations.

Full maximum likelihood estimation of τ^2. One method for estimating τ^2 relies on the fact that if either β or τ^2 were known, it would be easy to obtain the least squares estimate of the other. By starting with an initial value of τ^2 and estimating first β and then reestimating τ^2, a sequence of estimates can be obtained that converge to the maximum likelihood estimators of β and τ^2 (see Goldstein 1986).

Specifically, start with an initial estimate $\tau^2_{[0]}$ of τ^2, such as the method of moments estimator of the between-study variance given in equation (12.68). Then, if $\hat{\tau}^2_{[s]}$ is the estimate of τ^2 on the sth step, the $k \times k$ weight matrix on that step becomes

$$\mathbf{W}_{[s]} = \text{Diag}\left[w^*_{1[s]}, \ldots, w^*_{k[s]}\right], \tag{12.71}$$

the estimate $\mathbf{b}^*_{[s+1]}$ of β on the $(s+1)$st step is given by

$$\begin{aligned}\mathbf{b}^*_{[s+1]} &= \left(\mathbf{X}'\mathbf{W}_{[s]}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}_{[s]}\mathbf{T} \\ &= \left(\sum_{i=1}^{k} w^*_{i[s]}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^{k} w^*_{i[s]}\mathbf{x}_i T_i\right)\end{aligned} \tag{12.72}$$

where

$$w^*_{i[s]} = 1/\left(v_i + \hat{\tau}^2_{[s]}\right). \tag{12.73}$$

The vector $\mathbf{e}_{[s+1]} = (e_{1[s+1]}, \ldots, e_{k[s+1]})'$ of residuals on the $(s+1)$st step given by

$$\mathbf{e}_{[s+1]} = \mathbf{T} - \mathbf{X}\mathbf{b}^*_{[s+1]}, \tag{12.74}$$

is used to calculate $\hat{\tau}^2_{[s+1]}$, the generalized least squares estimate of τ^2 given $\mathbf{b}^*_{[s+1]}$ and the data, via

$$\hat{\tau}^2_{[s+1]} = \left[\sum_{i=1}^{k}\left(w^*_{i[s]}\right)^2\left(e^2_{i[s+1]} - v_i\right)\right] \Big/ \sum_{i=1}^{k}\left(w^*_{i[s]}\right)^2 \qquad (12.75)$$

setting any negative estimates of τ^2 to zero. Proceed until the estimates of τ^2 converge (that is, changes negligibly between two consecutive iterations).

Restricted maximum likelihood estimation of τ^2. One method for obtaining restricted maximum likelihood estimates τ^2 is quite similar to that described above for obtaining full maximum likelihood estimates of τ^2. To use this method start with an initial value of τ^2, such as the method of moments estimator of the variance given in (12.78). Then estimate $\boldsymbol{\beta}$ and reestimate τ^2 (but by a slightly different method than that used for full maximum likelihood). This leads to a sequence of estimates that converge to the restricted maximum likelihood estimator of τ^2 and consequently $\boldsymbol{\beta}$ (see Goldstein 1989). The weights at the sth step $w^*_{i[s]}$, the estimate of $\boldsymbol{\beta}$ at the $(s+1)$st step $\hat{\boldsymbol{\beta}}^*_{[s+1]}$, and the residual at the $(s+1)$st step $\mathbf{e}_{[s+1]}$ are computed exactly as in formulas (12.72), (12.73), and (12.74) for full maximum likelihood estimation. The only difference is that instead of computing the estimate of τ^2 at the $(s+1)$st step $\hat{\tau}^2_{[s+1]}$ via (12.75), compute $\hat{\tau}^2_{[s+1]}$ via

$$\hat{\tau}^2_{[s+1]} = \frac{\displaystyle\sum_{i=1}^{k}\left(w^*_{i[s]}\right)^2\left(e^2_{i[s+1]} - v_i\right) + \mathrm{tr}\left[\left(\sum_{i=1}^{k} w^*_{i[s+1]}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^{k}\left(w^*_{i[s+1]}\right)^2\mathbf{x}_i\mathbf{x}_i'\right)\right]}{\displaystyle\sum_{i=1}^{k}\left(w^*_{i[s+1]}\right)^2} \qquad (12.76)$$

The difference between equations (12.75) and (12.76) is the second term of the numerator, which adds a small constant to the estimate of τ^2 at each iteration. This iterative method relies on the fact that e_i^2 estimates $v_i + \tau^2$, but actually the expected value of $e_i^2 < v_i + \tau^2$. The constant corrects that bias and therefore reduces the overall bias of the restricted maximum likelihood estimate. As in the case of full maximum likelihood, the iteration proceeds until the estimates of τ^2 converge (that is, change negligibly between two consecutive iterations) and τ^2 is approximated by the value of $\hat{\tau}^2_{[s]}$ at the last iteration.

12.4.2.3.3 Testing the Significance of the Residual Variance Component. It is sometimes useful to test the statistical significance of the residual variance component τ^2 in addition to estimating it. The sampling distributions of the estimators of τ^2 given in the previous section are not close to normal unless k is large and thus a normal-score test assuming a normal distribution of $\hat{\tau}^2/\mathrm{SE}(\hat{\tau}^2)$ does not provide an adequate test. However, it can be shown that if the null hypothesis H_0: $\tau^2 = 0$ is true, then the weighted residual sum of squares Q_E given in equation (12.63) in conjunction with the fixed-effects analysis has a chi-square distribution with $k - p$ degrees of freedom (where p is the total number of predictors including the intercept). Therefore, the test of H_0 at level α is to reject if Q_E exceeds the $100(1-\alpha)$ percent point of the chi-square distribution with $(k-p)$ degrees of freedom.

12.4.2.4 Estimation of the Regression Coefficients

The mixed-effects linear model $\mathbf{T} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi} + \boldsymbol{\varepsilon}$ for the effect sizes can be written as

$$\mathbf{T} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta}, \qquad (12.77)$$

where $\boldsymbol{\eta} = \boldsymbol{\xi} + \boldsymbol{\varepsilon}$, which is analogous to the model that is the basis for ordinary least squares regression analysis. The distribution of $\boldsymbol{\eta} = \boldsymbol{\xi} + \boldsymbol{\varepsilon}$ has mean zero and diagonal covariance matrix given by

$$\mathrm{Diag}\left(v_1 + \tau^2, v_2 + \tau^2, \ldots, v_k + \tau^2\right). \qquad (12.78)$$

The elements of $\boldsymbol{\eta}$ are independent but not identically distributed. If the residual variance component τ^2 were known, the mixed model would become a special case of the fixed-effects model with the i^{th} observation having residual variance equal to $v_i + \tau^2$. When τ^2 has to be estimated (method of moments, full maximum likelihood, or restricted maximum likelihood) the estimate is substituted for τ^2 in (12.78). We then use the estimate of τ^2 to obtain a generalized least squares estimate of $\boldsymbol{\beta}$. Let the $k \times k$ diagonal matrix $\mathbf{W}$ be defined by

$$\mathbf{W}^* = \mathrm{Diag}\left[1/\left(v_1 + \hat{\tau}^2\right), \ldots, 1/\left(v_k + \hat{\tau}^2\right)\right]. \qquad (12.79)$$

The weighted least squares estimator $\mathbf{b}^*$ under the model (12.65) using the estimated weight matrix $\mathbf{W}^*$ is given by

$$\mathbf{b}^* = (\mathbf{X}'\mathbf{W}^*\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^*\mathbf{T} = \left(\sum_{i-1}^{k} w_i^*\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i-1}^{k} w_i^*\mathbf{x}_i T_i\right), \qquad (12.80)$$

where $w_i^* = 1/(v_i + \hat{\tau}^2)$. When k is large, $\mathbf{b}^*$ is approximately normally distributed with mean $\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Sigma}^*$ given by

$$\boldsymbol{\Sigma}^* = (\mathbf{X}'\mathbf{W}^*\mathbf{X})^{-1} = \left(\sum_{i-1}^{k} w_i^* \mathbf{x}_i \mathbf{x}_i'\right)^{-1}. \qquad (12.81)$$

Regardless of the method used to estimate τ^2, the estimate of $\boldsymbol{\beta}$ has the same approximate sampling distribution with mean $\boldsymbol{\beta}$ and covariance matrix given in formula (12.81).

12.4.2.4.1 Tests and Confidence Intervals for Individual Regression Coefficients. The standard normal distribution can be used to obtain tests of significance or confidence intervals for components of $\boldsymbol{\beta}$. If σ_{jj}^* is the *jth* diagonal element of $\boldsymbol{\Sigma}^*$, and $\mathbf{b}^* = (b_1^*, \ldots, b_p^*)'$ then an approximate $100(1 - \alpha)$ percent confidence interval for β_j, $1 \le j \le p$, is given by

$$b_j^* - C_{\alpha/2}\sqrt{\sigma_{jj}^*} \le \beta_j \le b_j^* + C_{\alpha/2}\sqrt{\sigma_{jj}^*} \qquad (12.82)$$

where $C_{\alpha/2}$ is the 100α percent two-tailed critical value of the standard normal distribution.

An approximate two-tailed test of the hypothesis that β_j equals some predefined value β_{j0} (typically 0), that is a test of the hypothesis

$$H_0: \beta_j = \beta_{j0},$$

uses the statistic

$$Z_j^* = \left(b_j^* - \beta_{j0}\right)\Big/\sqrt{\sigma_{jj}^*}, \qquad (12.83)$$

and rejects H_0 when the absolute value of Z_j^* exceeds the 100α percent critical value of the standard normal distribution. The usual theory for the normal distribution can be applied if one-tailed or simultaneous confidence intervals are desired.

Tests and confidence intervals can be made somewhat more accurate by using the variance $\sigma_{jj}^{KH} = \sigma_{jj}^* Q_E^*$ in place of σ_{jj}^*, where

$$Q_E^* = \mathbf{T}'\left(\mathbf{W}^* - \mathbf{W}^*\mathbf{X}(\mathbf{X}'\mathbf{W}^*\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^*\right)\mathbf{T}, \qquad (12.84)$$

whenever $Q_E^* > k - p$ and using critical values from student's t-distribution with $k - p$ degrees of freedom (see Knapp and Hartung 2003).

The estimates and standard errors can be computed using a standard weighted regression program such as SAS PROC GLM. The estimates of the regression coefficients and the weighted sums of squares will be correct for meta-analysis, but the standard errors of the regression coefficients given by the program will have to be corrected via equation (12.59). Alternatively, software such as *HLM* (v-known option) and specialized software such as *Comprehensive Meta-analysis* or the R package *Metaphor* (https://github.com/wviechtb/metafor) can do all of the required computations directly.

12.4.2.4.2 Tests for Blocks of Regression Coefficients. As in the fixed-effects model, we sometimes want to test whether a subset $\beta_1, \ldots, \beta_m$ of the regression coefficients are simultaneously zero, that is,

$$H_0: \beta_1 = \ldots = \beta_m = 0.$$

This test arises, for example, in stepwise analyses where it is desired to determine whether a set of m of the p predictor variables ($m \le p$) are related for effect size after controlling for the effects of the remaining predictor variables. To test this hypothesis, compute $\mathbf{b} = (b_1^*, \ldots, b_m^*)'$ and the statistic

$$Q^* = \left(b_1^*, \ldots, b_m^*\right)\left(\boldsymbol{\Sigma}_{11}^*\right)^{-1}\left(b_1^*, \ldots, b_m^*\right)', \qquad (12.85)$$

where Σ_{11}^* is the upper $m \times m$ submatrix of

$$\boldsymbol{\Sigma}^* = \begin{pmatrix} \boldsymbol{\Sigma}_{11}^* & \boldsymbol{\Sigma}_{12}^* \\ \boldsymbol{\Sigma}_{12}^* & \boldsymbol{\Sigma}_{22}^* \end{pmatrix}.$$

The test that $\beta_1 = \ldots = \beta_m = 0$ at the 100α percent significance level consists in rejecting the null hypothesis if Q exceeds the $100(1 - \alpha)$ percentage point of the chi-square distribution with m degrees of freedom. A somewhat more accurate test can be obtained by using the test statistic

$$\left(b_1^*, \ldots, b_m^*\right)\left(\boldsymbol{\Sigma}_{11}^* Q_E^*\right)^{-1}\left(b_1^*, \ldots, b_m^*\right)', \qquad (12.86)$$

where Q_E^* is given by (12.84).

If $m = p$, then the test given above is a test that all the β_j are simultaneously zero, that is $\boldsymbol{\beta} = 0$. In this case, the test statistic Q^* given in (12.85) becomes the weighted sum of squares due to regression

$$Q_R^* = \mathbf{b}^{*\prime}\,\boldsymbol{\Sigma}^{*-1}\mathbf{b}^*. \qquad (12.87)$$

The test that $\boldsymbol{\beta} = 0$ is simply a test of whether the weighted sum of squares due to the regression is larger than would be expected if $\boldsymbol{\beta} = 0$, and the test consists of rejecting the hypothesis that $\boldsymbol{\beta} = 0$ if Q_R exceeds the $100(1 - \alpha)$ percentage point of a chi-square with p degrees of freedom.

Example. Return to the example of the standardized mean differences for gender differences in conformity given in data set I. In this analysis we fit the linear model suggested by Becker (1986), who explains variation in effect sizes by a predictor variable that was the natural logarithm of the number of items on the conformity measure (column 4 of table 12.3). This variable is highly correlated ($r = 0.793$) with the percentage of male authors used as a predictor in the example given for the categorical model analysis. Starting with the fixed-effects analysis, we computed the method of moments estimate of the variance component. The weighted residual sum of squares given in (12.63) is $Q_E = 8.664$, and the constant c given in (12.69) was computed as $c = 203.493$, which leads to the method of moments estimator

$$\hat{\tau}^2 = \frac{8.664 - 8}{203.293} = 0.003.$$

Using this starting value for τ^2 with either the full or restricted maximum likelihood estimators yields an estimate of zero for τ^2 after the second iteration. Using the estimate $\hat{\tau}^2 = 0.003$ to compute random-effects weights and then using these weights into a weighted regression such as SAS PROC GLM with effect sizes and weights given in table 12.3, we computed a weighted regression analysis. The estimates of the regression coefficients were $b_0^* = -0.321$ for the intercept and $b_1^* = 0.212$ for the effect of the number of items. The standard errors of b_0^* and b_1^* could be computed in either of two ways. The $(\mathbf{X}'\mathbf{W}^*\mathbf{X})$ inverse matrix computed by SAS was

$$\begin{pmatrix} 0.0137 & -0.00456 \\ -0.00456 & 0.0021 \end{pmatrix}$$

and hence the standard errors can be computed as

$$S_0 = \sqrt{0.0137} = 0.117,$$
$$S_1 = \sqrt{0.0021} = 0.046.$$

Alternatively, we could have obtained the standard errors by correcting the standard errors printed by the program (which are incorrect for our purposes). The standard errors printed by the SAS program were $SE(b_0) = 0.117$ and $SE(b_1) = 0.046$, and the residual mean square from the analysis of variance for the regression was $MS_{ERROR} = 0.999$. Using formula (12.59) gives

$$S_0 = .117/\sqrt{0.999} = 0.117,$$
$$S_1 = .046/\sqrt{0.999} = 0.046.$$

A 95 percent confidence interval for the effect β_1 of the number of items using $C_{0.05} = 1.960$, $S_1 = 0.046$, and formula (12.82) is given by $0.212 \pm 1.960(0.046)$

$$0.122 \le \beta_1 \le 0.302.$$

Because the confidence interval does not contain zero, or, alternatively, because the statistic

$$Z_1^* = 0.212/0.046 = 4.609$$

exceeds 1.960, we reject the hypothesis that there is no relationship between number of items and effect size. Thus the number of items on the response measure has a statistically significant relationship to effect size.

After two iterations, both the maximum likelihood estimate and the restricted maximum likelihood estimates of τ^2 are zero, so both the maximum likelihood estimate and the restricted maximum likelihood analyses are identical to the fixed-effects analysis.

12.4.2.5 Robust Variance Estimation Another approach to inference about effect sizes is based on variance estimates computed from the empirical distribution of the effect-size estimates. This approach, similar to that used in econometrics (Wooldridge 2010) to obtain standard errors of regression coefficients, was adapted to meta-analysis by Larry Hedges, Elizabeth Tipton, and Matt Johnson (2010). It is particularly appealing for three reasons. First, it makes no assumptions about the (conditional or unconditional) distribution of the effect-size estimates, so it is robust to violations of assumptions that estimates or random effects have any specific (for example, normal) distributions. Second, it does not require that the study-level covariates be fixed as in other meta-regression models. Positing fixed covariates makes sense when the values of covariates may be set by the experimenter, but when the values of the covariates are sampled

along with those of the effect-size estimates, the assumption seems problematic. One may argue that the analysis is conditional upon the particular values of the covariates sampled. This may be intellectually defensible, but it seems particularly problematic to say on the one hand that studies are a random sample from a population while conditioning on the observed covariate values. A theorem that justifies robust variance computation makes the intellectual subterfuge of conditioning unnecessary. Third, robust variance computations can be used even when variance estimates for individual effects size estimates are not available (as occurs when reporting of statistics in studies is incomplete).

There are two disadvantages of this method. One is that it provides no method for computing weights to increase efficiency. However, if the variance of each effect-size estimate is known, then standard random-effects procedures can be used to compute (efficient) weights, and the robust variance estimates can be used in conjunction with these weights. A second disadvantage is that, unlike other approaches to inference in meta-analysis, the theory justifying the robust variance estimates assumes a large number of studies.

Suppose that $\hat{\boldsymbol{\beta}}^*$ is a mixed model estimate of $\boldsymbol{\beta}$ computed using weights w_i^*, $i = 1, \ldots, k$. Then the robust variance estimate $\mathbf{V}^R$ is given by

$$\mathbf{V}^R = \left(\frac{k}{k-p}\right)\left(\sum_{i-1}^{k} w_i^* \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \left(\sum_{i=1}^{k} \left(w_i^*\right)^2 e_i^2 \mathbf{x}_i \mathbf{x}_i'\right)\left(\sum_{i-1}^{k} w_i^* \mathbf{x}_i \mathbf{x}_i'\right)^{-1}, \quad (12.88)$$

where $e_i = T_i - \mathrm{x}_i \hat{\boldsymbol{\beta}}^*$ is the residual in the ith study and p is the total number of coefficients in the regression model including the intercept. Any of the methods provided in this chapter for constructing tests or confidence intervals, such as equation (12.82), from the distribution of $\hat{\boldsymbol{\beta}}^*$ can be used with the robust variance estimates. However, simulation studies suggest that tests and confidence intervals computed using critical values of student's t-distribution with $(k - p)$ degrees of freedom will yield tests with actual significance levels that are closer to nominal than standard normal critical values (Hedges, Tipton, and Johnson 2010).

Although the noted simulations suggest that the method can perform reasonably well for a single covariate even when the number of studies is as small as ten to twenty, it is difficult to know when the number of studies is large enough to support valid inferences when the number of covariates is larger. Tipton suggests an improvement to the original method that improve performance in small samples of studies (2015). The modified robust variance estimate is

$$\mathbf{V}^R = \left(\frac{k}{k-1}\right)\left(\sum_{i-1}^{k} w_i^* \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \left(\sum_{i=1}^{k} \left(w_i^*\right)^2 a_i e_i^2 \mathbf{x}_i \mathbf{x}_i'\right)\left(\sum_{i-1}^{k} w_i^* \mathbf{x}_i \mathbf{x}_i'\right)^{-1}, \quad (12.89)$$

where a_i is an adjustment given by

$$a_i = \left(1 - w_i^* \mathbf{x}_i \left(\sum_{i-1}^{k} w_i^* \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \mathbf{x}_i'\right)^{-1}.$$

Thus the robust analysis involves first computing a random-effects regression analysis using the method of moments or a distribution-free estimate of τ^2, then computing the variance of the regression coefficient estimates using formulas (12.86) or (12.87). Then approximate $100(1 - \alpha)$ percent confidence interval for β_j, $1 \le j \le p$, is given by

$$b_j^* - C_{\alpha/2}\sqrt{v_{jj}^{*R}} \le \beta_j \le b_j^* + C_{\alpha/2}\sqrt{\sigma_{jj}^{*\mathrm{R}}} \quad (12.90)$$

where $C_{\alpha/2}$ is the 100α percent two-tailed critical value of the t-distribution with ν_j degrees of freedom, where ν_j is given by (12.88), and $v_{jj}^{*\mathrm{R}}$ is the jth diagonal element of $\mathbf{V}^{\mathrm{R}}$ given by equation (12.88) or (12.89).

An approximate two-tailed test of the hypothesis that β_j equals some predefined value β_{j0} (typically 0), that is a test of the hypothesis

$$H_0: \beta_j = \beta_{j0},$$

uses the statistic

$$Z_j^* = \left(b_j^* - \beta_{j\mathrm{o}}\right) \Big/ \sqrt{\sigma_{jj}^{*R}}, \quad (12.91)$$

and rejects H_0 when the absolute value of Z_j^* exceeds the 100α percent critical value of the standard normal distribution. The usual theory for the normal distribution can be applied if one-tailed or simultaneous confidence intervals are desired.

Example. Return to the example of the standardized mean differences for gender differences in conformity given in data set I and the analysis explaining variation in effect sizes by a predictor variable that was the natural logarithm of the number of items on the conformity measure (column 4 of table 12.3). Here we use random-effects weights based on the method of moments estimate of τ^2 computed in the previous example, so that the weights become

$$w_i^* = 1/(v_i + 0.003).$$

Using these weights, the robust covariance matrix of $b^* = (\hat{\beta}_1^*, \hat{\beta}_2^*)$ given by (12.88) is

$$\begin{pmatrix} 0.01541 & -0.00461 \\ -0.00461 & 0.00177 \end{pmatrix}.$$

The standard errors of b_1^* and b_2^* (square roots of the diagonal elements) corresponding to the variance estimate based on (12.88) are 0.124 and 0.042.

Using the robust variance estimate and critical value $C_{\alpha/2} = 2.31$ of student's t-distribution with $10 - 2 = 8$ degrees of freedom, we obtain $100(1 - 0.05) = 95$ percent confidence intervals

$$\begin{aligned} -0.607 = -0.321 - 2.31 \times 0.124 &\le \beta_1 \\ &\le -0.321 + 2.31 \times 0.124 \\ &= -0.035, \end{aligned}$$

which does not include zero, so the intercept is statistically significant, and

$$\begin{aligned} 0.113 = 0.210 - 2.31 \times 0.042 &\le \beta_2 \\ &\le 0.210 + 2.31 \times 0.042 \\ &= 0.307. \end{aligned}$$

Computing the corresponding test statistics, we see that

$$Z_1^* = -0.321/0.124 = -2.589,$$

which exceeds 2.31 in absolute value, the critical value of student's t-distribution with $10 - 2 = 8$ degrees of freedom, so it is statistically significant at the 5 percent level using the test based on robust variance estimates. Similarly,

$$Z_2^* = 0.210/.0463 = 4.537,$$

which exceed 2.31, the critical value of the t-distribution with 8 degrees of freedom, so it is statistically significant at the 5 percent level using the test based on robust variance estimates.

12.4.2.6 Collinearity and Regression Diagnostics All of the problems that arise in connection with multiple regression analysis can also arise in meta-analysis. Just as in regression analysis in primary research, many diagnostics are available for regression analysis in meta-analysis (see, for example, Hedges and Olkin 1985, chapter 12; Viechtbauer and Cheng 2010). Some of these diagnostics are used to evaluate the goodness of fit of the regression model, including standardized residuals, the change in the test statistic for residual variation when a study is deleted, or the change in the estimated residual variance component when a study is deleted. Other diagnostics address the influence of a particular study on the results of the regression analysis, including the relative weight given to each study, the change in the regression coefficients or the fitted values when a study is deleted or more theoretical quantities such as the leverage of each study (diagonal elements of the weighted hat matrix $\mathbf{X}(\mathbf{X}'\mathbf{W}^*\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^*$).

Collinearity may degrade the quality of estimates of regression coefficients in primary research studies, wildly influencing their values and increasing their standard errors. The same procedures used to safeguard against excessive collinearity in multiple regression analysis in primary research are useful in meta-analysis. Examination of the correlation matrix of the predictors and the exclusion of some predictors that are too highly intercorrelated with the others can often be helpful. In some cases, predictor variables derived from critical study characteristics may be too highly correlated for any meaningful analysis using more than a very few predictors. It is important to recognize, however, that collinearity is a limitation of the data (reflecting little information about the independent relations among variables) and not an inadequacy of the statistical method. Thus highly collinear predictors based on study characteristics imply that the studies simply do not have the array of characteristics that might make it possible to ascertain precisely their joint relationship with study effect size. In our example about gender differences in conformity, the correlation coefficient between percentage of male authors and number of items is 0.664. The correlation between percentage of male authors and log items is even higher, 0.793.

12.5 QUANTIFYING EXPLAINED VARIATION

Although the Q_E statistic (or the Q_W statistic for models with categorical independent variables) is a useful test statistic for assessing whether any unexplained variation is statistically reliable, it is not a useful descriptive statistic for quantifying the amount of unexplained variation. Such quantifying typically involves imposing a model of randomness on this unexplained variation. This is equivalent to imposing a mixed model on the data for the purposes of describing unexplained variation. The details of such models are beyond the scope of this chapter (for a discussion, see chapter 16).

A descriptive statistic R_B (called the Birge ratio) has long been used in the physical sciences (see, for example, Birge 1932) and was recently proposed as a descriptive statistic for quantifying unexplained variation in medical meta-analysis with no covariates, where it was called H^2 (see Higgins and Thompson 2002). It is the ratio of Q_E (or Q_W) to its degrees of freedom, that is,

$$R_B = Q_E/(k-p-1),$$

(for a regression model with intercept) or

$$R_B = Q_W/(k-p),$$

(for a categorical model without intercept). The expected value of R_B is exactly 1 when the effect parameters are determined exactly by the linear model. When the model does not fit exactly, R_B tends to be larger than one. The Birge ratio has the crude interpretation that it estimates the ratio of the between-study variation in effects to the variation in effects due to (within-study) sampling error. Thus a Birge ratio of 1.5 suggests that there is 50 percent more between-study variation than might be expected given the within-study sampling variance.

The squared multiple correlation between the observed effect sizes and the predictor variables is sometimes used as a descriptive statistic. However, the multiple correlation may be misinterpreted in this context because the maximum value of the population multiple correlation is always less than one, and can be much less than one. The reason is that the squared multiple correlation is a measure of "variance (in the observed effect-size estimates) accounted for" by the predictors. But there are two sources of variation in the effect-size estimates: between-study (systematic) effects and within-study (nonsystematic or sampling) effects. We might write this partitioning of variation symbolically as

$$\mathrm{Var}[T] = \sigma\theta^2 + v,$$

where $\mathrm{Var}[T]$ is the total variance, σ_θ^2 is the between-study variance in the effect-size parameters and v is the within-study variance (the variance of the sampling errors). Only between-study effects are systematic and therefore only they *can* be explained via predictor variables. Variance due to within-study sampling errors cannot be explained. Consequently the maximum proportion of variance that could be explained is determined by the proportion of total variance that is due to between-study effects. Thus the maximum possible value of the squared multiple correlation could be expressed (loosely) as

$$\frac{\sigma_\theta^2}{\sigma_\theta^2 + v} = \frac{\sigma_\theta^2}{\mathrm{Var}[T]}.$$

Clearly this ratio can be quite small when the between-study variance is small relative to the within-study variance. For example, if the between-study variance (component) is 50 percent of the (average) within-study variance, the maximum squared multiple correlation would be

$$0.5/(0.5+1.0) = 0.33.$$

In this example, predictors that yielded an R^2 of 0.30 would have explained 90 percent of the *explainable* variance even though they explain only 30 percent of the *total* variance in effect estimates.

A better measure of explained variance than the conventional R^2 would be based on a comparison of the between-study variance in the effect-size parameters in a model with no predictors and that in a model with predictors. If $\sigma_{\theta 0}^2$ is the variance in the effect-size parameters in a model with no predictors and $\sigma_{\theta 1}^2$ is the variance in the effect-size parameters in a model with predictors, then the ratio

$$P_{MA}^2 = \frac{\sigma_{\theta 0}^2 - \sigma_{\theta 1}^2}{\sigma_{\theta 0}^2} = 1 - \frac{\sigma_{\theta 1}^2}{\sigma_{\theta 0}^2}$$

of the explained variance to the total variance in the effect-size parameters is an analog to the usual concept of

squared multiple correlation in the meta-analytic context. Such a concept is widely used in the general mixed model in primary data analysis (see, for example, Bryk and Raudenbush 1992, 65). The parameter P^2_{MA} is interpretable as the proportion of explainable variance that is explained in the meta-analytic model.

12.6 CONCLUSION

Fixed- and random- (or mixed-) effects approaches to meta-analysis provide a variety of techniques for statistically analyzing effect sizes. These techniques are analogous to fixed- and mixed-effects statistical methods commonly used in the analysis of primary data, such as variance and multiple regression. Consequently, familiar analysis strategies (such as contrasts from analysis of variance) or coding methods (such as dummy or effect coding from multiple regression analysis) can be used in meta-analysis just as they are in primary analyses. The choice between fixed- or random- (mixed-) effects models should be driven primarily by the kind of inference the meta-analyst wants to make. If the inference is restricted to the sample of studies that are observed (that is, conditional), the fixed-effects approach is appropriate. In contrast, if the inference drawn is about the population of studies from which the observed studies are considered to be a random (representative) sample, the random- or mixed-effects models are appropriate. Fixed-effects models may also be reasonable when the number of studies is too small to support the effective use of mixed- or random-effects models. In practice, it is not unusual for meta-analysts to determine whether effect-size estimates between studies show any systemic variability. If there is indeed systematic variation between studies that the analyst believes needs to be incorporated in the analyses, the random- or mixed-effects model seems appropriate.

12.7 REFERENCES

Becker, Betsy J. 1986. "Influence Again: An Examination of Reviews and Studies of Gender Differences in Social Influence." In *The Psychology of Gender: Advances Through Meta-Analysis*, edited by Janet S. Hyde and Marcia C. Linn. Baltimore, Md.: Johns Hopkins University Press.

Birge, Raymond T. 1932. "The Calculation of Errors by the Method of Least Squares." *Physical Review* 40: 207–27.

Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. 2009. *Introduction to Meta-Analysis.* Hoboken, N.J.: John Wiley & Sons.

———. 2013. "Comprehensive Meta-Analysis, Version 3." Accessed December 8, 2018. http://www.meta-analysis.com.

Bryk, Anthony S., and Stephen W. Raudenbush. 1992. *Hierarchical Linear Models*. Thousand Oaks, Calif.: Sage Publications.

Eagly, Alice H., and Linda L. Carli. 1981. "Sex of Researchers and Sex Typed Communication as Determinants of Sex Differences in Influenceability: A Meta-Analysis of Social Influence Studies." *Psychological Bulletin* 90(1): 1–20.

Goldstein, Harvey. 1986. "Multilevel Mixed Linear Model Analysis Using Iteratively Generalized Least Squares." *Biometrika* 73(1): 43–56.

———. 1987. *Multilevel Models in Educational and Social Research*. New York: Oxford University Press.

———. 1989. "Restricted (Unbiased) Iterative Generalised Least Squares Estimation." *Biometrika* 76(3): 622–23.

Hartung, Joachim, and Guido Knapp. 2001. "On Tests of the Overall Treatment Effect in Meta-Analysis with Normally Distributed Responses." *Statistics in Medicine* 20(12): 1771–82.

Hedges, Larry V. 1982a. "Fitting Categorical Models to Effect Sizes from a Series of Experiments." *Journal of Educational Statistics* 7(2): 119–37.

———. 1982b. "Fitting Continuous Models to Effect Size Data." *Journal of Educational Statistics* 7(4): 245–70.

———. 1983a. "A Random Effects Model for Effect Sizes." *Psychological Bulletin* 93(2): 388–95.

———. 1983b. "Combining Independent Estimators in Research Synthesis." *British Journal of Mathematical and Statistical Psychology* 36(1): 123–31.

Hedges, Larry V., and Ingram Olkin. 1985. *Statistical Methods for Meta-Analysis*. Orlando, Fl.: Academic Press.

Hedges, Larry V., Elizabeth Tipton, and Matt Johnson. 2010. "Robust Variance Estimation for Meta-Regression with Dependent Effect Size Estimators." *Journal of Research Synthesis Methods* 19(1): 39–65.

Hedges, Larry V., and Jack L. Vevea. 1998. "Fixed and Random Effects Models in Meta-Analysis." *Psychological Methods* 3(4): 486–504.

Higgins, Julian P. T., and Simon G. Thompson. 2002. "Quantifying Heterogeneity in Meta-Analysis." *Statistics in Medicine* 21(11): 1539–58.

Higgins, Julian, Simon G. Thompson, Jonathan J. Deeks, and Doug Altman. 2003. "Measuring Inconsistency in Meta-Analysis." *British Medical Journal* 327(1): 557–60.

Knapp, Guido, and Joachim Hartung. 2003. "Improved Tests for a Random Effects Meta-Regression with a Single Covariate." *Statistics in Medicine* 22(17): 2693–710.

Konstantopoulos, Spyros. 2011. "Fixed Effects and Variance Components Estimation in Three-Level Meta-Analysis?" *Research Synthesis Methods* 2(1): 61–76.

———. 2013. "Meta-Analysis." In *Handbook of Quantitative Methods for Educational Research*, edited by Timothy Teo. Rotterdam: Sense Publishers.

Longford Nick T. 1987. "A Fast Scoring Algorithm for Maximum Likelihood Estimation in Unbalanced Mixed Models with Nested Random Effects." *Biometrika* 74: 817–27.

Miller, Rupert G., Jr. 1981. *Simultaneous Statistical Inference*, 2nd ed. New York: Springer-Verlag.

Raudenbush, Stephen W., and Anthony S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd ed. Thousand Oaks, Calif.: Sage Publications.

Tipton, Elizabeth. 2015. "Small Sample Adjustments for Robust Variance Estimation with Meta-Regression." *Psychological Methods* 20(3): 375–93.

Viechtbauer, Wolfgang. 2005. "Bias and Efficiency of Meta-Analytic Estimators in the Random-Effects Model." *Journal of Educational and Behavioral Statistics* 30(3): 261–94.

Viechtbauer, Wolfgang, and Mike W. L Cheng. 2010. "Outlier and Influence Diagnostics for Meta-Analysis." *Research Synthesis Methods* 1(2): 112–25.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, Mass.: MIT Press.

13

STOCHASTICALLY DEPENDENT EFFECT SIZES

LARRY V. HEDGES
Northwestern University

CONTENTS

13.1 INTRODUCTION

Previous chapters on statistical analysis of effect sizes have focused on situations in which each study yields only a single estimate of effect size and all estimates are independent. However, it is possible that effect-size estimates are not statistically independent. In the most abstract sense, an effect-size estimate T can be decomposed into an effect-size parameter θ and an estimation error $\varepsilon \equiv T - \theta$. Thus two effect-size estimates T_1 and T_2 can be decomposed into $T_1 = \theta_1 + \varepsilon_1$ and $T_2 = \theta_2 + \varepsilon_2$. The two effect-size estimates T_1 and T_2 can be statistically dependent because the estimation errors ε_1 and ε_2 are not independent (for example, the ε_i are correlated), because the effect-size parameters θ_1 and θ_2 are random but not independent (for example, the θ_i are correlated) or both. The first source of dependence (correlated estimation errors) has attracted the most attention in meta-analysis, but the second source of dependence (correlated study-level random effects) can be a serious concern in some situations.

13.1.1 Dependence Due to Correlated Estimation Errors

Dependence occurs in three common situations. First, each individual may be assessed using several different measures of outcome (for example, different measures of mathematics achievement), and an effect size may be computed using data on each of the several measures. Second, the same individuals may be measured at several different points in time (such as just after treatment and then at various follow-up periods) using the same measuring instruments, and an effect size may be computed at each time point. Third, in studies with multiple treatment groups, an effect size may be computed by comparing each of the several treatment groups with the same control group, and because the estimates so computed share a common control group, they are not independent.

13.1.2 When Estimation Errors Are Not Independent

In each of the cases mentioned, effect sizes may be sorted into groups so that there is only one effect size per study and thus all of the effect-size estimates in a particular group are independent. For example, groups might be effect sizes based on the same measure (such as mathematics computation), the same follow-up interval (such as approximately six months after treatment), or the same particular treatment (such as the standard rather than the enhanced variation of treatment versus control). The summary of independent effect-size estimates across groups can be accomplished via standard meta-analytic methods such as those described in chapter 12.

However, reviewers often want to carry out analyses that involve combining data across groups of effect-size estimates—analyses that involve effect-size estimates that are not all independent. Two types of combined analyses are the most common. One involves estimating a mean effect size across all types of outcomes or treatment variations (such as mathematics computation and problem solving, different follow-up intervals, or variations of the treatment). Such analyses are often desired to answer the most global questions about whether the treatment had an impact. A second kind of analysis involves estimating differential effects of treatment. This type is often desired to answer questions about whether the treatment has a bigger effect on some outcomes than others, at some follow-up intervals than others, or whether certain variations of treatment have bigger effects.

Five strategies are commonly used for handling effect-size data involving such non-independence. The first is to explicitly model correlations among the effect sizes using multivariate methods (see, for example, Hedges and Olkin 1985; Kalaian and Raudenbush 1996). This strategy is the most elegant, providing the most efficient estimates of effects and accurate results of significance tests. Unfortunately, because it requires knowledge of the covariance structure of the effect-size estimates (which in turn requires knowledge of the dependence structure of the raw data in each study), the information needed to implement this strategy is rarely available.

The second strategy is to estimate the fixed effects (for example, the mean or the regression coefficients in a linear model) using standard weighted meta-analysis methods, but to use robust computations of the variances that take into account the dependence of the effect-size estimates within studies (see, for example, Hedges, Tipton, and Johnson 2010). These robust methods do not require any knowledge of the covariance structure of estimates within studies and they are relatively easy to compute.

The third strategy is to first compute a within-study summary from non-independent effect-size estimates and then summarize the (independent) summaries across

studies. For example, to compute the overall mean effect size (across effect sizes of all types from all studies) one might first compute the mean effect size within each study as a kind of synthetic effect size for the study. The effect sizes within a study will not be independent, but the synthetic effect sizes (the study-average effect sizes) from different studies will be independent, and consequently conventional meta-analytic methods can be used to combine the summary effect sizes across studies.

Similarly, one might compute the difference between the average effects of two types (outcome measures, follow-up intervals, or treatment types) by first computing the difference between the two effects within each study as another type of synthetic effect size. The effect sizes within a study are not independent, but the synthetic effect sizes (difference between effect sizes of the two types) from different studies will be independent.

In either case, because the synthetic effect sizes are independent, they can be combined across studies using conventional methods for meta-analysis. There is, however, one problem in using conventional meta-analytic methods to combine the synthetic effect sizes: the standard errors of these synthetic effect sizes depend on the correlation structure of the within-study effect sizes from which they are computed, and thus are typically not known.

The fourth strategy is to use Bayesian methods. These have a particular advantage in that they can introduce (weakly or strongly) informative prior distributions to help deal with the missing data issues that often arise in the context of multivariate meta-analyses. Although there is much to recommend this approach, we do not consider it in this chapter (but see Wei and Higgins 2011).

The fifth strategy is to ignore the fact that some of the effect-size estimates are not independent, and to use the same meta-analytic procedures that would have been used had the effect-size data been independent. This approach is naïve and usually misguided, but may not be too misleading if relatively few studies report more than one effect size. Moreover, in some cases, it may lead to conservative results for tests of the difference between average effects of different types.

Although the third and fourth strategies (use of synthetic effect sizes and ignoring dependence) seem to be widely used for dealing with non-independent effect-size estimates, they have serious limitations. The first strategy (full multivariate analysis) is the most elegant, but it is also often difficult or impossible to implement. A major exception is when dependence arises when a common control group is compared with several treatments (see, for example, Gleser and Olkin 2009). The second strategy (robust variance estimation) has considerable merit and is probably the best practical approach in most meta-analytic situations.

This chapter describes how to handle dependent effect sizes within studies in meta-analyses conceived broadly. Because possible analyses are numerous, I describe them primarily in terms of linear models for effect sizes (meta-regression), which includes both estimation of the mean effect size and categorical (analysis of variance style) models as special cases.

13.1.3 Dependence Among Study Effects

This chapter focuses on dependence related to correlated estimation errors. Another model of dependence, however, relates to dependence across studies, which arises through the study-level random effects when groups or clusters of studies exhibit less variation in their effect-size parameters than the collection of effect-size parameters as a whole. For example, the entire collection of effect sizes may result from studies conducted by different investigators. If an investigator contributes multiple studies, then some methodological features might be the same in all the studies. and the effect-size parameters (the true effect sizes) produced by one investigator might vary less than those produced by others. Such a situation induces a correlation among effect-size parameters from the same investigator. Because the structure of the data is that effect sizes (hierarchically) nested within investigators, and the effect sizes generated by the same investigator are correlated, this model is sometimes called the hierarchical dependence model (see Stevens and Taylor 2009).

Other kinds of hierarchical dependence are also possible, but empirical evidence that they occur is scant. For example, experiments involving different samples of individuals but reported in the same publication might be considered hierarchically dependent. Similarly, studies conducted by the same laboratory or by groups of researchers sharing common methodological conventions (such as a research mentor and former doctoral students) might be considered hierarchically dependent.

Note that this form of dependence arises through the effect-size parameters, not the estimation errors. However, if studies in such a hierarchical structure measure several effect-size estimates with correlated estimation errors,

both kinds of dependence (that induced by correlations among effect-size parameters and correlations induced by estimation errors) can occur in the same data.

Although hierarchical dependence potentially has important consequences in some situations, it has not been of great interest in most meta-analyses. Moreover, because this form of dependence requires estimation of complex variance component structures, it requires a fairly large number of studies for precise estimation and many meta-analytic datasets have too little information (in the statistical sense) for estimation. Consequently, we do not consider it further. However such dependence can arise and should be at least considered as a possibility. The robust estimation methods considered in section 13.4 of this chapter can be used to provide valid analyses for hierarchical dependence models (see Hedges, Tipton, and Johnson 2010). However, other analyses of hierarchical dependence structures are also available (see Stevens and Taylor 2009; Konstantopoulos 2011).

13.2 MULTIVARIATE DATA STRUCTURES

The structure of effect-size data in univariate meta-analysis is reasonably simple, but the data structure in multivariate meta-analysis is more complex. In univariate meta-analysis, there may be study-level covariates that moderate (explain variation in) the effect parameters, but all of the effect-size parameters and estimates refer to the same construct. That is, they are all effect sizes of a treatment on the same outcome. However, the assumption in multivariate meta-analysis is that some effect sizes refer to one treatment effect construct and that others refer to different treatment effect constructs.

For example, studies in education might measure outcomes on different achievement domains (for example, reading, mathematics, and science) leading to three types of effect-size constructs (effects on reading, effects on mathematics, and effects on science). Alternatively, studies might measure outcomes at different follow-up periods or reflecting comparisons between variations of treatment, resulting in different effect-size constructs. If some or all of the outcome measurements used to compute different effect-size estimates in a study come from the same individuals, then the effect-size estimates will have correlated estimation errors.

Whether effect sizes refer to the same or different constructs is a theoretical decision determined by the breadth of the construct defined in the problem formulation aspect of the review. For example, for some purposes it might be useful to define the effect size of a particular treatment on a broad construct of mathematics achievement, including achievement in algebra, geometry, and arithmetic computation. For more refined purposes, it might be preferable to differentiate the construct of mathematics achievement into achievement in algebra, geometry, and arithmetic computation as separate effect-size constructs.

Because it is frequently the case that every study in a meta-analysis may not provide estimates of all of the effect-size constructs, the data structure is more complex than in univariate meta-analysis or even in multivariate analysis of primary data. Therefore we describe a notation for identifying which effect-size constructs are represented in each study (so-called selection matrices) to handle this complexity of multivariate effect-size data.

13.2.1 Multivariate Distribution of Effect Sizes

Exceptions aside, such as methods for handling sparse 2×2 tables (the Mantel-Haenszel method, for example), in most of this book we stress the unity of statistical methods for meta-analysis. We discuss, in detail, different effect-size estimates and how to compute their variances in chapter 11. However, once the effect-size estimates and their variances are computed, other analyses are identical for all effect-size estimates. Just as the sampling distributions (means and variances) of univariate effect-size estimates depend on details of the designs, their multivariate analogs (the joint distributions of effect-size estimates) also depend on details for the design.

Previous editions of this handbook have included explicit formulas for computing the covariance matrix of effect-size estimates based on knowledge of the covariance matrix of the original measurements. Often these computations suggest that the effect-size estimates have (very) approximately the same correlation as the underlying measurements. In this edition, we decided to omit these formulas—for four reasons. First, the formulas are different for each type of effect size, which means that a complete treatment would be extensive. Second, the formulas are somewhat different depending on the details of the effect-size calculation, making a complete treatment even more extensive. Third, and perhaps most important, these formula are often of very little use because they depend on the correlations among the outcome constructs being measured and these underlying correlations among outcomes are often not very well known. Finally, new

methods (particularly the robust methods described later) make the use of these computations less important because they can provide valid multivariate analyses that do not depend on knowledge of the correlation structure of the correlated effect-size estimates within studies.

Expositions of the multivariate distribution of various effect-size estimators are available elsewhere. Harold Hotelling provides an exposition of the multivariate structure of correlated correlation coefficients (1940). J. J. Neil and Olive Dunn give an exposition of the multivariate structure of correlated z-transformed correlation coefficients (1975). Leon Gleser and Ingram Olkin offer an exposition of the multivariate structure of both log risk ratios and log odds ratios (2000) and the multivariate structure of standardized mean differences (2009).

13.3 FULL MULTIVARIATE METHODS FOR DEPENDENT EFFECT SIZES

In this section, we briefly sketch the procedure for the full multivariate analysis of effect sizes based on a linear random-effects model. The analysis for estimation of the mean effect can be carried out by using a design matrix with an intercept and no predictors (that is, the design matrix is a vector of ones). Fixed-effects analyses can be carried out by constraining the covariance matrix $\boldsymbol{\Psi}$ of the random effects $\boldsymbol{\xi}_i$ to be zero.

13.3.1 Models and Notation

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$ correspond to the entire collection of effect-size parameters whose estimates are observed in any of the studies, and assume that each study observes an estimate of some nonempty subset of those parameters. For example, there might be studies with outcomes of reading, mathematics, and science achievement, but some studies observed effect-size estimates of only math and science achievement. Alternatively, there might be studies that observed the outcome at immediate, six-month, and one-year follow-up intervals, but some that observed effect-size estimates only for immediate and one-year intervals.

Suppose that there are $k \leq m$ studies with $1 \leq p_i \leq m$ possibly correlated effect-size estimates arising in the ith study. Let T_{ij} be the jth effect-size estimate from the ith study, respectively, and denote the p_i dimensional column vector of effect-size estimates from the ith study by $\mathbf{T}_i = (T_{i1}, \ldots, T_{ip_i})'$. Not every study will necessarily estimate every one of the possible effect-size parameters. We could describe this situation by using a $p_i \times m$ selection matrix $\mathbf{A}_i$ whose jth row has unity in the sth column if T_{ij} estimates θ_s, and zero otherwise.

For example, if $m = 5$ so that $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)'$ and the ith study has estimates of θ_1, θ_3, and θ_4, but not θ_2 or θ_5, then $p_i = 3$, T_{i1} estimates θ_1, T_{i2} estimates θ_3, and T_{i3} estimates θ_4. The matrix $\mathbf{A}_i$ is therefore

$$\mathbf{A}_i = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

and thus the vector $\mathbf{T}_i$ of estimates of the ith study estimates $\mathbf{A}_i\boldsymbol{\theta} = (\theta_1, \theta_3, \theta_4)'$.

Assume that each $\mathbf{T}_i$ has a p_i-variate normal distribution with mean $\mathbf{A}_i\boldsymbol{\theta}$ with known $p_i \times p_i$ covariance matrix $\boldsymbol{\Sigma}_i$, that is,

$$\mathbf{T}_i - \mathbf{A}_i\boldsymbol{\theta} = \boldsymbol{\varepsilon}_i \sim \mathrm{N}(\mathbf{0}, \Sigma_i), i = 1, \ldots, k, \qquad (13.1)$$

so that the vector of (correlated) estimation errors for the ith study $\boldsymbol{\varepsilon}_i = \mathbf{T}_i - \mathbf{A}_i\boldsymbol{\theta}$ has covariance matrix $\boldsymbol{\Sigma}_i$. The studies need not all have the same number of effect sizes. Assume that the effect-size parameter vector $\mathbf{A}_i\boldsymbol{\theta}$ for the *ith* experiment depends on a $p_i \times q$ matrix of $q \geq 1$ fixed concomitant variables

$$\mathbf{A}_i\mathbf{X}_i = \begin{pmatrix} x_{i11} & \cdots & x_{i1q} \\ \vdots & \ddots & \vdots \\ x_{ip_i1} & \cdots & x_{ip_iq} \end{pmatrix}, \qquad (13.2)$$

so that

$$\mathbf{A}_i\boldsymbol{\theta} = \mathbf{A}_i\mathbf{X}_i\boldsymbol{\beta} + \mathrm{A}_i\boldsymbol{\xi}_i \qquad (13.3)$$

$\boldsymbol{\xi}_i = (\xi_1, \ldots, \xi_m)'$ is an $m \times 1$ vector of study-level random effects (only p_i of which are realized because only p_i of the m components of $\boldsymbol{\theta}$ are realized in the ith study), $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)'$ is a $q \times 1$ vector of unknown regression coefficients, and the $m \times q$ matrix $\mathbf{X}_i$ can be conceived as the matrix $\mathbf{A}_i\mathbf{X}_i$ with $(m - p_i)$ rows of zeros inserted in the row corresponding to each component of θ that is *not* estimated by the ith study.

Consider again the example where $m = 5$, $\theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)'$ and the ith study has estimates of θ_1, θ_3, and θ_4,

but not θ_2 or θ_5, then $p_i = 3$, T_{i1} estimates θ_1, T_{i2} estimates θ_3, and T_{i3} estimates θ_4. The matrix $\mathbf{A}_i$ is

$$\mathbf{A}_i = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Because θ_2 and θ_5 are not estimated in study i, the matrix $\mathbf{X}_i$ is defined by starting with $\mathbf{A}_i\mathbf{X}_i$, and inserting a row of zeros as the second and fifth rows to create $\mathbf{X}_i$ as

$$\mathbf{X}_i = \begin{pmatrix} x_{i11} & \cdots & x_{i1q} \\ 0 & \cdots & 0 \\ x_{i21} & \cdots & x_{i2q} \\ x_{i31} & \cdots & x_{i3q} \\ 0 & \cdots & 0 \end{pmatrix}.$$

Just as the sth row of $\mathbf{A}_i\boldsymbol{\theta}$ may refer to different components of $\boldsymbol{\theta}$ in different studies, the values of the covariates in the sth row of $\mathbf{A}_i\mathbf{X}_i$ are intended to predict different components of $\boldsymbol{\theta}$ in different studies.

Denote the $p = p_1 + \ldots + p_k$ dimensional column vector of sample effect sizes by $\mathbf{T} = (\mathbf{T}_1', \ldots, \mathbf{T}_k')'$, and the $m \times kq$ selection matrix $\mathbf{A} = \text{diag}(\mathbf{A}_1, \ldots, \mathbf{A}_m)$, and the $m \times q$ design matrix $\mathbf{X}$ by stacking the matrices $\mathbf{X}_i$

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_k \end{pmatrix},$$

so that $\mathbf{AX}$ is the full design matrix. Thus the total number of effect sizes and estimates across all k studies is $p \geq k$ and when $p = \text{k}$ then there is one effect size per study and all the effect sizes are independent.

Assume that the effect-size parameters $\boldsymbol{\theta}$ are determined by a linear model of the form

$$\boldsymbol{\theta} = \mathbf{AX}\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\xi} = \begin{pmatrix} \boldsymbol{\theta}_1 \\ \vdots \\ \boldsymbol{\theta}_k \end{pmatrix} = \begin{pmatrix} \mathbf{A}_1\mathbf{X}_1 \\ \vdots \\ \mathbf{A}_k\mathbf{X}_k \end{pmatrix}\boldsymbol{\beta} + \begin{pmatrix} \mathbf{A}_1\boldsymbol{\xi}_1 \\ \vdots \\ \mathbf{A}_k\boldsymbol{\xi}_k \end{pmatrix}, \quad (13.4)$$

so that the effect-size estimates $\mathbf{T}$ are given by a linear model of the form

$$\mathbf{T} = \mathbf{AX}\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\xi} + \boldsymbol{\varepsilon} = \begin{pmatrix} \mathbf{T}_1 \\ \vdots \\ \mathbf{T}_k \end{pmatrix} = \begin{pmatrix} \mathbf{A}_1\mathbf{X}_1 \\ \vdots \\ \mathbf{A}_k\mathbf{X}_k \end{pmatrix}\boldsymbol{\beta} + \begin{pmatrix} \mathbf{A}_1\boldsymbol{\xi}_1 \\ \vdots \\ \mathbf{A}_k\boldsymbol{\xi}_k \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_k \end{pmatrix}, \quad (13.5)$$

where $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{ip_i})'$ is a vector of the p_i estimation errors of the effect-size estimates in the ith study defined in (13.1), where $\boldsymbol{\varepsilon}_i$ and $\boldsymbol{\xi}_i$ are normally distributed with covariance matrices $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\Psi}$, respectively, and are independent of one another.

13.3.2 Regression Coefficients and Covariance Components

The regression coefficient vector $\boldsymbol{\beta}$ and the covariance component matrix $\boldsymbol{\Psi}$ can be estimated by weighted least squares as in the case of the univariate mixed model. The usual procedure is to first estimate the covariance component matrix $\boldsymbol{\Psi}$ and then reweight to estimate the regression coefficient vector $\boldsymbol{\beta}$ and its covariance matrix $\boldsymbol{\Sigma}_\beta$. There are usually advantages (among them software availability) in considering the problem as a special case of the hierarchical linear model considered in chapter 12 in conjunction with univariate mixed-model analyses. The multivariate mixed-model analyses can be carried out as instances of the multivariate hierarchical linear model (see Thum 1997), estimating parameters by the method of by maximum likelihood. However a simpler alternative is available.

If the sampling error covariance matrix $\boldsymbol{\Sigma}_i$ is known, it is possible to transform the within-study model so that the sampling errors are independent with the same variances as the components of the $\mathbf{T}_i$ (see Raudenbush, Becker, and Kalaian 1988). Note that it is conventional in meta-analysis to treat variances as known (because they depend on samples sizes, which are known), but the covariances in the $\boldsymbol{\Sigma}_i$ are typically unknown. The correlation matrix of the estimation errors is

$$\mathbf{P}_i = \mathbf{D}_i^{-1}\boldsymbol{\Sigma}_i\mathbf{D}_i^{-1}, \quad (13.6)$$

where $\mathbf{D}_i$ is a diagonal matrix of the standard errors (square roots of the variances) of the components of $\mathbf{T}_i$, $\mathbf{D}_i = \text{diag}\left(\sqrt{v_{i1}}, \ldots, \sqrt{v_{ip_i}}\right)$. For each study perform the

Cholesky factorization of the correlation matrix of the estimation errors $\mathbf{P}_i$ so that

$$\mathbf{P}_i = \mathbf{F}_i\mathbf{F}_i', \tag{13.7}$$

where $\mathbf{F}_i$ is a known matrix (since $\mathbf{P}_i$ is a known matrix). The transformed effect-size vector $\mathbf{Z}_i$ given by

$$\mathbf{Z}_i = \mathbf{D}_i\mathbf{F}_i^{-1}\mathbf{D}_i^{-1}\mathbf{T}_i = \mathbf{C}_i\mathbf{T}_i \tag{13.8}$$

where $\mathbf{C}_i = \mathbf{D}_i\mathbf{F}_i^{-1}\mathbf{D}_i^{-1}$ has a sampling error vector

$$\tilde{\boldsymbol{\varepsilon}}_i = \mathbf{D}_i\mathbf{F}_i^{-1}\mathbf{D}_i^{-1}\boldsymbol{\varepsilon}_i = \mathbf{C}_i\boldsymbol{\varepsilon}_i \tag{13.9}$$

which has covariance matrix $\mathbf{D}_i^2 = \text{diag}\,(v_{i1}, \ldots, \mathbf{v}_{ipi})$. When there is only a single effect-size estimate from a study, the transformation matrix is understood to be the identity so that $\mathbf{C}_i = 1$ and $\mathbf{D}_i = v_{i1}$.

Premultiplying both sides of the model given in formula (13.5) by $\mathbf{C}_i$, the within-study model for the transformed effect-size vector Z_i becomes

$$\mathbf{C}_i\mathbf{T}_i = \mathbf{C}_i\mathbf{A}_i\boldsymbol{\theta} + \mathbf{C}_i\boldsymbol{\varepsilon}_i$$

Thus one might write the within-study (level I) model as

$$\mathbf{Z}_i = \mathbf{C}_i\mathbf{A}_i\boldsymbol{\theta}_i + \tilde{\boldsymbol{\varepsilon}}_i, \tag{13.10}$$

where the transformed effect-size estimates $\mathbf{Z}_i$ are independent with the same variances as the effect sizes in $\mathbf{T}_i$, and the effect-size parameter vector $\boldsymbol{\theta}$ is the same as in the original model.

Thus the within-study model (13.10) along with the between-study (level II) model

$$\boldsymbol{\theta}_i = \mathbf{A}_i\mathbf{X}\boldsymbol{\beta} + \mathbf{A}_i\boldsymbol{\xi} \tag{13.11}$$

is a conventional two-level hierarchical linear model with independent estimation errors at the first level. Therefore conventional software can be used to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\Psi}$ by the method of full or restricted maximum likelihood (as in HLM, SAS Proc Mixed, Stata, or the R package nlme) or iteratively reweighted least squares (as in ML3).

13.3.3 Tests and Confidence Intervals

The interpretation of results of the analysis is identical to that of any other multilevel model analysis. The regression coefficients are the fixed effects and their estimates $b_1, \ldots, b_q$ and their standard errors $S_1, \ldots, S_q$ will be given by the program. A $100(1 - \alpha)$ percent confidence interval for β_j is given by

$$b_j - C_{\alpha/2}\sqrt{S_j} \le \beta_j \le b_j + C_{\alpha/2}\sqrt{S_j}, \tag{13.12}$$

where C_α is the $100(1 - \alpha)$ percent point of t-distribution with $k - q$ degrees of freedom.

The test of the hypothesis

$$\mathrm{H}_0\colon \beta_j = 0$$

uses the test statistic

$$t_j = b_j \big/ \sqrt{S_j} \tag{13.13}$$

where S_j is the standard error of b_j, which is compared with critical values of student's t-distribution with $k - q$ degrees of freedom.

The variance and covariance components of $\mathbf{C}_i\mathbf{A}_i\boldsymbol{\xi}$ are the elements of $\boldsymbol{\Psi}$. The diagonal elements ψ_{jj} are the variance components and the off-diagonal elements ψ_{st} correspond to the covariances between the random effects, which may be more interpretable when expressed as the correlations between the random effects. The correlation between the sth and tth random is

$$\psi_{st} \big/ \sqrt{\psi_{ss}\psi_{tt}}\,.$$

One aspect of this procedure that complicates interpretation is that the random effects, as the predictors are transformed by the matrices $\mathbf{C}_i$ in this analysis (for a discussion of this issue and an extended tutorial, see Begos 2015).

Example. Consider the $k = 26$ randomized experiments on SAT coaching reported by Sema Kalaian and Stephen Raudenbush (1996). Five of these studies reported outcomes for both SAT verbal scores (SATV) and SAT mathematics scores (SATM), eight reported results only for SATM, and fifteen reported results only for SATV. The data from these studies are reproduced in table 13.1. Suppose we were interested in estimating the mean effect size for SATV and SATM, then $\boldsymbol{\theta} = (\theta_1, \theta_2)'$, where $\theta_1 =$ the effect size for SATV and θ_2 is the mean effect size for SATM. Define the selection matrix $\mathbf{A}_i$ so that it selects the

Table 13.1 Data from Twenty-Six Randomized Experiments on the Effects of SAT Coaching

Study	n^T	n^C	T	x	v	$\bar{T}_{i\bullet}$	$\bar{v}_{i\bullet}$
1	28	22	0.22	0	0.0817	0.22	0.0817
2	39	40	0.09	0	0.0507	0.09	0.0507
3	22	17	0.14	0	0.1045	0.14	0.1045
4	48	43	0.14	0	0.0442	0.14	0.0442
5	25	74	–0.01	0	0.0535	–0.01	0.0535
6	37	35	0.14	0	0.0557	0.14	0.0557
7	24	70	0.18	0	0.0561	0.18	0.0561
8	16	19	0.01	0	0.1151	0.01	0.1151
9	43	37	0.01	0	0.0503	0.01	0.0503
10	19	13	0.67	0	0.1366	0.67	0.1366
11	16	11	–0.38	0	0.1561	–0.38	0.1561
12	20	12	–0.24	0	0.1342	–0.24	0.1342
13	39	28	0.29	0	0.0620	0.29	0.0620
14	38	25	0.26	1	0.0669	0.26	0.0669
15	18	13	–0.41	1	0.1352	–0.41	0.1352
16	19	13	0.08	1	0.1297	0.08	0.1297
17	37	22	0.30	1	0.0732	0.30	0.0732
18	19	11	–0.53	1	0.1482	–0.53	0.1482
19	17	13	0.13	1	0.1360	0.13	0.1360
20	20	12	0.26	1	0.1344	0.26	0.1344
21	20	13	0.47	1	0.1303	0.47	0.1303
22	145	129	0.13	0	0.0147	0.13	0.0147
22	145	129	0.12	1	0.0147		
23	72	129	0.25	0	0.0218	0.16	0.0217
23	72	129	0.06	1	0.0216		
24	71	129	0.31	0	0.0221	0.20	0.0220
24	71	129	0.09	1	0.0219		
25	13	14	0.00	0	0.1484	0.04	0.1484
25	13	14	0.07	1	0.1484		
26	16	17	0.13	0	0.1216	0.31	0.1232
26	16	17	0.48	1	0.1248		

SOURCE: Author's tabulation.
NOTE: $x = 0$ if the effect size is for SATV and $x = 1$ if the effect size is for SATM.

effect size for θ_1 if the effect size is estimating SATV, and θ_2 if the effect size is estimating SATM. Thus the selection matrix $\mathbf{A}_i$ for studies one to thirteen in table 13.1 is

$$\mathbf{A}_i = (1 \quad 0), \tag{13.14}$$

the selection matrix A_i for studies fourteen to twenty-one is

$$\mathbf{A}_i = (0 \quad 1), \tag{13.15}$$

and the selection matrix A_i for studies twenty-two to twenty-six is

$$\mathbf{A}_i = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{13.16}$$

The overall $\mathbf{A}$ matrix therefore consists of a block diagonal of thirteen 1×2 $\mathbf{A}_i$ matrices of form (13.14), eight 1×2 $\mathbf{A}_i$ matrices of form (13.15), and five 2×2 $\mathbf{A}_i$ matrices of form (13.16). The design matrices $\mathbf{X}_i$ for all of the studies are 2×2 matrices and the overall design matrix $\mathbf{X}$ is a matrix of $k = 26$ stacked 2×2 matrices.

With this design matrix, the intercept β_1 is interpreted as the effect size on SATV and the slope β_2 is interpreted as the effect size for SATM. The object of the analysis is to estimate the mean effect size of coaching on SATV and SATM, develop confidence intervals for each, and test the hypothesis that the coaching effect is different from zero.

Because the first twenty-one studies have only a single effect size, there is only one estimation error in each of these studies, $\mathbf{C}_i = 1$ and $\mathbf{Z}_i = \mathbf{T}_i$. However, studies twenty-two to twenty-six have two effect sizes each and each of these estimation error covariance matrices must be factored. Using the correlation $\rho = 0.7$ between the effect sizes on SATV and SATM, so that the covariance between effect sizes on SATV and SATM is $0.7\sqrt{v_{i1}v_{i2}}$, we obtain the covariance matrices, transformation matrices, and transformed effect-size estimates for studies twenty-two to twenty-six as

$$\boldsymbol{\Sigma}_{22} = \begin{pmatrix} 0.0147 & 0.0103 \\ 0.0103 & 0.0147 \end{pmatrix}, \mathbf{C}_{22} = \begin{pmatrix} 1.000 & 0.000 \\ -0.980 & 1.400 \end{pmatrix},$$

$$\mathbf{Z}_{22} = \mathbf{C}_{22}\mathbf{T}_{22} = \begin{pmatrix} 0.13 \\ 0.04 \end{pmatrix}.$$

Similarly, we obtain

$$\boldsymbol{\Sigma}_{23} = \begin{pmatrix} 0.0218 & 0.0152 \\ 0.0152 & 0.0216 \end{pmatrix}, \mathbf{C}_{23} = \begin{pmatrix} 1.000 & 0.000 \\ -0.976 & 1.400 \end{pmatrix},$$

$$\mathbf{Z}_{23} = \mathbf{C}_{23}\mathbf{T}_{23} = \begin{pmatrix} 0.25 \\ -0.16 \end{pmatrix},$$

$$\boldsymbol{\Sigma}_{24} = \begin{pmatrix} 0.0221 & 0.0154 \\ 0.0154 & 0.0219 \end{pmatrix}, \mathbf{C}_{24} = \begin{pmatrix} 1.000 & 0.000 \\ -0.976 & 1.400 \end{pmatrix},$$

$$\mathbf{Z}_{24} = \mathbf{C}_{24}\mathbf{T}_{24} = \begin{pmatrix} 0.31 \\ -0.18 \end{pmatrix},$$

$$\boldsymbol{\Sigma}_{25} = \begin{pmatrix} 0.1484 & 0.1039 \\ 0.1039 & 0.1484 \end{pmatrix}, \mathbf{C}_{25} = \begin{pmatrix} 1.000 & 0.000 \\ -0.980 & 1.400 \end{pmatrix},$$

$$\mathbf{Z}_{25} = \mathbf{C}_{25}\mathbf{T}_{25} = \begin{pmatrix} 0.00 \\ 0.10 \end{pmatrix},$$

and

$$\boldsymbol{\Sigma}_{26} = \begin{pmatrix} 0.1216 & 0.0862 \\ 0.0862 & 0.1248 \end{pmatrix}, \mathbf{C}_{26} = \begin{pmatrix} 1.000 & 0.000 \\ -0.993 & 1.400 \end{pmatrix},$$

$$\mathbf{Z}_{26} = \mathbf{C}_{26}\mathbf{T}_{26} = \begin{pmatrix} 0.13 \\ 0.54 \end{pmatrix}.$$

In this case, it is simple to obtain simple algebraic formulas for the transformation from $\mathbf{T}_i$ to $\mathbf{Z}_i$—in this case, $Z_{i1} = T_{i1}$ and

$$Z_{12} = \frac{T_{i2} - \rho T_{i1}(v_{i2}/v_{i1})}{\sqrt{1-\rho^2}}.$$

The equation for Z_{i2} shows that the second element of the second row of $\mathbf{C}$ is the same for each study because it depends only on ρ, which is the same for each study. It also shows that the first element of the second row of each $\mathbf{C}$ matrix is quite similar for every study because it depends on ρ (which is constant) and the ratio v_{i2}/v_{i1}, which is nearly the same for every study. In cases with different correlations across studies or larger numbers of correlated effect sizes within a study, the pattern of transformations will be more complex.

The vector of effect-size estimates $\mathbf{T}$ consists of T_1 to T_{21} stacked on top of one another, stacked on top of $\mathbf{Z}_{22}$ to $\mathbf{Z}_{26}$, so that $\mathbf{Z} = (T_1, \ldots, T_{21}, \mathbf{Z}'_{22}, \ldots, \mathbf{Z}'_{26})'$.

Using HLM with the V Known option using the variances of the effect sizes in table 13.1, and the effect-size estimates given in $\mathbf{Z}$, we obtain estimates that the between-studies variance components are both zero (that is variances of ξ_1 and ξ_2 are both zero). This implies that the appropriate weights to apply are the inverses estimation error variances (fixed-effects weights).

With these weights, the analysis yields an estimate of β_1 (the mean effect on SATV) of $b_1 = 0.16$ with a standard error of 0.051 and a 95 percent confidence interval of $0.06 \le \beta_1 \le 0.26$. This corresponds to a test statistic of $Z_1 = 3.13$, which exceeds the 5 percent critical value of the standard normal distribution so we can reject the null hypothesis that the coaching effect on SATV is zero.

With these weights, the analysis yields an estimate of β_2 (the mean effect on SATM) of $b_2 = 0.09$ with a standard error of 0.057 and a 95 percent confidence interval of $-0.03 \le \beta_2 \le 0.20$. This corresponds to a test statistic of $Z_2 = 1.51$, which does not exceed the 5 percent critical value of the standard normal distribution so we cannot reject the null hypothesis that the coaching effect on SATV is zero.

13.3.4 Approximate Covariance Matrices

Although using the actual covariance matrices is desirable for the unequivocally correct results, simulation studies suggest that the results of multivariate meta-regressions were relatively insensitive to incorrect values of the within-study correlations (see, for example, Ishak et al. 2008; Riley 2009). This is consistent with our experience in using multivariate methods in meta-analysis. Thus a reasonable approach to multivariate meta-analysis in many situations (where estimation error covariances are not known precisely) is to use an approximate or working covariance matrix, which may be of simpler structure than the actual covariance matrix. The purpose of this working correlation is to acknowledge in the analysis some degree of correlation among estimates to provide a better approximation to the covariance matrix than assuming independence of estimates within studies would. This permits incorporating more information (more effect-size estimates) in the meta-analysis than if we were required to have all the estimates be independent, but does not substantially overestimate the amount of information as would be if the correlated estimates were treated as if they were independent.

Whenever working covariance matrices are used, they should be chosen based on some knowledge of the likely correlation among estimates. Moreover, it is advisable to do sensitivity analyses to understand how strongly the results of the meta-analysis depend on the approximate correlation structure chosen.

For example, if a study has several effect-size estimates related to constructs measured by different cognitive tests, one might posit that the (working or approximate) correlation has the same value ρ between estimates of any two of them. The approximate correlation matrix may be incorrect,

but if it is reasonably close to the correct value, the resulting analysis may be approximately correct, and may not be very sensitive to the precise choice of the correlation used to define the working covariance matrix.

Example. Return to the $k = 26$ randomized experiments on SAT coaching reporting outcomes for either SAT verbal scores or SAT mathematics scores, or both, as reported in table 13.1. Earlier, we carried out a multivariate analysis of these data assuming that the exact correlation between effect-size estimates on SATV and SATM within each study was $\rho = 0.7$. Table 13.2 shows the results obtained from multivariate analyses assuming $\rho = 0.0$, 0.5, 0.6, 0.7, and 0.8. The rows of the table correspond to different values of ρ. The left vertical panel of the table shows the estimate of the average effect size on SATV (b_1), its standard error (S_1), the lower and upper 95 percent confidence intervals (lcl_1 and ucl_1), and the test statistic for testing that $\theta_1 = 0$. The right vertical panel of the table shows the estimate of the average effect size on SATM (b_2), its standard error (S_2), the lower and upper 95 percent confidence intervals (lcl_2 and ucl_2), and the test statistic for testing that $\theta_2 = 0$. The table shows that the differences in the estimates of θ_1 associated with different values of ρ (and their standard errors) are negligible. The differences in the estimates of θ_2 associated with different values of ρ are larger, but still not substantial. This illustrates that the results of the meta-analysis is relatively (but not entirely) insensitive to small differences in ρ.

13.4 ROBUST VARIANCE ESTIMATION

All of the estimates of regression coefficients (including maximum likelihood estimates) are equivalent to weighted least squares with some weight matrix. Usually the weight matrix is derived to obtain maximum efficiency and involves variance or covariance component estimates. In this section, we describe a method of estimation of the variances of weighted least squares estimates that is valid when the number of studies is large, regardless of the within-study covariance structure or the sampling distribution of the effect-size estimates (see chapter 12).

This approach is appealing in the multivariate situation for four reasons. First, it does not require knowledge of the correlation structure of the estimation errors or the random effects. Second, it makes no assumptions about the (conditional or unconditional) distribution of the effect-size estimates, so it is robust to violations of assumptions that estimates or random effects have any specific distribution. Third, it does not require that the study-level covariates be fixed as in other meta-regression models. Fourth, robust variance computations can be used even when variance estimates for individual effects-size estimates are not available (such as when reporting of statistics in studies is incomplete).

The approach, as noted in chapter 12, has two disadvantages. One, it does not offer a way to compute weights to increase efficiency. However, if the variance of each effect-size estimate is known, then standard random-effects procedures can be used, and the robust variance estimates can be used in conjunction with these weights. A second disadvantage is that the formal derivation of the validity of the method requires more studies. However, simulations suggest that the method performs reasonably well for a single covariate even when the number of studies is as small as ten to twenty, and modifications of the method described here can assure its validity in even smaller numbers of studies (Hedges, Tipton, and Johnson 2010).

Table 13.2 Results of Multivariate Meta-Analyses of SAT Coaching Data

	Coefficients for SATV					Coefficients for SATM				
ρ	b_1	S_1	lcl_1	ucl_1	Z_1	b_2	S_2	lcl_2	ucl_2	Z_2
0.0	0.15	0.053	0.05	0.26	2.93	0.11	0.064	–0.12	0.24	0.04
0.5	0.16	0.052	0.06	0.26	3.03	0.09	0.060	–0.02	0.21	0.98
0.6	0.16	0.051	0.06	0.26	3.08	0.09	0.059	–0.03	0.21	1.22
0.7	0.16	0.051	0.06	0.26	3.13	0.09	0.057	–0.03	0.20	1.49
0.8	0.16	0.050	0.06	0.26	3.04	0.09	0.058	–0.03	0.20	1.84

SOURCE: Author's tabulation.
NOTE: Data are for various values of the correlation ρ between effect sizes on SATV and SATM.

13.4.1 Models and Notation

Using the earlier notation, denote the total residual for the ith study as $\boldsymbol{\eta}_i = \mathbf{A}_i\boldsymbol{\xi}_i + \boldsymbol{\varepsilon}_i$ and $\boldsymbol{\eta} = (\boldsymbol{\eta}_1', \ldots, \boldsymbol{\eta}_k')'$. Here we assume the same model as before, except that we relax the assumption that the effect-size estimates have normal distributions. Then we can write the linear model for the estimates as

$$\begin{pmatrix} \mathbf{T}_1 \\ \vdots \\ \mathbf{T}_k \end{pmatrix} = \begin{pmatrix} \mathbf{A}_1\mathbf{X}_1 \\ \vdots \\ \mathbf{A}_1\mathbf{X}_k \end{pmatrix}\boldsymbol{\beta} + \begin{pmatrix} \boldsymbol{\eta}_1 \\ \vdots \\ \boldsymbol{\eta}_k \end{pmatrix}, \tag{13.17}$$

where the $\boldsymbol{\eta}_i$ are independent but $(\boldsymbol{\eta}_1', \ldots, \boldsymbol{\eta}_k')'$ has an unknown covariance matrix.

13.4.2 Robust Variance Estimator

We can compute the weighted least squares estimate of $\boldsymbol{\beta}$ with any weight matrix. In meta-analysis, we usually use weights that are selected to yield the most efficient estimate of $\boldsymbol{\beta}$. We consider the problem of selecting weights later in this chapter. For now, assume an arbitrary weight matrix $\mathbf{W}$. This need not be diagonal, but here we assume $\mathbf{W} = \text{diag}(\mathbf{W}_1, \ldots, \mathbf{W}_k)$, where $\mathbf{W}_i = \text{diag}(w_{i1}, \ldots, w_{ip_i})$ is the $p_i \times p_i$ weight matrix for the ith study. The weighted least squares estimate of $\boldsymbol{\beta}$ is

$$\mathbf{b} = \left(\sum_{i=1}^{k} \mathbf{X}_i'\mathbf{A}_i'\mathbf{W}_i\mathbf{A}_i\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{k} \mathbf{X}_i'\mathbf{A}_i'\mathbf{W}_i\mathbf{T}_i\right). \tag{13.18}$$

The exact variance of $\mathbf{b}$ is

$$\left(\sum_{i=1}^{k} \mathbf{X}_i'\mathbf{A}_i'\mathbf{W}_i\mathbf{A}_i\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{k} \mathbf{X}_i'\mathbf{A}_i'\mathbf{W}_i\left(\boldsymbol{\Sigma}_i + \mathbf{A}_i\boldsymbol{\Psi}\mathbf{A}_i'\right)\mathbf{W}_i\mathbf{A}_i\mathbf{X}_i\right)\left(\sum_{i=1}^{k} \mathbf{X}_i'\mathbf{A}_i'\mathbf{W}_i\mathbf{A}_i\mathbf{X}_i\right)^{-1},$$

where $\boldsymbol{\Psi}_i$ is the covariance matrix of $\boldsymbol{\eta}_i$.

Robust standard errors are obtained by substituting the matrix of cross products of within-study residuals in the jth study for $(\boldsymbol{\Sigma}_j + \mathbf{A}_i\boldsymbol{\Psi}\mathbf{A}_i')$, that is

$$\mathbf{V}^R = \left(\frac{k}{k-q}\right)\left(\sum_{i=1}^{k} \mathbf{X}_i'\mathbf{A}_i'\mathbf{W}_i\mathbf{A}_i\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{k} \mathbf{X}_i'\mathbf{A}_i'\mathbf{W}_i\mathbf{e}_i\mathbf{e}_i'\mathbf{W}_i\mathbf{A}_i\mathbf{X}_i\right)\left(\sum_{i=1}^{k} \mathbf{X}_i'\mathbf{A}_i'\mathbf{W}_i\mathbf{A}_i\mathbf{X}_i\right)^{-1}, \tag{13.19}$$

where $\mathbf{e}_i = \mathbf{T}_i - \mathbf{X}_i\mathbf{b}$ is the $(p_i \times 1)$ residual vector in the ith study. Although the $p_i \times p_i$ matrix $\mathbf{e}_i\mathbf{e}_i'$ is a rather poor estimate of $(\boldsymbol{\Sigma}_i + \mathbf{A}_i\boldsymbol{\Psi}\mathbf{A}_i')$, it is good enough that (13.19) converges in probability to the correct value as $k \to \infty$. Even when the predictor values and weights are random, this estimate converges almost surely to $\boldsymbol{\beta}$ as $k \to \infty$ (see Hedges, Tipton, and Johnson 2010).

Despite related simulations, as discussed in chapter 12, it is difficult to know when the number of studies is large enough to support valid inferences when the number of studies is not large (see Hedges, Tipton, and Johnson 2010). Two modifications can improve performance in small samples of studies (Tipton 2015). The first is a modification of the robust variance estimate itself. The second is a computation of the effective degrees of freedom for the standard error to be used in place of $(k - p)$ for determining critical values.

The modified robust variance estimate is

$$\mathbf{V}^{RM} = \left(\frac{k}{k-q}\right)\mathbf{M}^{-1}\left(\sum_{i=1}^{k} \mathbf{X}_i'\mathbf{A}_i'\mathbf{W}_i\mathbf{B}_i\mathbf{e}_i\mathbf{e}_i'\mathbf{B}_i\mathbf{W}_i\mathbf{A}_i\mathbf{X}_i\right)\mathbf{M}^{-1}, \tag{13.20}$$

where

$$\mathbf{M} = \sum_{i=1}^{k} \mathbf{X}_i'\mathbf{A}_i'\mathbf{W}_i\mathbf{A}_i\mathbf{X}_i,$$

$$\mathbf{B}_i = \left(\mathbf{I}_i - \mathbf{A}_i\mathbf{X}_i\mathbf{M}\mathbf{X}_i'\mathbf{A}_i'\mathbf{W}_i\right)^{-1/2},$$

and $\mathbf{I}_i$ is a $p_i \times p_i$ identity matrix.

The effective degrees of freedom can be different for each regression coefficient. The effective degrees of freedom for the jth regression coefficient $(1 \le j \le q)$ are given by

$$v_j = \frac{\left[tr\left(\mathbf{Q}_j\mathbf{W}^{-1}\right)\right]^2}{tr\left(\mathbf{Q}_j\mathbf{W}^{-1}\mathbf{Q}_j\mathbf{W}^{-1}\right)}, \tag{13.21}$$

where the $k \times k$ matrix $\mathbf{Q}_j$ is given by

$$\mathbf{Q}_j = \sum_{i=1}^{k} (\mathbf{I} - \mathbf{H})_i' \mathbf{B}_i \mathbf{W}_i \mathbf{A}_i \mathbf{X}_i \mathbf{M}^{-1} l_j l_j' \mathbf{M}^{-1} \mathbf{X}_i' \mathbf{A}_i' \mathbf{W}_i \mathbf{B}_i (\mathbf{I} - \mathbf{H})_i . \tag{13.22}$$

l_j is a $q \times 1$ vector whose jth row is unity and all other rows are zero, and $(\mathbf{I} - \mathbf{H})_i$ denotes the p_i rows of the $p \times p$ matrix

$$\mathbf{H} = \mathbf{I} - \mathbf{AX}(\mathbf{X'A'WAX})^{-1}\mathbf{X'A'W}$$

associated with study i.

The effective degrees of freedom can be substantially smaller than $k - q$, which will have a substantial impact on the critical values. The effective degrees of freedom will typically be fractional, not integers, so care must be taken when using some computer programs (such as Microsoft Excel), which round or truncate fractional degrees of freedom to integers and thus would give inaccurate computations. When the effective degrees of freedom are smaller than four, substantial caution is warranted in drawing inferences using the robust standard errors.

13.4.2.1 Tests and Confidence Intervals The square roots of the diagonal elements of $\mathbf{V}^{RM}$ are the robust standard errors of the elements of $\mathbf{b}$. The robust test of the hypothesis

$$\mathrm{H}_0: \beta_j = 0$$

uses the test statistic

$$t_j^{RM} = b_j \Big/ \sqrt{v_{jj}^{RM}} \tag{13.23}$$

where v_{jj}^{RM} is the jth diagonal element of $\mathbf{V}^{RM}$, which is compared with critical values of student's t-distribution with ν_j degrees of freedom. The maximum degrees of freedom is the number of clusters (not effect-size estimates) minus the number of coefficients (including the intercept) in the regression model. A robust $100(1 - \alpha)$ percent confidence interval for β_j is given by

$$b_j - C_{\alpha/2}\sqrt{v_{jj}^{RM}} \leq \beta_j \leq b_j + C_{\alpha/2}\sqrt{v_{jj}^{RM}}, \tag{13.24}$$

where C_α is the $100(1 - \alpha)$ percentage point of t-distribution with ν_j degrees of freedom.

13.4.2.2 Weighting and Robust Estimates Weighting is used in meta-analysis for two purposes. The first and most important is to increase efficiency of estimates. Because estimates from different studies typically have very different precision, we give more weight to studies whose estimates have greater precision. For example, inverse variance weights are selected to obtain the most efficient weighted mean (or weighted regression coefficient estimates). The second, and incidental, function of weights in meta-analysis is to compute the variances of the combined estimate. When inverse variance weights are used in either fixed- or random-effects models, the variance of the weighted mean is the reciprocal of the sum of the weights. Similarly, the inverse of the weighted sum of squares and cross products matrix, the $(\mathbf{X'WX})^{-1}$ matrix, is the covariance matrix of the regression coefficients.

The robust estimate of the covariance matrix of the regression coefficients obviates the second function of weights altogether. However, although it provides valid variance estimates for any weights, it provides no obvious method for selecting weights and particularly selecting weights that will lead to efficient estimates.

Because weights proportional to the inverse covariance matrix produce combined estimates with the smallest variance, it would often make sense to use them if possible. Unfortunately, computing inverse covariance weights requires knowing the covariance matrix of the estimates, which depends in part on the covariance structure of the correlated estimates.

That this covariance structure is rarely known is exactly the motivation for the robust variance estimates. We suggest approximating the optimal weights in a way that does not involve the within-study covariances (except when they are used to estimate between-study variance components). In deriving approximations, it is useful to remember that any weights that are reasonably close to optimal will often give estimates that are nearly as efficient as the optimal weights, and the robust variance estimates will be valid. However, in some situations, this weighting will depart from optimal weighting. For example, if the effect sizes on different outcomes have substantially different between-study variance components, this weighting does not allow differential weighting of the different outcomes and thus will depart from optimal weighting.

The variance of all effect-size estimates depends heavily on within-study sample size. Sample sizes typically vary substantially across studies. Correlated effects within a study arising from multiple measurements on the same individuals are likely to be based on similar (if not identical) sample sizes. It follows that much more variation is likely in the precision of effect-size estimates between

studies than within studies. Thus a sensible approach might be to give equal weights to all of the correlated estimates within a study. However, it does not make sense to assign each estimate in study i a weight proportional to the inverse of the total weight of that study, because that strategy would assign more total weight to studies having more estimates. This could depart substantially from optimal weighting if some studies have a large number of highly correlated estimates.

It is more sensible to bound the total weight that any single study can receive. Suppose that each study has p effect-size estimates with equal variances v and equal correlations ρ between estimates. If we were estimating the mean effect in this situation, it would be optimal to assign a weight of $1/v[1 + (p - 1)\rho]$ to each effect in the study. This optimal weight tends to $1/vp = (1/v)/p$ (the average of the inverse variance weights in the study) as ρ tends to 1.

The correlation ρ involved here is the unconditional correlation, which includes correlation induced by the (study-level) random effects, which is likely to be larger than the correlation induced by correlated errors of estimation alone. If the variance of the estimation errors is v and the correlation between estimates induced by estimation error (the correlation of estimates T_{is} and T_{it} $(s \neq t)$ conditional on the study-level parameters θ_{is} and θ_{it}) is ρ_C and the variance of the study-level effects (the between-studies variance component) is τ^2, the unconditional correlation between estimates ρ_U is

$$\rho_U = \left(v\rho_c + \tau^2\right)/\left(v + \tau^2\right) = \rho_c + (1 - \rho_c)\left[\tau^2/(v + \tau^2)\right]$$
$$= \rho_c + (1 - \rho_c)I^2,$$

where here I^2 is the index of heterogeneity defined in chapter 11, but here without multiplying by 100 percent. This makes clear that ρ_U is an increasing function of τ^2 that tends to 1 as τ^2 becomes large. Thus (under this model) when the between-study variance $\tau^2 > 0$, $\rho_U > \rho_C$ (and can be much larger), so that the optimal weights will be closer to $1/vp$ than might be imagined considering ρ_C alone.

One implication of this special case is that, if estimates are reasonably highly correlated, little efficiency will typically be lost by bounding the total weight for each cluster by $1/\bar{v}$, where $\bar{v}$ is the average variance of the estimates in that cluster. This would amount to assigning each estimate a weight of

$$w_{ij} = \frac{1}{p_i \bar{v}_{i\bullet}} = 1 \Big/ \sum_{j=1}^{p_i} v_{ij}, \qquad (13.25)$$

where $\bar{v}_{i\bullet}$ is the average of the variances in study i. If the value of ρ were known or could be imputed, using a weight of $1/\bar{v}_{i\bullet}[1 + (p_i - 1)\rho]$ could lead to more efficient estimates (at least for the mean effect size).

This strategy would be sufficient if we wished only to develop fixed-effects weights that ignored between-study variation. Such weights will lead to (approximately) efficient estimates only if there is no between-study variation. In many cases, we will wish to develop random-effects weights (that include between-study variation), that will lead to (approximately) efficient estimates when there is between-study variation. To do so, it will be necessary to estimate the between-study variance component or components, which we address in the following two sections.

13.4.2.3 Estimates of Variance Components for Weighting One approach to estimation of between-study variance components for computation of efficient weights is to impute values for a convenient choice of covariance structure (a working covariance matrix) and compute optimal weights given between-studies variance components estimated for that structure. It is important that the accuracy of the weights as approximations of the optimal weights has no impact on the accuracy of the robust standard errors. They will be accurate for any choice of weights. The robust standard errors will just be smallest for the most efficient weights.

A reasonable choice for a working covariance matrix is one in which the estimates in each study have the same estimation error variance equal to the average ($\bar{v}_{i\bullet}$), the fixed-effects weights given to each effect-size estimate in the ith study are equal to $1/p_i$ times the reciprocal of that variance (so that the total weight for the ith study is $1/\bar{v}_{i\bullet}$), and the correlation of any pair of effect-size estimates in the same study is ρ. Thus we posit the linear model and error covariance structure (13.5), except that now each of the study-specific covariance matrices can be described in terms of a between-study variance component τ^2 and a between-effect-within-study correlation ρ. Specifically,

$$\left(\tau^2 + \rho\bar{v}_i/p_i\right)\mathbf{J}_i + \left[(1 - \rho)\bar{v}_i/p_i\right]\mathbf{I}_i,$$

where $\mathbf{I}_j$ is a $p_j \times p_j$ identity matrix, $\mathbf{J}_j$ is a $p_j \times p_j$ matrix of 1's.

In this case, the estimate of τ^2 can be computed from a preliminary meta-regression using all of the effect-size estimates, giving equal weights to each effect-size estimate from the same study, so that the weight given to each effect-size estimate in the ith study is $w_i = 1/p_i\bar{v}_{i\bullet}$. In this case, the weighted residual sum of squares from the meta-regression Q_E is given by

$$Q_E = \sum_{i=1}^{k} \mathbf{T}_i'\mathbf{W}_i\mathbf{T}_i - \left(\sum_{i=1}^{k} \mathbf{T}_i'\mathbf{W}_i\mathbf{A}_i\mathbf{X}_i\right)\left(\sum_{i=1}^{k} \mathbf{X}_i'\mathbf{A}_i'\mathbf{W}_i\mathbf{A}_i\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{k} \mathbf{X}_i'\mathbf{A}_i'\mathbf{W}_i\mathbf{T}_i\right) \tag{13.26}$$

and the residual variance component estimate is

$$\hat{\tau}^2 = \frac{Q_E - k + \text{tr}\left[\mathbf{R}\left(\sum_{i=1}^{k} \frac{w_i}{p_i}\mathbf{X}_i'\mathbf{X}_i\right)\right] + \rho\,\text{tr}\left[\mathbf{R}\left(\sum_{i=1}^{k} \frac{w_i}{p_i}\mathbf{X}_i'\mathbf{J}_i\mathbf{X}_i - \mathbf{X}_i'\mathbf{X}_i\right)\right]}{\sum_{i=1}^{k} p_i w_i - \text{tr}\left[\mathbf{R}\left(\sum_{i=1}^{k} w_i^2\mathbf{X}_i'\mathbf{J}_i\mathbf{X}_i\right)\right]}, \tag{13.27}$$

where $\mathbf{R}$ is the inverse of the $\mathbf{X}'\mathbf{W}\mathbf{X}$ matrix given by

$$\mathbf{R} = \left(\sum_{i=1}^{k} w_i\mathbf{X}_i'\mathbf{X}_i\right)^{-1},$$

$\mathbf{J}_i = \mathbf{1}\mathbf{1}'$ is a $m \times m$ matrix of 1's, tr($\mathbf{X}$) is the trace of the square matrix $\mathbf{X}$, and negative values of the estimate are set to zero. These computations are somewhat simplified by recognizing that $\mathbf{X}_i'\mathbf{X}_i = \mathbf{X}_i'\mathbf{A}_i'\mathbf{A}_i\mathbf{X}_i$ and $\mathbf{X}_i'\mathbf{J}_i\mathbf{X}_i = \mathbf{X}_i'\mathbf{Y}_i'\tilde{\mathbf{J}}_i\mathbf{Y}_i\mathbf{X}_i$, where $\tilde{\mathbf{J}}_i$ is a $p_i \times p_i$ matrix of ones.

It is wise to then conduct sensitivity analyses to ensure that estimates are not highly sensitive to choice of that covariance structure. However, the effect of the correlation among estimates on the estimate of τ^2 given in formula (13.27) occurs entirely through the term

$$\rho\left[\frac{\text{tr}\left[\mathbf{R}\left(\sum_{i=1}^{k} \frac{w_i}{p_i}[\mathbf{X}_i'\mathbf{J}_i\mathbf{X}_i - \mathbf{X}_i'\mathbf{X}_i]\right)\right]}{\sum_{i=1}^{k} p_i w_i - \text{tr}\left[\mathbf{R}\left(\sum_{i=1}^{k} w_i^2\mathbf{X}_i'\mathbf{J}_i\mathbf{X}_i\right)\right]}\right], \tag{13.28}$$

which is typically small (see Hedges, Tipton, and Johnson 2010). Moreover, in simulation studies, the results of multivariate meta-regressions proved relatively insensitive to incorrect values of the within-study correlations (Ishak et al. 2008). Our experience is also that these estimates are not highly sensitive to choices of working covariance structure in many situations, but it is clear that estimates of between-study variance components will be most sensitive to within-study correlation structure when between-study covariance components are small compared to within-study covariances (see Riley 2009).

Example. Return to the example of the $k = 26$ randomized experiments on SAT coaching (Kalaian and Raudenbush 1996). Five of these studies reported outcomes for both SAT verbal scores and SAT mathematics scores, eight studies reported results only for SATM, and fifteen studies reported results only for SATV. The data from these studies is reproduced in table 13.1. Suppose we were interested in estimating the mean effect size for SATV and SATM, then $m = 2$, $\boldsymbol{\theta} = (\theta_1, \theta_2)'$, where θ_1 = the effect size for SATV and θ_2 is the mean effect size for SATM. Define the selection matrix $\mathbf{A}_i$ so that it selects the effect size for θ_1 if the effect size is estimating SATV, and θ_2 is the effect size is estimating SATM. Thus the selection matrix A_i for studies one to thirteen is

$$\mathbf{A}_i = (1 \quad 0), \tag{13.29}$$

the selection matrix Y_i for studies 14 to 21 is

$$\mathbf{A}_i = (0 \quad 1), \tag{13.30}$$

and the selection matrix $\mathbf{A}_i$ for studies 22 to 26 is

$$\mathbf{A}_i = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{13.31}$$

The overall $\mathbf{A}$ matrix therefore consists of thirteen stacked 1×2 $\mathbf{A}_i$ matrices of form (13.29), stacked on top of eight stacked 1×2 $\mathbf{A}_i$ matrices of form (13.30), stacked on top of five stacked 2×2 $\mathbf{A}_i$ matrices of form (13.31). The design matrices $\mathbf{X}_i$ for all of the studies are 2×2 matrices and the overall design matrix is a matrix of $k = 26$ stacked 2×2 matrices.

With this matrix, the intercept β_1 is interpreted as the effect size on SATV and the slope β_2 is interpreted as the

effect size for SATM. The object of the analysis is to estimate the slope and intercept (the effect size of coaching on SATV and SATM), develop confidence intervals for each and test the hypothesis that the coaching effect is different from zero.

We began by estimating τ^2 to estimate efficient weights for the analysis. Starting with the assumption that the correlation between effect sizes on SATV and SATM were approximately the same as the correlations between measurements of SATV and SATM, we estimated weights assuming a working correlation of $\rho = 0.7$. This yielded an estimate of τ^2 equal to zero. To test the sensitivity of this value, we repeated the computation using working correlation values between $\rho = 0$ and $\rho = 0.95$, inclusive and obtained the same result.

Using weights computed with $\tau^2 = 0$, we obtained the estimate $b_1 = 0.140$, with a robust standard error of 0.0338 and 13.1 degrees of freedom. The 95 percent critical value of the t-distribution with 13.1 degrees of freedom is 2.159. Thus a 95 percent confidence interval for the intercept β_1 is given by

$$\begin{aligned} 0.066 &= 0.140 - 2.159 \times 0.0339 \\ &\leq \beta_1 \leq 0.140 + 2.159 \times 0.0339 \\ &= 0.213. \end{aligned}$$

The robust significance test for β_1 uses the statistic

$$t_1 = 0.140/0.0338 = 4.12,$$

which exceeds the critical value 2.159, so we can reject the hypothesis that $\beta_1 = 0$. Similarly, we obtained the estimate $b_2 = 0.114$, with a robust standard error of 0.0489 and 7.69 degrees of freedom. The 95 percent critical value of the t-distribution with 7.69 degrees of freedom is 2.322. Thus a 95 percent confidence interval for the intercept is given by

$$\begin{aligned} -0.000 &= 0.114 - 2.322 \times 0.0489 \\ &\leq \beta_2 \leq 0.114 + 2.322 \times 0.0489 \\ &= 0.227. \end{aligned}$$

The robust significance test for β_2 uses the statistic

$$Z_2 = |0.114|/0.0489 = 2.32,$$

which does not exceed the critical value 2.322, so we cannot reject the hypothesis that $\beta_2 = 0$.

These results differ slightly from those of the full multivariate analysis, but their qualitative conclusions are the same. They differ slightly for two reasons. First, a slightly different (and marginally less efficient) weighting scheme is used here. Second, the full multivariate analysis uses information about the within-study covariance matrix from each study that the robust variance estimates do not depend upon. In general, the results will usually be quite similar.

These analyses could be computed directly with a package that permits manipulation of matrices, such as MATLAB, R, or the matrix languages in statistical packages such as SAS, SPSS, or Stata. The formulas have been preprogrammed in several packages that compute the robust variance estimates such as the R package *robumeta* or the Stata macro *robumeta*.

13.5 ELIMINATING DEPENDENCE

Perhaps the most frequently used approach to handling statistical dependence among effect-size estimates from the same study is to compute a single summary of those estimates. I call such estimates synthetic effect-size estimates. Because there is only one synthetic effect-size estimate for each study, and because studies are usually taken to be independent, synthetic effect sizes can be analyzed by conventional (univariate) meta-analysis methods.

13.5.1 Model and Notation

Suppose that the ith study has p_i effect-size estimates $(T_1, \ldots, T_{p_i})$, which are estimates of the effect-size parameters $(\theta_1, \ldots, \theta_{p_i})$, and the estimation error variance of the T_{ij} is denoted by v_{ij}. The reviewer may wish to use information from all of the effect estimates, but there may be no information about the covariances of the T_{ij}, so full multivariate methods cannot be used.

13.5.2 Estimating Mean Effect Size

If the object of the meta-analysis is to estimate the average of the effect sizes across studies, then the reviewer might wish to create a synthetic effect-size estimate for the ith study by combining the T_{ij} into single estimate. Because different estimates within the same study are likely to be based on similar sample sizes and therefore

have similar variances, it may be sensible to take the unweighted average

$$\overline{T}_i = \frac{1}{p_i}\sum_{j=1}^{p_i} T_{ij}. \tag{13.32}$$

Then the synthetic (average) estimate can be used in a conventional meta-analysis just like any other effect-size estimate.

One difficulty is that the variance of $\overline{T}_i$ depends on the covariance structure and is therefore not known precisely. The usual approach in this case is to use a conservative overestimate of the variance of $\overline{T}_i$ as an approximate variance estimate. A reasonable approach is to use the average of the variances v_{ij} of the T_{ij} within-study i as a synthetic variance, that is,

$$\overline{v}_{i\bullet} = \frac{1}{p_i}\sum_{j=1}^{p_i} v_{ij}. \tag{13.33}$$

If the effect-size estimates are positively correlated, $\overline{v}_{i\bullet}$ will be larger than the true variance of $\overline{T}_i$ so that the variance of the weighted mean in the meta-analysis will be overestimated. Confidence intervals will therefore be wider than the exact confidence intervals, and statistical significance will be understated (p-values will be too large) in the overall meta-analysis.

It is probably useful to realize that although $\overline{v}_{i\bullet}$ overestimates the true variance of $\overline{T}_i$, the overestimate may not be substantial. Consider a case in which all of the estimates in the ith study have the same variance v_i and any two of the estimates have the same positive correlation ρ. This may be an idealization, but it is an idealization of exactly the case where one might wish to combine effect-size estimates within a study: Each outcome measure is a measure of the same construct and all of the measures are equally valid. In such a case, the actual variance of $\overline{T}_i$ is

$$\overline{v}_i\left[\frac{1+(p_i-1)\rho}{p_i}\right] = \overline{v}_i\left[\rho+(1-\rho)\left(\frac{1}{p_i}\right)\right]. \tag{13.34}$$

If $\rho = 0$, then the actual variance of $\overline{T}_i$ is $\overline{v}_{i\bullet}/p_i$ which may be very different from $\overline{v}_{i\bullet}$. However, the situations in which it makes sense to combine information across different effect sizes about the same construct are the situations in which the intercorrelation ρ is relatively large (effects that are uncorrelated presumably are not measuring the same construct). Thus it seems implausible, on substantive grounds, that a reviewer would choose to combine effect-size estimates where ρ was small or negative. If ρ is near 1, equation (13.34) implies that the exact variance of $\overline{T}_i$ is close to $\overline{v}_{i\bullet}$. For example, if $\rho = 0.8$ (a reasonable value for different cognitive tests of the same construct) and $p_i = 2$, then the true value of the variance of $\overline{T}_i$ is $0.9\overline{v}_{i\bullet}$ and even if $p_i = 10$, the true value of the variance of $\overline{T}_i$ is $0.82\overline{v}_{i\bullet}$.

The weighted mean effect size will still be unbiased, only its uncertainty (as expressed in the variance, confidence interval width, and p-values) will be affected. The amount by which the overall meta-analysis is affected depends on the amount of correlated effect-size data and the extent of the correlation. Clearly, if only a few studies have any correlated estimates (and therefore synthetic effect-size estimates with synthetic overestimates of their variance), the impact will be quite small. If there are many studies, each with many correlated estimates (and therefore synthetic effect-size estimates with synthetic overestimates of their variance), the impact can be substantial (for an extensive study of the impact of using synthetic effect-size estimates on inference in meta-analysis, see Hedges 2007).

The use of synthetic effect sizes like $\overline{T}_i$ with synthetic variances like $\overline{v}_{i\bullet}$ also has an impact on statistics like tests of heterogeneity. It can be shown that if the correlations among effect-size estimates are positive, the test of heterogeneity using the Q statistics given in chapter 12 rejects the hypothesis of homogeneity less often than nominal. That is, the actual p-value of the Q test is smaller than nominal (for details, see Hedges 2007).

Example. Return to the example of the $k = 26$ randomized experiments on SAT coaching (Kalaian and Raudenbush 1996; for the data, see table 13.1). Five of these studies reported outcomes for both SAT verbal scores and SAT mathematics scores, eight studies reported results only for SATM, and thirteen reported results only for SATV. The vast majority of the studies (twenty-one of twenty-six) estimated only one effect size. We might choose to deal with the five studies that estimated effects on both SATV and SATM by creating a synthetic effect size (the average of the two effects) and a synthetic variance (the average of the variances) for each of these five studies. The three synthetic effects are given in the last column of table 13.1.

Computing first the test of heterogeneity we see that the Q statistic is $Q = 13.560$, and comparing this with the critical values of the chi-squared distribution with 25 degrees of freedom, we see that this value is not statistically significant at the 5 percent level. In fact, values this large would be expected to occur almost 97 percent of the time by chance if there were perfect homogeneity

of effects and the effects were all independent. The effects are not all independent, however, because three studies produce two effects each, effects that are probably correlated. The actual p-value associated with this significance test is probably smaller than 0.97. Because $Q = 13.53$ is less than its nominal degrees of freedom ($k - 1 = 25$), the method of moments estimator of τ^2 is zero.

Because the estimate of τ^2 is 0, the fixed- or random-effects weighted averages are identical. Here $\overline{T}_\bullet = 0.130$ and $v_\bullet = 0.0023$ for a standard error of 0.048. The 95 percent confidence interval computed for θ is

$$0.035 = 0.130 - 1.96 \times 0.048 \leq \theta \leq 0.130 + 1.96 \times 0.048 = 0.224.$$

The test statistic for testing the hypothesis that $\theta = 0$ is given by

$$Z = 0.130/0.048 = 2.569.$$

Comparing 2.569 with 1.96, the 95 percent critical value of the standard normal distribution, we can reject the hypothesis that $\theta = 0$.

13.6 CONCLUSION

Conventional methods for meta-analysis assume that the effect-size estimates are statistically independent. To obtain valid analyses of effect sizes that are statistically dependent, special methods are needed. Full multivariate methods are an elegant approach to analysis of dependent effect-size data, but these methods require extensive data on correlations among effect-size estimates that are frequently unavailable. Assuming simplified, but approximate, correlation structure among the estimates within studies is simpler and can provide reasonably accurate approximate analyses if it is based on sensible empirical evidence about correlations. Robust variance estimation methods are good practical alternatives that provide valid analyses without assumptions about the form of the correlation structure among dependent effect-size estimates. The increasing availability of software for robust analyses should help make these methods more accessible to users.

13.7 REFERENCES

Begos, Pantelis G. 2015. "Meta-Analysis in Stata Using Gllamm." *Research Synthesis Methods* 6(4): 310–32.

Gleser, Leon J., and Ingram Olkin. 2000. "Meta-Analysis of 2 × 2 Tables with Multiple Treatment Groups." In *Meta-Analysis in Medicine and Health Policy*, edited by Don Berry and Dalene Strangl. New York: Marcel Dekker.

———. 2009. "Stochastoically Dependent Effect Sizes." In *The Handbook of Research Synthesis and Meta-Analysis*, 2nd ed., edited by Harris Cooper, Larry V. Hedges, and Jeffrey C. Valentine. New York: The Russell Sage Foundation.

Hedges, Larry V. 2007. "Meta-Analysis." In *The Handbook of Statistics*, vol. 26, edited by C. Radhakrishna Rao and Sandip Sinharay. Amsterdam: Elsevier.

Hedges, Larry V., and Ingram Olkin. 1985. *Statistical Methods for Meta-Analysis*. Orlando, Fl.: Academic Press.

Hedges, Larry V., Elizabeth Tipton, and Matt Johnson. 2010. "Robust Variance Estimation for Meta-Regression with Dependent Effect Size Estimators." *Journal of Research Synthesis Methods* 1(1): 39–65.

Hotelling, Harold. 1940. "The Selection of Variates for Use in Prediction with Some Comments on the General Problem of Nuisance Parameters." *Annals of Mathematical Statistics* 11(3): 271–83.

Ishak, K. Jack, Robert W. Platt, Lawrence Joseph, and James A. Hanley. 2008. "Impact of Approximating or Ignoring Within-Study Covariances in Multivariate Meta-Analyses." *Statistics in Medicine* 27(5): 670–86.

Kalaian, Sema A., and Stephen W. Raudenbush. 1996. "A Multivariate Mixed Linear Model for Meta-Analysis." *Psychological Methods* 1(3): 227–35.

Konstantopoulos, Spyros. 2011. "Fixed Effects and Variance Components Estimation in Three Level Meta-Analysis." *Research Synthesis Methods* 2(1): 61–76.

Neil, J. J., and Olive J. Dunn. 1975. "Testing Equality of Dependent Correlation Coefficients." *Biometrics* 31: 531–43.

Raudenbush, Stephen W., Betsy J. Becker, and Hripsime A. Kalaian. 1988. "Modeling Multivariate Effect Sizes." *Psychological Bulletin* 103(11): 111–20.

Riley, Richard D. 2009. "Multivariate Meta-Analysis: The Effect of Ignoring Within-Study Correlation." *Journal of the Royal Statistical Society, Series A* 172(4): 789–811.

Stevens, John R., and Alan M. Taylor. 2009. "Hierarchical Dependence in Meta-Analysis." *Journal of Educational and Behavioral Statistics* 34(1): 46–73.

Thum, Yeow Meng. 1997. "Hierarchical Linear Models for Multivariate Outcomes." *Journal of Educational and Behavioral Statistics* 22(1): 77–108.

Tipton, Elizabeth. 2015. "Small Sample Adjustments for Robust Variance Estimation with Meta-Regression." *Psychological Methods* 20(3): 375–93.

Weu, Yinghui, and Julian P. T. Higgins. 2011. "Bayesian Multivariate Meta-Analysis with Multiple Outcomes." *Statistics in Medicine* 32(17): 2911–34.

14

BAYESIAN META-ANALYSIS

REBECCA M. TURNER
University College London, University of Cambridge

JULIAN P. T. HIGGINS
University of Bristol

CONTENTS

14.1 INTRODUCTION

The statistical methods for meta-analysis presented in previous chapters are based on a frequentist (or classical) approach to statistical inference. In a frequentist approach to inference, unknown parameters are viewed as fixed quantities, which can be estimated from observed data, subject to the uncertainty that results from sampling error. In a Bayesian approach, unknown parameters are viewed as random variables, with associated probability distributions that represent beliefs about the plausibility of different parameter values. A Bayesian analysis requires the analyst to provide an initial probability distribution, known as the prior distribution, which expresses their beliefs about the plausibility of different parameter values before making use of the evidence provided by the data. Evidence from the data is then used to update the prior distribution, through Bayes's theorem, and to obtain the posterior distribution on which Bayesian inference is based.

Over recent years, Bayesian approaches to meta-analysis and evidence synthesis have increased in popularity. A Bayesian meta-analysis allows direct probability statements to be made regarding parameters of interest and enables prediction of effects in future studies; these advantages are particularly valuable in meta-analyses performed to inform decision making. By choosing informative prior distributions for particular parameters, analysts can combine evidence from the observed data brought together specifically for the meta-analysis with external evidence from other sources, in order that statistical inference is based on all available evidence. A practical advantage is that Markov chain Monte Carlo (MCMC) simulation-based methods facilitate Bayesian estimation of complex models, meaning that a Bayesian approach can offer greater flexibility in modeling than a frequentist approach.

In the following section, we provide a brief introduction to Bayesian inference.

14.2 BAYESIAN INFERENCE

Bayesian inference is based on the posterior distribution of the parameters of interest, which combines evidence from the data with existing prior beliefs. Under a Bayesian approach, we begin by specifying a joint probability distribution, $f(\boldsymbol{\theta})$, which describes our prior beliefs about the values of all unknown model parameters $\boldsymbol{\theta}$. The information provided by the data $\mathbf{y}$ is represented by the likelihood function $f(\mathbf{y}|\boldsymbol{\theta})$. The likelihood function is the same one that is used in frequentist inference, and represents the statistical model, that is the way in which the data are thought to have arisen, conditional on the values of the parameters $\boldsymbol{\theta}$. In Bayesian inference, the prior distribution is updated by evidence from the data, and a posterior distribution $f(\boldsymbol{\theta}|\mathbf{y})$ is obtained. Bayes's theorem describes the relationship between the posterior distribution, the likelihood function, and the prior distribution:

$$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\mathbf{y})}$$

where $f(\mathbf{y})$ is a constant ensuring that the distribution $f(\boldsymbol{\theta}|\mathbf{y})$ integrates to 1.

In most analyses, it is not possible to express the posterior distribution in closed form and Bayesian inference requires evaluation of complicated integrals. Since the 1990s, however, the availability of powerful MCMC simulation methods (Smith and Roberts 1993; Best et al. 1996; Brooks 1998) within accessible software, such as

WinBUGS has facilitated Bayesian analysis (Lunn et al. 2000; see also box 14.1). Bayesian estimation based on MCMC methods provides an extremely flexible framework for data analysis, meaning that some applied problems that cannot be addressed using a frequentist approach can be handled using a Bayesian approach. However, Bayesian analysis requires careful choice of prior distributions, since these can be very influential in small samples. Prior distributions are discussed in more detail later in the chapter.

14.3 GENERIC BAYESIAN MODELS FOR META-ANALYSIS

14.3.1 Fixed-Effect Meta-Analysis Model

In a fixed-effect (or common effect, or equal-effects) meta-analysis, all studies included are assumed to provide estimates of the same underlying effect, and differences among the estimated effect sizes are assumed to result only from sampling error. Suppose that the data to be combined in the meta-analysis comprise k effect size estimates $y_1, \ldots, y_k$. We assume a common underlying effect θ in all studies. The Bayesian fixed-effect meta-analysis model is

$$\begin{aligned} y_i &\sim Normal\left(\theta, \sigma_i^2\right) \qquad i = 1, \ldots, k \\ \theta &\sim P_\theta \end{aligned} \tag{14.1}$$

where σ_i^2 is the variance of the effect estimate in study i and P_θ is the prior distribution for the underlying effect θ. By convention, the within-study variances are assumed to be known and are therefore replaced in equation 14.1 by the estimated within-study variances $\hat{\sigma}_i^2$, since this usually makes little difference in practice unless many of the studies are small (Hardy and Thompson 1996):

$$\begin{aligned} y_i &\sim Normal\left(\theta, \hat{\sigma}_i^2\right) \qquad i = 1, \ldots, k \\ \theta &\sim P_\theta \end{aligned} \tag{14.2}$$

Box 14.1 Markov Chain Monte Carlo Methods

The principle of Markov chain Monte Carlo (MCMC) simulation is that we construct a Markov chain for which the stationary distribution is the target posterior distribution $f(\boldsymbol{\theta}|\mathbf{y})$. A Markov chain is a sequence of random variables satisfying the condition that, conditional on the present value, future values are independent of past values.

Several algorithms are available for constructing Markov chains with a specified stationary distribution. One of the most widely used of these algorithms is Gibbs sampling (Gelfand and Smith 1990). Gibbs sampling works by sampling each parameter in turn, each time conditioning on the most recent values of all other parameters. A set of initial values $\boldsymbol{\theta}^{(0)} = \{\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_r^{(0)}\}$ is chosen at the beginning of the process. A new value for θ_1 is sampled from the full conditional distribution for θ_1 given the current values of the other model parameters $\{\theta_2^{(0)}, \theta_3^{(0)}, \ldots, \theta_r^{(0)}\}$. Next, a new value for θ_2 is sampled given the current values of the other parameters $\{\theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_r^{(0)}\}$ and so on until the last parameter θ_r is sampled given $\{\theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_{r-1}^{(1)}\}$. This process completes one iteration of the Gibbs sampler and a transition from $\boldsymbol{\theta}^{(0)}$ to $\boldsymbol{\theta}^{(1)}$. The process is repeated numerous times and a sequence of samples $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$ is generated.

Eventually, the Markov chain of samples should converge to the target posterior distribution. The initial iterations are referred to as *burn-in*, a phase in which the chain is judged not yet to have reached convergence, and these are discarded. Once the chain has converged, a large number of iterations are carried out to provide a sample from the posterior distribution. Summaries of this sample such as means, medians, and percentiles are used to provide estimates of summary measures for the posterior distribution. Markov chains can sometimes get stuck in a particular area as a result of the initial values chosen and may not converge to the correct stationary distribution. It is therefore preferable to run multiple chains starting from diverse sets of initial values and to use formal methods to check whether they have converged to the same distribution, which may then be assumed to be the target posterior distribution.

SOURCE: Authors' compilation.

When fitting the fixed-effect model (14.2) using a frequentist approach to estimation, the common effect θ is estimated by a weighted average of the studies' effect estimates, with weights set equal to the reciprocals of their within-study variances $\hat{\sigma}_i^2$. When taking a Bayesian approach to estimation, we begin by choosing prior distributions for all unknown parameters in the model. In model (14.2), we need to choose a prior distribution P_θ for the one unknown parameter θ. We discuss specification of this prior distribution in more detail in a later section.

14.3.2 Random-Effects Meta-Analysis Model

In most meta-analyses, differences among the study-effect estimates are known to arise from causes other than sampling error alone. For example, study designs may differ in populations recruited or timing of outcome measurement, differences in the conduct of the research may lead to variation in results, or the studies may be affected by methodological flaws that result in biases. The assumption of a common underlying effect across studies is then unrealistic, and it is inappropriate to combine the results using a fixed-effect meta-analysis model. The random-effects meta-analysis model is a commonly used alternative model, which includes a set of random effects to represent the variability of true effects across studies:

$$\begin{aligned} y_i &\sim Normal\left(\theta_i, \hat{\sigma}_i^2\right) \qquad i = 1, \ldots, k \\ \theta_i &\sim Normal\left(\mu, \tau^2\right) \\ \mu, \tau &\sim P_{\mu,\tau} \end{aligned} \tag{14.3}$$

In this model, the θ_i represent the true effects in each of the k studies, assumed to be normally distributed with mean μ and variance τ^2. The between-study variance τ^2 describes the extent of heterogeneity among the studies. A Bayesian random-effects meta-analysis relies on making an assumption of *exchangeability* for the true study effects, which is a judgment that the effects are similar, but nonidentical, and that the rankings of their magnitudes cannot be predicted in advance (Higgins, Thompson, and Spiegelhalter 2009). Conventionally, the prior distributions for the mean μ and variance τ^2 are assumed to be independent, so we will write $\mu \sim P_\mu$, $\tau \sim P_\tau$.

14.3.3 Random-Effects Meta-Regression Model

Given interest in exploring the influence of study-level characteristics on the magnitude of effect, it may be useful to fit meta-regression models. Bayesian meta-regression can be performed as an extension of the Bayesian meta-analyses discussed in the previous sections. We extend the random-effects model to a random-effects meta-regression model by introducing study-level covariates x_{li}:

$$\begin{aligned} y_i &\sim Normal\left(\theta_i, \hat{\sigma}_i^2\right) \qquad i = 1, \ldots, k \\ \theta_i &\sim Normal\left(\sum_{m=0}^{p} \beta_m x_{mi}, \tau_{res}^2\right) \qquad i = 1, \ldots, k \\ \boldsymbol{\beta} &\sim P_{\boldsymbol{\beta}}, \tau_{res} \sim P_{\tau_{res}} \end{aligned} \tag{14.4}$$

where x_{0i} is usually set equal to 1 so that the model includes an intercept β_0; the remaining x_{mi} are continuous or dichotomous study-level characteristics (possibly including dummy coding of categorical study-level characteristics). In this model, prior distributions are required for the set of meta-regression coefficients $\boldsymbol{\beta}$ and for τ_{res}^2, which represents the residual between-study heterogeneity remaining after accounting for that explained by the study-level covariates.

14.3.4 Prior Distributions

Prior distributions for the model parameters should be chosen carefully in any Bayesian analysis. Several general approaches to specifying prior distributions are available; we introduce these in decreasing order of popularity.

The most common option is to choose a "vague" prior distribution that is intended to provide little information relative to the information provided by the observed data. Vague prior distributions are used when analysts wish to obtain findings that are based on the data alone and close to those that would result from a frequentist analysis. In general, no consensus has been reached on the optimal choice of vague prior distribution for particular model parameters (Kass and Wasserman 1996). Bayesian statisticians have explored formal methods for choosing "reference" or "non-informative" priors, which could provide a default choice of a prior representing ignorance in particular models. However, many priors chosen using such methods have been shown to have undesirable properties and it is now generally accepted that choosing priors which are

approximately "flat" over the range of values that the parameters are likely to take (for example, by specifying very large variances) is a sensible approach (Kass and Wasserman 1996). When sample sizes are large, the likelihood will dominate the posterior distribution, and therefore findings based on different possible vague prior distributions will be very similar. When the data provide little information about particular parameters, prior distributions are more influential and sensitivity of results to the choice of prior distribution should be investigated.

Alternatively, we could choose an informative prior distribution based on external data sources available before the analysis is carried out. It can be quite beneficial to use informative rather than vague prior distributions, particularly for parameters about which the new data provide little or no information. Two standard sources of information that can be represented in a prior distribution are historical (or external) data and expert beliefs. For some parameters, a prior distribution can be constructed from relevant external data, perhaps from similar previous studies. In other cases, analysts may wish to incorporate prior beliefs by eliciting opinion from experts about the likely values of particular parameters.

A third type of prior distribution is one chosen expressly to represent a hypothetical viewpoint. David Spiegelhalter, Lawrence Freedman, and Mahesh Parmar demonstrate construction of a "skeptical" prior distribution that represents an opinion that the treatment difference in a randomized trial is unlikely to exceed a chosen value, and an "enthusiastic" prior distribution that represents an opposing opinion that the treatment difference is likely to favor one intervention over its comparator (1994). Such prior distributions are useful for exploring how much evidence would be required to convince a skeptic that a new treatment is effective, or to convince an enthusiast that it is not. Andrew Gelman and his colleagues suggest using "weakly informative" priors, constructed to include less information than is available from external data, in order that the prior influences the analysis only when the data are sparse (2009).

We briefly mention the concept of *conjugate* priors, leading to a posterior that has the same distributional form as the prior. Conjugate priors can be chosen in simple analyses, such as a single binary endpoint from one study (for which a Beta prior would be the conjugate choice), but are not always available for more complex analyses.

14.3.4.1 Choosing a Prior Distribution for the Overall Effect First, we consider how to choose a prior distribution for the common effect θ in model (14.2) or the mean effect μ in model (14.3). For convenience, we will refer to μ. Similar considerations for choice of a vague prior distribution apply to regression coefficients, β_m, in model (14.4).

In many meta-analyses, a vague prior distribution for μ will be preferred so that inference for the effect of primary interest is based on the observed data alone rather than being informed additionally by external information. We discuss the use of informative prior distributions for the overall effect later in the chapter.

The scale of the effect measure should be considered when choosing a prior distribution for the overall effect μ. For many effect measures, a wide normal distribution centered at the null effect would be an appropriate choice as a vague prior distribution. The variance of the normal distribution should be chosen to be large relative to likely values of the effect estimates to ensure that the prior distribution is approximately flat over the range of values supported by the likelihood. Plausible values of the effect measure should be considered when choosing a prior distribution.

For example, we consider choosing a prior distribution in meta-analysis of a binary outcome such as an adverse event or recurrence of disease, assuming the odds ratio to be the effect measure of interest. Model 14.2 will be fitted on the log odds ratio scale, so μ will represent the overall mean log odds ratio across studies. Suppose we choose $\mu \sim Normal\ (0,\ 10^4)$ as a prior distribution. This expresses a belief that 95 percent of values of μ will lie in the range $0 \pm 1.96 \times \sqrt{10^4}$, that is (–196,196).When transformed to the odds ratio scale, the corresponding interval (exp(–196),exp(196)) covers an extremely wide range of values. In most meta-analysis data sets, this distribution would be regarded as approximately flat over the range supported by the likelihood and therefore would be a suitable choice of vague prior distribution.

A wide normal distribution is an appropriate choice in meta-analysis of many common effect measures such as (log) risk ratios, risk differences, mean differences, standardized mean differences, Fisher-transformed correlations, (log) rate ratios and (log) hazard ratios. Note, however, that for mean differences, the distribution $\mu \sim Normal\ (0,\ 10^4)$ just described might not be wide enough if the units of measurement are such that values outside the range (–196,196) are plausible. This might be the case, for example, for an outcome such as resource use measured in U.S. dollars, or when small units (such as inches or seconds) are used.

14.3.4.2 Choosing a Prior Distribution for the Between-Study Variance Many meta-analyses contain only a few studies, and in these analyses the data provide

little information on the magnitude of the between-study variance τ^2. This means that the prior distribution for τ^2 is often very influential and should be chosen with particular care (Lambert et al. 2005). An advantage of performing a Bayesian meta-analysis is that relevant external evidence on the likely magnitude of τ^2 can be incorporated as an informative prior distribution, as we discuss later. In this section, we discuss choosing a vague prior distribution for τ^2.

As when choosing a prior distribution for the overall mean effect, it is important to consider what beliefs are represented by the prior distribution. Here, we need to check first that we have a clear understanding of what different values of τ^2 mean. Following the reasoning of Spiegelhalter and his colleagues, we note that 95 percent of the study effects θ_i will lie within the interval $\mu \pm 1.96\tau$ (asymptotically), and therefore that the 2.5 percent and 97.5 percent values of θ_i may be assumed to be 3.92τ apart (Spiegelhalter, Abrams, and Myles 2004, 168). For effect measures that are expressed as ratios and modeled on the log ratio scale, such as odds ratios, the value $\exp(3.92\tau)$ can be seen as representing the ratio of the 97.5 percent value to the 2.5 percent value of the distribution of true odds ratios across studies. For example, the value $\tau = 1$ corresponds to a ratio of approximately 50 between the 97.5 percent and 2.5 percent values of the studies' underlying odds ratios, which would seem implausible in most meta-analyses. When the effect measure is a mean difference, the meta-analysis model is fitted on the scale of the original outcome measure, and it is helpful to consider the magnitude of τ^2 relative to the (within-study) between-participant variances. For example, if the average between-participant standard deviation is 0.5, then a value of $\tau = 100$ would probably be considered implausibly high, but not if the average between-participant standard deviation is 500.

A common choice is to specify a uniform distribution for the between-study standard deviation τ. The upper limit of the distribution should be chosen with care to ensure that an appropriate range of values is covered. If the number of studies is small, it is likely that the results of the meta-analysis will be sensitive to the choice of prior distribution for τ^2 (11). In such meta-analyses a sensitivity analysis is recommended. As alternative choices of vague prior distribution, we could declare a half-normal prior distribution for τ or an inverse gamma prior distribution for τ^2. The inverse gamma distribution for τ^2 is a conjugate prior, conditional on the other parameters in the model. Choice of vague prior for the between-study variance is discussed in more detail by other authors (Gelman 2006; Spiegelhalter, Abrams, and Myles 2004, 170).

The variance τ^2_{res} in a meta-regression model (14.4) represents the residual between-study heterogeneity remaining after adjustment for study-level covariates. Vague priors considered for the heterogeneity variance τ^2 in a random-effects meta-analysis model (14.3) would also be suitable choices for τ^2_{res}.

14.3.5 Implementation

In the examples throughout this chapter, we undertake Bayesian estimation using MCMC simulation methods within the WinBUGS software (Lunn et al. 2000). We provide an introduction to MCMC methods and the Gibbs sampler in box 14.1.

We report posterior median values as central estimates for all model parameters; posterior distributions are often skewed and so the median is usually a more useful summary than the mean. As interval estimates, we report 95 percent credible intervals (CrI). Any interval containing 95 percent probability under the posterior distribution may be regarded as a 95 percent credible interval. Here, we use the standard approach of reporting the interval defined by the 2.5 percent and 97.5 percent percentiles of the posterior distribution. An alternative approach would be to report a *highest posterior density* interval, which is the narrowest interval containing the chosen probability, but these intervals are more difficult to compute.

In each analysis, we run three chains starting from widely dispersed values (as discussed in box 14.1) and use a diagnostic to check whether convergence has been reached (Brooks and Gelman 1998). Each set of results reported is based on running all three chains for one hundred thousand iterations following a burn-in period of ten thousand iterations, which in our examples was enough to ensure convergence in every analysis.

For model comparison, we use the deviance information criterion (DIC) (Spiegelhalter et al. 2002). The DIC is equal to the sum of the posterior mean deviance, $\bar{D}$, which measures model fit, and the effective number of parameters, p_D, which measures model complexity. Models with smaller DIC values are preferred; differences of 5 or more are suggested to be important, while there is less reason to choose between models if the DIC differs by less than 5. A tool for calculating the DIC value is available in WinBUGS.

14.3.6 Example

14.3.6.1 Fixed-Effect Meta-Analysis To illustrate application of the methods discussed, we reanalyze the data from a published meta-analysis including sixteen studies (D'Amico et al. 2009). This meta-analysis was performed to evaluate the effectiveness of a combination of topical plus systemic antibiotics versus no antibiotics in intensive care unit patients, with respect to prevention of respiratory tract infections. The odds ratio is the chosen effect measure. Table 14.1 shows the raw data and the observed study results. When calculating the log odds ratio and variance in the Jacobs 1992 study (row 7), which has a cell count of zero for events in the treated group, we have taken the standard approach of adding 0.5 to the numbers of events and nonevents in each group.

We first fit a fixed-effect model (14.2) to the respiratory tract infections data, with a normal(0,10^4) prior distribution placed on θ. The common log odds ratio θ is estimated as –1.09 (95 percent CrI –1.27 to –0.91), which corresponds to an odds ratio of 0.34 (95 percent CrI 0.28 to 0.40). This analysis shows evidence that antibiotics are beneficial for prevention of respiratory tract infections. This result is very close (identical to two decimal places) to the result obtained using a conventional frequentist inverse-variance method for fitting model (14.2), which produces an estimate of –1.09 (95 percent CI –1.27 to –0.91) for θ.

14.3.6.2 Random-Effects Meta-Analysis We now fit a random-effects meta-analysis model (14.3) to the respiratory tract infections data. We place a normal(0,10^4) prior distribution on μ, the same as for θ in the fixed-effect model, and initially place a uniform(0,2) prior distribution on τ. The mean log odds ratio μ is estimated as –1.29 (95 percent CrI –1.73 to –0.94), which corresponds to an odds ratio of 0.28 (95 percent CrI 0.18 to 0.39), and the between-study heterogeneity variance τ^2 is estimated as 0.26 (95 percent CrI 0.01 to 1.08). The odds ratio estimate and interval are similar to those from a frequentist random-effects meta-analysis, which produces an estimate of 0.28 (95 percent CI 0.21 to 0.38) for μ. The interval estimate for μ is wider in the Bayesian analysis because it takes into account the higher central estimate for τ^2 (table 14.2) and the uncertainty in estimating τ^2.

In table 14.2, we compare the results based on the uniform(0,2) prior with those obtained from two different choices of vague prior for the between-study variance: a half-normal(0,0.5^2) prior for τ or a gamma(0.001,0.001) prior for $1/\tau^2$. The central estimate for τ^2 is somewhat

Table 14.1 Respiratory Tract Infections Data

	Antibiotic Prophylaxis		No Prophylaxis		Log Odds	Var(Log
Study	Events	Total	Events	Total	Ratio	Odds Ratio)
1 Abele-Horn 1997	13	58	23	30	–2.43	0.29
2 Aerdts 1991	1	28	29	60	–3.23	1.10
3 Blair 1991	12	161	38	170	–1.27	0.12
4 Boland 1991	14	32	17	32	–0.38	0.25
5 Cockerill 1992	4	75	12	75	–1.22	0.36
6 Finch 1991	4	20	7	24	–0.50	0.51
7 Jacobs 1992	0	45	4	46	–2.27	2.27
8 Kerver 1988	5	49	31	47	–2.84	0.32
9 Krueger 2002	91	265	149	262	–0.92	0.03
10 Palomar 1997	10	50	25	49	–1.43	0.21
11 Rocha 1992	7	47	25	54	–1.59	0.24
12 Sanchez-Garcia 1992	32	131	60	140	–0.84	0.07
13 Stoutenbeek 2007	62	201	100	200	–0.81	0.04
14 Ulrich 1989	7	55	26	57	–1.75	0.23
15 Verwaest 1997	22	193	40	185	–0.76	0.08
16 Winter 1992	3	91	17	92	–1.89	0.42

SOURCE: Raw data published in D'Amico et al. 2009. Statistics calculated by authors.

Table 14.2 Random-Effects Meta-Analysis of Respiratory Tract Infections Data

	Combined OR Estimate (95 Percent CI/CrI)	Heterogeneity Variance Estimate (95 Percent CI/CrI)
Frequentist random-effects meta-analysis (DerSimonian and Laird estimation)	0.28 (0.21, 0.38)	0.18 (0.04, 1.20)
Bayesian random-effects meta-analysis, uniform(0,2) prior for τ	0.28 (0.18, 0.39)	0.26 (0.01, 1.08)
Bayesian random-effects meta-analysis, half-normal(0,0.5^2) prior for τ	0.28 (0.19, 0.39)	0.19 (0.004, 0.70)
Bayesian random-effects meta-analysis, gamma(0.001,0.001) prior for $1/\tau^2$	0.29 (0.19, 0.39)	0.16 (0.002, 0.81)

SOURCE: Authors' calculations.

sensitive to choice of vague prior, and the upper limit of the 95 percent credible interval for τ^2 changes more substantially. In this example, because of the reasonably large number of sixteen studies, the central and interval estimates for the odds ratio are not sensitive to choice of vague prior for τ^2 and change only slightly.

14.3.6.3 Random-Effects Meta-Regression As an example of random-effects meta-regression, we investigate the influence of a study-level covariate in the respiratory tract infections meta-analysis (table 14.1). The Cochrane review authors classified the studies according to whether concealment of randomized treatment allocation was adequate (studies 2, 3, 6, 9, 12–16) or not (studies 1, 4, 5, 7–11), and performed subgroup analyses to explore the impact of quality of allocation concealment on treatment effect (D'Amico et al. 2009). To investigate this here, we fit a random-effects meta-regression including a single study-level covariate x_{1i}, taking the value 1 for studies in which allocation concealment was inadequate and 0 otherwise:

$$\begin{aligned} y_i &\sim Normal\left(\theta_i, \hat{\sigma}_i^2\right) && i = 1, \ldots, k \\ \theta_i &\sim Normal\left(\beta_0 + \beta_1 x_{1i}, \tau_{res}^2\right) && i = 1, \ldots, k \\ \beta_0 &\sim P_{\beta_0}, \beta_1 \sim P_{\beta_1}, \tau \sim P_\tau && \end{aligned} \tag{14.5}$$

We choose a uniform(0,2) prior for τ_{res}, together with normal(0,10^4) distributions for β_0 and β_1. The difference β_1 between average treatment effect in studies with inadequate allocation concealment compared to that in studies with adequate allocation concealment has a central estimate of –0.58 (95 percent CrI –1.22 to 0.16). No strong evidence therefore exists of a difference between these two subgroups of studies. The residual between-study variance τ_{res}^2 is estimated as 0.12 (95 percent CrI 0.0003 to 0.88). This is lower than in the random-effects meta-analysis of these data (table 14.2), because some of the heterogeneity has been explained by the covariate x_{1i}. By calculating the exponentials of β_0 and $\beta_0 + \beta_1$, we find that the combined odds ratio in studies with adequate allocation concealment is estimated as 0.35 (95 percent CrI 0.21 to 0.49), while that in studies with inadequate allocation concealment is estimated as 0.20 (95 percent CrI 0.11 to 0.33).

14.3.6.4 Model Comparison We use the deviance information criterion to compare the fit of three different models fitted to the respiratory tract infections data above (table 14.3). The posterior mean deviance $\bar{D}$ is equal to

Table 14.3 Comparison of Models Fitted to Respiratory Tract Infections Data

Model	Posterior Mean Deviance ($\bar{D}$)	Effective Number of Parameters (p_D)	DIC Value
Fixed-effect meta-analysis	40.3	1.0	41.3
Random-effects meta-analysis	24.5	9.7	34.2
Random-effects meta-regression	26.6	8.3	34.9

SOURCE: Authors' calculations.

24.5 in the random-effects model and to 40.3 in the fixed-effect model. The much lower value for $\bar{D}$ shows that the random-effects model provides a better fit to the study data, which is unsurprising given that between-study heterogeneity is moderately high. In the fixed-effect model, the effective number of parameters p_D equals 1, representing the treatment effect. The effective number of parameters in a random-effects model must lie between the number of parameters in the fixed-effect model and the number of parameters in a model estimating independent treatment effects for all studies. In this data set, p_D must lie between 1 and 16 in the random-effects model, and has been calculated as 9.7. The DIC values for the fixed-effect and random-effects models are 41.3 and 34.2 respectively, indicating that the random-effects model is preferred. The DIC value for the random-effects meta-regression model is 34.9, which is close to the DIC value for the random-effects model, so there is no reason to choose between these two models on the basis of fit.

14.3.7 Estimating Effects for Specific Studies in Random-Effects Meta-Analysis

In the random-effects model (14.3), the θ_i represent the true effects in each of the k studies included in the meta-analysis. When using Bayesian estimation to fit this model, we obtain a posterior distribution for each θ_i, which is informed not only by the results observed for study i, but also by the fitted random-effects distribution. The posterior distributions for the θ_i are said to "borrow strength" from the other studies in the meta-analysis. In comparison with the original observed study results, uncertainty is reduced and the interval estimates for the θ_i become narrower. The central estimates obtained for the θ_i move closer together, toward the overall mean effect. This is known as shrinkage and the posterior summaries for the θ_i are often referred to as shrunken study-specific estimates.

In the respiratory tract infections example in table 14.1, the Aerdts 1991 study (row 2) had an extreme observed log odds ratio of –3.23 (95 percent CI –5.29 to –1.17) and this was imprecisely estimated because the study was small. The corresponding shrunken study-specific estimate (from the analysis with uniform(0,2) prior for τ) is –1.62 (95 percent CrI –3.02 to –0.76), which is substantially closer to the mean log odds ratio. By contrast, the Blair 1991 study (row 3) had an observed log odds ratio of –1.27 (95 percent CI –1.96 to –0.58), close to the mean; the corresponding shrunken estimate is –1.26 (95 percent CrI –1.86 to –0.71), with central estimate almost unchanged but a slightly narrower interval estimate.

In many meta-analyses, the primary focus is on summarizing the evidence across studies, and interest is scant in estimating the effects within individual studies. However, when a meta-analysis is carried out to address a particular target question, it may be that one particular study is closest to the target setting, for example, with respect to population or treatments compared. In this situation, the focus of the meta-analysis could be on this one study's effect estimate and on increasing its precision through borrowing information from other similar studies.

14.3.8 Predicting the Effect in a New Study

An important advantage of a Bayesian random-effects meta-analysis is that it enables us to predict the effect expected in a future study. In study design, prediction from a meta-analysis of existing evidence can be used to estimate the probability that a new planned study of a given size will produce statistically significant results. The predictive distribution for the effect in a new study is obtained directly from the random-effects distribution, $\theta^{new} \sim Normal\ (\mu, \tau^2)$, under the assumption that the new study can be considered exchangeable with the studies in the meta-analysis.

In the respiratory tract infections example, the log odds ratio expected in a new study is estimated as –1.26 (95 percent CrI –2.61 to –0.10) (based on the analysis with uniform(0,2) prior for τ), which corresponds to an odds ratio of 0.28 (95 percent CrI 0.07 to 0.91). Prediction of an effect in a new individual study is associated with much higher uncertainty than the uncertainty associated with estimating the mean of the random distribution of study effects. When using the outputs from a random-effects meta-analysis in decision modeling, it is important to consider whether the mean effect or the predictive effect for a new study should be used (Ades, Lu, and Higgins 2005).

14.4 BAYESIAN APPROACHES FOR SPECIFIC TYPES OF DATA

The random-effects meta-analysis model (14.3) is a generic model and can be used to perform meta-analysis of any effect measure for which observed estimates and within-study variances are available from each study. In this model and in the fixed-effect model (14.2), normality is assumed for the within-study likelihood. We now

discuss using exact within-study likelihoods for specific types of outcome data.

14.4.1 Binary Data

When analyzing binary outcome data, we can model the within-study likelihood as binomial rather than assuming normality (Smith, Spiegelhalter, and Thomas 1995). This is straightforward when the effect measure of interest is the odds ratio. The numbers of events r_{i0}, r_{i1} and total numbers of patients n_{i0}, n_{i1} in the treatment arms are modeled directly, rather than modeling estimated log odds ratios and their variances. The random-effects meta-analysis model is now written as

$$\begin{aligned} r_{i0} &\sim Binomial(p_{i0}, n_{i0}) \\ r_{i1} &\sim Binomial(p_{i1}, n_{i1}) \\ \text{logit}(p_{i0}) &= \alpha_i - \theta_i/2 \\ \text{logit}(p_{i1}) &= \alpha_i + \theta_i/2 \\ \theta_i &\sim Normal(\mu, \tau^2) \qquad i = 1, \ldots, k \\ \mu &\sim P_\mu, \tau \sim P_\tau, \alpha_i \sim P_{\alpha_i} \end{aligned} \tag{14.6}$$

Advantages of this approach are that we avoid the need to adjust for zero cells when estimating log odds ratios and their variances, and that the within-study likelihood is modeled exactly rather than assumed normal. In many data sets, the difference between the results obtained from models (14.3) and (14.6) will be minimal, but greater when studies are small or event rates are extreme.

Methods for performing Bayesian meta-analysis of risk differences or relative risks while assuming binomial within-study likelihoods were proposed by David Warn, Simon Thompson, and David Spiegelhalter (2002). Implementation of a binomial likelihood model is less straightforward for risk differences or relative risks than for odds ratios because care has to be taken to ensure that the values of the risk differences or relative risks are appropriately constrained.

14.4.1.1 Example We fit model (14.6) to the respiratory tract infections example, specifying vague normal(0,10^4) priors for the average log odds α_i in each study as well as for μ, and a uniform(0,2) prior for τ. As previously, an initial five thousand iterations were discarded as burn-in and estimates were based on the following hundred thousand. The log odds ratio is estimated as –1.40 (95 percent CrI –1.88 to –1.01), which corresponds to an odds ratio of 0.25 (95 percent CrI 0.15 to 0.36); τ^2 is estimated as 0.40 (95 percent CrI 0.08 to 1.41). In this example, the results differ quite substantially between models (14.3) and (14.6), and it would be preferable to use the exact binomial likelihood model.

14.4.2 Continuous Data

For a continuous outcome, an exact likelihood model can be fitted if we have access to the observed mean and its standard error from each arm of each study. An assumption of normality for the mean value in each arm is based on the central limit theorem if sample sizes are large enough. In the random-effects model (14.7), y_{ij} represents the observed mean in arm j of study i, and $\hat{\sigma}_{ij}$ represents its standard error, and mean differences are modeled.

$$\begin{aligned} y_{i0} &\sim Normal(\alpha_i - \theta_i/2, \hat{\sigma}_{i0}^2) \\ y_{i1} &\sim Normal(\alpha_i + \theta_i/2, \hat{\sigma}_{i1}^2) \\ \theta_i &\sim Normal(\mu, \tau^2) \\ \mu &\sim P_\mu, \tau \sim P_\tau, \alpha_i \sim P_{\alpha_i} \end{aligned} \tag{14.7}$$

We specify vague normal(0,10^4) distributions for the average means α_i in each study. Choices of prior distribution for the overall effect μ and the between-study variance τ^2 are discussed earlier in the chapter.

14.4.3 Rate Data

In a meta-analysis comparing rates of a particular event over time, we model the number of events y_{ij} in arm j of trial i; this depends on the rate λ_{ij} at which events occur in the trial arm and the exposure time E_{ij}. In model (14.8), we present a random-effects meta-analysis model on the log rate difference scale:

$$\begin{aligned} y_{ij} &\sim Poisson(\lambda_{ij} E_{ij}) \\ \log(\lambda_{i0}) &= \alpha_i - \theta_i/2 \\ \log(\lambda_{i1}) &= \alpha_i + \theta_i/2 \\ \theta_i &\sim Normal(\mu, \tau^2) \\ \mu &\sim P_\mu, \tau \sim P_\tau, \alpha_i \sim P_{\alpha_i} \end{aligned} \tag{14.8}$$

This model assumes that the rate of events in each study arm remains constant over the length of exposure.

14.5 INFORMATIVE PRIOR DISTRIBUTIONS

14.5.1 Informative Prior Distributions for the Between-Study Variance

Earlier in this chapter, we use vague prior distributions for all unknown parameters in the models presented to reflect a lack of prior knowledge in advance of seeing the data. We now consider choosing informative prior distributions for the between-study heterogeneity variance τ^2 in random-effects meta-analysis. In meta-analyses including only a few studies, τ^2 is imprecisely estimated. Frequentist estimation of the combined effect does not take this imprecision into account, and Bayesian analyses are sensitive to the choice of vague prior for τ^2. It would be preferable for Bayesian meta-analyses to use informative prior distributions for τ^2, informed by relevant external evidence on the likely magnitude of between-study heterogeneity (Higgins and Whitehead 1996).

Recently, predictive distributions for the heterogeneity expected in a variety of medical settings have been published. These were constructed by modeling the data from a large collection of meta-analyses. We will demonstrate how these distributions can be used as informative prior distributions for heterogeneity in a new meta-analysis.

In a meta-regression model (14.4), the between-study variance τ^2_{res} represents the residual between-study heterogeneity remaining after adjustment for study-level covariates. It is unlikely that a data-based predictive distribution for residual heterogeneity in a particular meta-regression model would ever be available, given that we would need multiple examples of the same model fitted in the same setting. However, we could use an informative prior recommended for total heterogeneity τ^2 in a particular setting as an informative prior for residual heterogeneity τ^2_{res} (Jackson et al. 2014). We would expect some heterogeneity to be explained by the covariates and consequently τ^2_{res} is likely to be smaller than τ^2. This prior would therefore be a conservative choice, since the prior distribution would support somewhat larger values of τ^2_{res} than are likely to occur and would lead to wider intervals for the meta-regression coefficients β_m.

14.5.1.1 Empirical Data-Based Priors We previously used data from the Cochrane Database of Systematic Reviews to construct predictive distributions for between-study heterogeneity in a variety of different medical settings and for several types of effect measure. To construct distributions for heterogeneity of log odds ratios, we used data from 14,886 meta-analyses, including data from 77,237 individual studies in total (Turner et al. 2012). We fitted hierarchical models, in which all 14,886 meta-analyses were performed simultaneously; in a regression model, we explored the effects of meta-analysis characteristics on the magnitude of heterogeneity. We find that type of outcome and type of intervention comparison evaluated in the meta-analysis both influenced the magnitude of between-study heterogeneity. For example, heterogeneity variances for meta-analyses in which the outcome was all-cause mortality were lower than heterogeneity variances for other outcomes. We presented predictive distributions for nine broad settings, defined by outcome type and intervention comparison type. In a later paper, we presented predictive distributions for eighty narrower settings, again defined by outcome type and intervention comparison type (Turner et al. 2015). Similar methods have been used to construct predictive distributions for heterogeneity of standardized mean differences in meta-analyses of continuous outcomes (Rhodes, Turner, and Higgins 2016).

14.5.1.2 Example We first return to the example presented in table 14.1 and choose a suitable informative prior distribution for the extent of heterogeneity expected in this setting. The outcome in this meta-analysis is respiratory tract infections and the analysis compares topical plus systemic antibiotic prophylaxis versus no prophylaxis. Selecting from the eighty settings we presented, we find that the example meta-analysis fits well into the outcome category of Infection/onset of new disease and the intervention comparison category of Pharmacological versus Placebo/control. A log-normal(-2.49, 1.52^2) distribution was recommended for between-study heterogeneity in this setting. When specifying this prior for τ^2 in a random-effects meta-analysis, the heterogeneity variance is estimated as 0.15 (95 percent CrI 0.01 to 0.62). The central estimate has reduced in comparison with the analyses based on vague priors for τ^2 (table 14.2), and the interval estimate is narrower because additional information has been provided through the prior distribution. The estimate for the combined log odds ratio μ has changed slightly to -1.25 (95 percent CrI -1.63 to -0.95), with corresponding odds ratio 0.29 (95 percent CrI 0.20 to 0.39).

Next, we repeat the meta-regression analysis carried out in section 14.3.6.3, using the same informative prior distribution for τ^2_{res}. The estimate for τ^2_{res} reduces to 0.07 (95 percent CrI 0.004 to 0.45) as a result of incorporating external information. Correspondingly, the interval estimates for the meta-regression coefficients have narrowed; the difference β_1 is now estimated as -0.59 (95 percent CrI -1.16 to 0.02).

Table 14.4 Recurrence of Violence Data

	CBT		Control			
Study	Events	Total	Events	Total	Log Odds Ratio	Var(Log Odds Ratio)
1 Bronx 2005	20	202	11	218	0.73	0.15
2 Brooklyn 2000	13	129	100	386	–1.14	0.10
3 Broward 2000	52	216	45	188	0.01	0.05
4 San Diego Navy 2000	63	218	75	214	–0.28	0.04

SOURCE: Raw data published in Smedslund et al. 2011. Statistics calculated by authors.

Many meta-analyses contain much smaller numbers of studies than the respiratory tract infections example. Jonathan Davey and his colleagues find that 75 percent of meta-analyses published in the Cochrane Database of Systematic Reviews in the first 2008 issue include no more than five studies (2011). To illustrate use of an informative prior for heterogeneity when the data provide very limited information, we reanalyze the data from a meta-analysis including four studies (Smedslund et al. 2011). This meta-analysis evaluated the effectiveness of cognitive behavioral therapy for men who physically abuse their partner, with respect to recurrence of violence (table 14.4). The odds ratio is the chosen effect measure.

In a frequentist random-effects meta-analysis using method-of-moments estimation, the heterogeneity variance estimate is moderately high at 0.31, but extremely imprecise (95 percent CI 0.07 to 8.15, calculated using the Q-profile method). The usual frequentist analysis does not allow for this imprecision in estimation of the combined odds ratio, so the confidence interval for μ is inappropriately narrow (table 14.5). If we fit the random-effects model using Bayesian estimation, with a vague prior specified for τ^2, we find that estimates and interval estimates for τ^2 and interval estimates for the combined log odds ratio μ are sensitive to the choice of vague prior (table 14.5).

Although this meta-analysis is not in a medical research setting, we use a predictive distribution for broad categorizations of outcomes and interventions (Turner et al. 2012). The CBT versus control comparison fits well into the category of nonpharmacological comparisons and recurrence of violence would have been categorized as

Table 14.5 Random-Effects Meta-Analysis of Recurrence of Violence Data

	Combined OR Estimate (95 Percent CI/CrI)	Heterogeneity Variance Estimate (95 Percent CI/CrI)
Frequentist random-effects meta-analysis (DerSimonian and Laird estimation)	0.82 (0.45, 1.52)	0.31 (0.07, 8.15)
Bayesian random-effects meta-analysis, uniform(0,2) prior for τ	0.83 (0.27, 2.54)	0.74 (0.06, 3.51)
Bayesian random-effects meta-analysis, half-normal(0,0.5^2) prior for τ	0.82 (0.41, 1.67)	0.30 (0.02, 1.27)
Bayesian random-effects meta-analysis, gamma(0.001,0.001) prior for $1/\tau^2$	0.82 (0.30, 2.37)	0.42 (0.01, 5.97)
Bayesian random-effects meta-analysis, log-normal(–3.95,1.79^2) prior for τ^2	0.82 (0.39, 1.74)	0.29 (0.03, 1.94)

SOURCE: Authors' calculations.

a subjective outcome. We will assume that levels of between-study heterogeneity for meta-analyses evaluating this type of outcome and intervention comparison in medical research are similar to those in social science. A log-normal($-2.01,1.64^2$) distribution was recommended for heterogeneity in this setting (22). When the informative prior distribution was used, the central estimate for τ^2 changed to 0.29, with 95 percent credible interval 0.03 to 1.94. In this example, we prefer to incorporate relevant external information on the likely values of τ^2 than to estimate the combined treatment difference using a very imprecise estimate of τ^2 in a frequentist analysis. The combined odds ratio has changed to 0.82 (95 percent CrI 0.39 to 1.74) in the Bayesian meta-analysis.

14.5.2 Informative Prior Distributions for the Overall Effect

In most meta-analyses, researchers prefer to use a vague rather than informative prior for the overall effect so that the parameter of primary interest is estimated only from the current data set. However, an informative prior may be chosen if certain relevant external studies cannot directly be included in the meta-analysis, if there is a desire to incorporate expert opinion on the magnitude of the effect, or if there is interest in assessing the impact of the observed data on one or more particular prior distributions. We describe a published example of the first scenario in the following section. If researchers wish to incorporate expert opinion on the effect of primary interest, it is advisable to seek detailed guidance on elicitation of opinion-based prior distributions (for a review of this area, see Spiegelhalter, Abrams, and Myles 2004). The impact of data on an intentionally skeptical prior distribution has been demonstrated in interpretation of a meta-analysis of clinical trials of intravenous magnesium after acute myocardial infarction (Higgins and Spiegelhalter 2002).

14.5.2.1 Example Alex Sutton and Keith Abrams present results from an evaluation of the effectiveness of electronic fetal heart rate monitoring (EFM), in which relevant observational evidence was used to construct an informative prior for the treatment effect in a meta-analysis of randomized trials (2001). Data were available from nine randomized trials comparing perinatal mortality rates between expectant mothers who received EFM during labor and mothers who did not. The risk difference (per thousand births) was the effect measure of interest and a Bayesian random-effects meta-analysis (see model 14.3) of the randomized trial data was fitted. Initially, a vague normal($0,10^6$) prior was chosen for the combined risk difference μ and a gamma(0.001,0.001) prior was chosen for $1/\tau^2$. This analysis produced an estimate of 1.07 (95 percent CrI –2.53 to 1.71) for μ.

In addition to that available from randomized trials, considerable evidence on the effectiveness of EFM was available from observational studies. These studies were believed to be of lower quality than the trials, so it was not considered appropriate to include them in the meta-analysis. Sutton and Abrams discuss how such evidence could instead be incorporated as an informative prior in the meta-analysis of the trial data (2001). A random-effects meta-analysis of the observational studies was performed and produced an estimate of –1.64 for the risk difference per thousand births, with a standard error of 0.45. Assuming normality, this result is translated into a normal($-1.64,0.45^2$) prior for μ. A random-effects meta-analysis of the trial data, incorporating this prior, now produces an estimate of 0.42 (95 percent CrI –2.19 to –0.55) for μ. The original result based on the trial evidence alone has shifted some way toward the result based on observational evidence. Sutton and Abrams also discuss how to downweight the prior distribution based on observational evidence to reduce its influence on the meta-analysis.

14.5.3 Informative Prior Distributions for Other Quantities

We have discussed how to introduce external information on two of the standard meta-analysis model parameters. It is also possible to extend the standard models specifically to incorporate relevant external information on other quantities. In this section, we discuss two examples.

14.5.3.1 Allowing for Within-Study Biases Studies within a meta-analysis often vary in quality, and flaws in them can potentially lead to biased estimation of the overall effect. Given concerns about study quality, researchers may want to make allowance for within-study biases in the meta-analysis. In a bias-adjusted meta-analysis, relatively less weight is given to the studies judged to be at high risk of bias, and thus the results of lower quality studies have less influence on estimation of the combined effect. The seminal work in this area was carried out by David Eddy and his colleagues, who presented models allowing for multiple biases in meta-analysis (Eddy, Hasselblad, and Schachter 1992). Model (14.9) is a simple extension of the standard random-effects

model, including study-specific bias parameters β_i representing the bias resulting from one characteristic:

$$\begin{aligned} y_i &\sim Normal\left(\theta_i + \beta_i, \hat{\sigma}_i^2\right) \qquad i = 1, \ldots, k \\ \theta_i &\sim Normal\left(\mu, \tau^2\right) \\ \mu &\sim P_\mu, \tau \sim P_\tau, \beta_i \sim P_{\beta_i} \end{aligned} \tag{14.9}$$

Informative prior distributions for the β_i could be based on empirical evidence on the expected magnitude of certain biases or on expert opinion. Nicky Welton and her colleagues constructed distributions for the bias associated with a particular methodological flaw by using evidence from an external collection of meta-analyses (Welton et al. 2009). In their model, studies in a new meta-analysis are judged according to whether they are at low or high risk of bias due to a given flaw. The study-specific bias β_i is set to zero in studies judged to be at low risk of bias. In studies at high risk of bias, the bias parameter β_i is given an informative prior distribution, which is derived from a hierarchical model fitted to the biases associated with this flaw in each of a large collection of meta-analyses. Welton and her colleagues illustrate their method by adjusting for the bias associated with inadequate or unclear allocation concealment in a meta-analysis of twenty-one randomized trials. A distribution for the bias affecting the sixteen trials judged to have inadequate or unclear allocation concealment was derived from an external collection of thirty-three meta-analyses including 250 trials.

Rebecca Turner and colleagues considered basing prior distributions for the study-specific bias parameters β_i on expert opinion rather than empirical evidence (2009). Under this method, the methodological quality of each study in the new meta-analysis was assessed in detail and expert opinion was used to construct a distribution for the bias expected to affect each study's results. As when using empirical evidence, the weighting of the studies in the meta-analysis was altered by allowance for the expected levels of bias, and the lower quality studies had relatively less influence on the overall bias-adjusted result.

14.5.3.2 Allowing for Within-Subject Correlation Keith Abrams and his colleagues demonstrate how a Bayesian approach can be used to allow for heterogeneous reporting of study results (2000). Their motivating example was a meta-analysis assessing the impact of testing positive or negative in a screening program on levels of long-term anxiety. The measure of interest was the change in anxiety between baseline and follow-up. Two of the six studies in the meta-analysis had reported estimates of change, together with standard deviations. The other four studies had reported only baseline and follow-up levels of anxiety. These results can be used to calculate estimates of change, but calculation of the correct standard deviations requires knowledge of the within-subject correlations between baseline and follow-up. Abrams and his colleagues fit a Bayesian meta-analysis model in which the variance $\hat{\sigma}_i^2$ of each estimated change from baseline (unless already reported) was assumed to depend on the unknown within-subject correlation ρ:

$$\begin{aligned} \hat{\sigma}_i^2 &= \frac{(n_{i0}-1)V(d_{i0}) + (n_{i1}-1)V(d_{i1})}{n_{i0}+n_{i1}-2} \\ V(d_{ij}) &= V(x_{ij}) + V(y_{ij}) - 2\rho\sqrt{V(x_{ij})V(y_{ij})},\ j = 0, 1 \end{aligned} \tag{14.10}$$

In expression (14.10), x_{ij}, y_{ij} and d_{ij} represent the baseline measurement, follow-up measurement and change from baseline respectively, in trial arm j of study i. An informative prior for ρ was derived by performing a meta-analysis of six estimates of within-subject correlation for anxiety levels and using a moment-based approach to fit a gamma distribution to the results obtained.

A meta-analysis of change from baseline in anxiety is then performed, combining evidence from all six studies of the impact of screening results on change in anxiety levels, while allowing appropriately for within-subject correlation in anxiety over time.

14.6 DISCUSSION

We have demonstrated some of the advantages of Bayesian meta-analysis: prediction of the effect in a future study, incorporation of external evidence on between-study heterogeneity or the combined effect, and full allowance for uncertainty in estimation. An additional advantage is flexibility in modeling. Because Bayesian estimation of more complex models is easily achieved using MCMC methods, the basic models presented here can be extended to perform network meta-analysis (Higgins and Whitehead 1996; Dias et al. 2013) or multivariate meta-analysis (Wei and Higgins 2013), or to allow for varying types of study within the meta-analysis, by adding an extra layer of variation to the hierarchical models (Prevost, Abrams, and Jones 2000).

A challenge of Bayesian meta-analysis is that selection of prior distributions requires care and the results of the meta-analysis may be sensitive to the choices made, particularly in small data sets. If relevant external evidence is available for the between-study heterogeneity variance, we recommend that an informative prior is used for this parameter in preference to a vague prior. However, published predictive distributions for heterogeneity are currently available only for health-related meta-analyses, not for every type of effect measure. Whether using informative or vague prior distributions, sensitivity of the meta-analysis results should always be explored. Another drawback of using MCMC methods to perform Bayesian estimation is that these methods are computationally intensive and require analysts to ensure that convergence has been reached. Alternative implementations of Bayesian meta-analysis may be available, based, for example, on numerical integration or importance sampling, though these methods have so far been described for only a limited range of models (Turner et al. 2015).

In summary, Bayesian meta-analysis offers a number of useful benefits. The opportunity to incorporate external information on heterogeneity is particularly valuable in meta-analyses including few studies; incorporating external information on the combined effect may be worthwhile when relevant evidence is available from studies that cannot be directly included in the meta-analysis. When meta-analyses are performed to inform decision making, the predictive distribution for the effect in a future study is often required, and this is available only from Bayesian meta-analysis. These benefits may be weighed against the need to choose prior distributions with care and to carry out large numbers of simulations.

14.7 REFERENCES

Abrams, Keith R., Paul C. Lambert, Bruno Sansó, and Chris Shaw. 2000. "Meta-Analysis of Heterogeneously Reported Study Results: A Bayesian Approach." In *Meta-Analysis in Medicine and Health Policy*, edited by Dalene K. Stangl and Don A. Berry. New York: Marcel Decker.

Ades, A. E., Guobing Lu, and Julian P. T. Higgins. 2005. "The Interpretation of Random-Effects Meta-Analysis in Decision Models." *Medical Decision Making* 25(6): 646–54.

Best, Nicky G., David J. Spiegelhalter, Andrew Thomas, and Carol E. G. Brayne. 1996. "Bayesian Analysis of Realistically Complex Models." *Journal of the Royal Statistical Society, Series A* 159(2): 323–42.

Brooks, Stephen P. 1998. "Markov Chain Monte Carlo Method and Its Application." *The Statistician* 47(1): 69–100.

Brooks, Stephen P., and Andrew Gelman. 1998. "General Methods for Monitoring Convergence of Iterative Simulations." *Journal of Computational and Graphical Statistics* 7(4): 434–55.

D'Amico, Roberto, Silvia Pifferi, Valter Torri, Luca Brazzi, Elena Parmelli, and Alessandro Liberati. 2009. "Antibiotic Prophylaxis to Reduce Respiratory Tract Infections and Mortality in Adults Receiving Intensive Care." *Cochrane Database of Systematic Reviews* 4.

Davey, Jonathan, Rebecca M. Turner, Mike J. Clarke, and Julian P. T. Higgins. 2011. "Characteristics of Meta-Analyses and Their Component Studies in the *Cochrane Database of Systematic Reviews*: A Cross-Sectional, Descriptive Analysis." *BMC Medical Research Methodology* 11: 160.

Dias, Sofia, Alex J. Sutton, A. E. Ades, and Nicky J. Welton. 2013. "Evidence Synthesis for Decision Making 2: A Generalized Linear Modeling Framework for Pairwise and Network Meta-Analysis of Randomized Controlled Trials." *Medical Decision Making* 33: 607–17.

Eddy, David M., Victor Hasselblad, and Ross Schachter. 1992. *Meta-Analysis by the Confidence Profile Method: The Statistical Synthesis of Evidence*. San Diego, Calif.: Academic Press.

Gelfand, Alan E., and Adrian F. M. Smith. 1990. "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association* 85(410): 398–409.

Gelman, Andrew. 2006. "Prior Distributions for Variance Parameters in Hierarchical Models." *Bayesian Analysis* 1(3): 515–33.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. 2009. "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models." *Annals of Applied Statistics* 2(4): 1360–83.

Hardy, Rebecca J., and Simon G. Thompson. 1996. "A Likelihood Approach to Meta-Analysis with Random Effects." *Statistics in Medicine* 15: 619–29.

Higgins, Julian P. T., and David J. Spiegelhalter. 2002. "Being Sceptical About Meta-Analyses: A Bayesian Perspective on Magnesium Trials in Myocardial Infarction." *International Journal of Epidemiology* 31(1): 96–104.

Higgins, Julian P. T, Simon G. Thompson, and David J. Spiegelhalter. 2009. "A Re-Evaluation of Random-Effects Meta-Analysis." *Journal of the Royal Statistical Society, Series A* 172(1): 137–59.

Higgins, Julian P. T., and Anne Whitehead. 1996. "Borrowing Strength from External Trials in a Meta-Analysis." *Statistics in Medicine* 15(24): 2733–49.

Jackson, Dan, Rebecca Turner, Kirsty Rhodes, and Wolfgang Viechtbauer. 2014. "Methods for Calculating Confidence and Credible Intervals for the Residual Between-Study Variance in Random Effects Meta-Regression Models." *BMC Medical Research Methodology* 14(1): 103.

Kass, Robert E., and Larry Wasserman. 1996. "The Selection of Prior Distributions by Formal Rules." *Journal of the American Statistical Association* 91(435): 1343–70.

Lambert, Paul C., Alex J. Sutton, Paul R. Burton, Keith R. Abrams, and David R. Jones. 2005. "How Vague Is Vague? A Simulation Study of the Impact of the Use of Vague Prior Distributions in MCMC Using WinBUGS." *Statistics in Medicine* 24(15): 2401–28.

Lunn, David J., Andrew Thomas, Nicky Best, and David Spiegelhalter. 2000. "WinBUGS—A Bayesian Modelling Framework: Concepts, Structure, and Extensibility." *Statistics and Computing* 10(4): 325–37.

Prevost, Teresa C., Keith R. Abrams, and David R. Jones. 2000. "Hierarchical Models in Generalized Synthesis of Evidence: An Example Based on Studies of Breast Cancer Screening." *Statistics in Medicine* 19(24): 3359–76.

Rhodes, Kirsty, Rebecca M. Turner, and Julian P. T. Higgins. 2016. "Predictive Distributions Were Developed for the Extent of Heterogeneity in Meta-Analyses of Continuous Outcome Data." *Journal of Clinical Epidemiology* 68(1): 52–60.

Smedslund, Geir, Therese K. Dalsbo, Asbjørn K. Steiro, Aina Winsvold, and Jocelyne Clench-Aas. 2011. "Cognitive Behavioural Therapy for Men Who Physically Abuse Their Female Partner." *Cochrane Systematic Reviews* 2.

Smith, Adrian F. M., and Gareth O. Roberts. 1993. "Bayesian Computation Via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods." *Journal of the Royal Statistical Society, Series B* 55(1): 3–23.

Smith, Teresa C., David J. Spiegelhalter, and Andrew Thomas. 1995. "Bayesian Approaches to Random-Effects Meta-Analysis: A Comparative Study." *Statistics in Medicine* 14(24): 2685–99.

Spiegelhalter, David J., Keith R. Abrams, and Jonathan P. Myles. 2004. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Chichester, UK: John Wiley & Sons.

Spiegelhalter, David J., Nicola G. Best, Bradley P. Carlin, Angelika van der Linde. 2002. "Bayesian Measures of Model Complexity and Fit (with Discussion)." *Journal of the Royal Statistical Society, Series B* 64(4): 583–639.

Spiegelhalter, David J., Lawrence S. Freedman, and Mahesh K. B. Parmar. 1994. "Bayesian Approaches to Randomized Trials." *Journal of the Royal Statistical Society, Series A* 157(3): 357–416.

Sutton, Alex J., and Keith R. Abrams. 2001. "Bayesian Methods in Meta-Analysis and Evidence Synthesis." *Statistical Methods in Medical Research* 10(4): 277–303.

Turner, Rebecca M., Jonathan Davey, Mike J. Clarke, Simon G. Thompson, and Julian P. T. Higgins. 2012. "Predicting the Extent of Heterogeneity in Meta-Analysis, Using Empirical Data from the Cochrane Database of Systematic Reviews." *International Journal of Epidemiology* 41(3): 818–27.

Turner, Rebecca M., Dan Jackson, Yinghui Wei, Simon G. Thompson, and Julian P. T. Higgins. 2015. "Predictive Distributions for Between-Study Heterogeneity and Simple Methods for Their Application in Bayesian Meta-Analysis." *Statistics in Medicine* 34(6): 984–98.

Turner, Rebecca M., David J. Spiegelhalter, Gordon C. S. Smith, and Simon G. Thompson. 2009. "Bias Modelling in Evidence Synthesis." *Journal of the Royal Statistical Society, Series A* 172(1): 21–47.

Warn, David E., Simon G. Thompson, and David J. Spiegelhalter. 2002. "Bayesian Random Effects Meta-Analysis of Trials with Binary Outcomes: Methods for the Absolute Risk Difference and Relative Risk Scales." *Statistics in Medicine* 21(11): 1601–23.

Wei, Yinghui, and Julian P. T. Higgins. 2013. "Bayesian Multivariate Meta-Analysis with Multiple Outcomes." *Statistics in Medicine* 32(17): 2911–34.

Welton, Nicky J., A E. Ades, John B. Carlin, Douglas G. Altman, and Jonathan A. C. Sterne. 2009. "Models for Potentially Biased Evidence in Meta-Analysis Using Empirically Based Priors." *Journal of the Royal Statistical Society, Series A* 172(1): 119–36.

15

CORRECTING FOR THE DISTORTING EFFECTS OF STUDY ARTIFACTS IN META-ANALYSIS AND SECOND ORDER META-ANALYSIS

FRANK L. SCHMIDT
University of Iowa

HUY LE
University of Texas at San Antonio

IN-SUE OH
Temple University

CONTENTS

15.1 ARTIFACTS THAT DISTORT OBSERVED STUDY RESULTS

Every study has imperfections, many of which bias the results. In some cases, we can define precisely what a methodologically ideal study would be like, and thus can say that the effect size obtained from any real study will differ to some extent from the value that would have been obtained had the study been methodologically perfect. Although it is important to estimate and eliminate bias in individual studies, it is even more important to remove such errors in research syntheses such as meta-analyses.

Some authors have argued that meta-analysts should not correct for study imperfections because the purpose of meta-analysis is only to provide a description of study findings, not an estimate of what would have been found in methodologically ideal studies. However, the errors and biases that stem from study imperfections are artifactual; they stem from imperfections in our research methods, not from the underlying relationships that are of scientific interest (Rubin 1990). Thus, scientific questions are better addressed by estimates of the results that would have been observed had studies been free of methodological biases (Cook et al. 1992; Schmidt and Hunter 2015, 34–36; Rubin 1990; Schmidt 1992). For example, in correlational research, the results most relevant to evaluation of a scientific theory are those that would be obtained from a study using an infinitely large sample from the relevant population (that is, the population itself) and using measures of the independent and dependent variables that are free of measurement error and perfectly construct valid. Such a study would be expected to provide an exact estimate of the relation between constructs in the population of interest; such an estimate is maximally relevant to the testing and evaluation of scientific theories (and also to theory construction). Thus corrections for biases and other errors in study findings due to study imperfections (which we call artifacts) are essential to the development of valid cumulative knowledge. The increasing use of estimates from meta-analysis as input into causal modeling procedures further underlines the importance of efforts to ensure that meta-analysis findings are free of correctable bias and distortion (for example, see Colquitt, LePine, and Noe 2002; Becker and Schram 1994). In the absence of such corrections, the results of path analyses and other causal modeling procedures are biased in complex and often unpredictable ways (Coffman and MacCallum 2005).

The goal of research is accurate estimation of relationships between constructs. Correcting for research artifacts provides the most accurate estimates of correlations at the true score level. It is sometimes argued that true scores are not accurate estimates of construct scores. However, the evidence shows that in all but unusual cases the correlation between true scores and the relevant construct is very high, that is, in the high 0.90s (Schmidt, Le, and Oh 2009). Hence true scores provide good estimates of construct scores. The discussion in this chapter is oriented to the scientific examination of theories and hypotheses. In some areas of applied work in which the goal is not theory evaluation but instead empirical prediction of human performance (for example, in educational or work performance), it is not necessary or appropriate *to correct the observed relationships for error of measurement in the predictive scales used*. Such applied cases are noted again later where appropriate.

Most artifacts with which we are concerned have been studied in the field of psychometrics. The goal is to develop methods of calibrating each artifact and correcting for its effects. The procedures for correcting for these artifacts can be complex, but software for applying them is available (Schmidt and Le 2014). The procedures summarized in this chapter are more fully detailed elsewhere (Schmidt and Hunter 2015). They are presented there for both the correlation coefficient and the standardized mean difference (*d* value statistic); however, for economy of presentation this chapter presents only their application to correlations. The procedures described in this chapter can

also be applied to other effect-size statistics such as odds ratios and related risk statistics.

15.1.1 Unsystematic Artifacts

Some artifacts produce a systematic effect on the study effect size and some cause unsystematic (random) effects. Even within a single study, it is possible to correct for a systematic effect. Unsystematic effects usually cannot be corrected in single studies and sometimes may not be correctable even at the level of meta-analysis. The two major unsystematic artifacts are sampling error and data errors.

15.1.1.1 Sampling Errors It is not possible to correct for the effect of sampling error in a single study. The confidence interval gives an idea of the potential size of the sampling error, but the magnitude of the sampling error in any one study is unknown and hence cannot be corrected. However, the effects of sampling error can be greatly reduced or eliminated in meta-analysis if the number of studies (k) is large enough to produce a large total sample size, because sampling errors are random and average out across studies. If the total sample size in the meta-analysis is not large, one can still correct for the effects of sampling error, though the correction is less precise and some smaller amount of sampling error will remain in the final meta-analysis results, a second-order sampling error (see Schmidt and Hunter 2015, chapter 9; Schmidt and Oh 2013). A meta-analysis that corrects only for sampling error and ignores other artifacts is a partial meta-analysis and is therefore called a bare-bones meta-analysis.

15.1.1.2 Data Errors and Outliers Bad data in meta-analysis stem from a variety of errors in handling data: primary data used in a study may be erroneous due to transcription errors, coding errors, and so on; the initial results of the analysis of the primary data in a particular study may be incorrect due to computational errors, transcriptional errors, computer program errors, and so on; the study results as published may have errors caused by transcriptional error by the investigator, by a typist, or by a printer; or a meta-analyst may miscopy a result or make a computational error. Data errors are apparently quite common (Gulliksen 1986; Tukey 1960). Sometimes such errors can be detected and eliminated using outlier analysis, but outlier analysis is problematic in meta-analysis because it is often impossible to distinguish between data errors and large sampling errors. Deletion of data with large sampling errors can bias corrections for sampling error (Schmidt and Hunter 2015, 235–36).

15.1.2 Systematic Artifacts

Many artifacts have a systematic influence on study effect size parameters and their estimates. If such an effect can be quantified, often there is an algebraic formula for the effect of the artifact. Most algebraic formulas can be inverted, producing a correction formula. The resulting correction removes the bias created by the artifact and estimates the effect size that would have been obtained had the researcher carried out a study without the corresponding methodological limitation.

Correction for an artifact requires knowledge about the size of the effect of that artifact. Correction for each new artifact usually requires at least one new piece of information. For example, to correct for the effects of random error of measurement in the dependent variable, we need to know the reliability of the dependent variable in the primary studies. Many primary studies do not present information on the artifacts in the study, but often this information (for example, scale reliability) is available from other sources. Even when artifact information is presented in the study, it is not always of the required type. For example, in correcting for the influence of measurement error, it is important to use the appropriate type of reliability coefficient. Use of an inappropriate coefficient will lead to a correction that is at least somewhat erroneous (Schmidt and Hunter 2015, 115–21), usually an undercorrection.

There are at least ten systematic artifacts that can be corrected if the artifact information is available. The correction can be made within each study individually if artifact information is available for all (or nearly all) studies individually. If so, then the meta-analysis is performed on these corrected values. If this is not the case, the correction can be made at the level of the meta-analysis if the distribution of artifact values across studies can be estimated.

As pointed out elsewhere in this volume, study effect sizes can be expressed in a variety of ways, the two most frequently used indices being the correlation coefficient and the standardized mean difference (d value and its variations). For ease of explication, artifact effects and corrections are discussed in this chapter in terms of correlations. The same principles apply to standardized mean differences, although it is often more difficult to make appropriate corrections for artifacts affecting the independent variable in true experiments (see Schmidt and Hunter 2015, chapters 6, 7, and 8).

15.1.2.1 Single Artifacts Most artifacts attenuate the population correlation ρ. The amount of attenuation

depends on the artifact. For each artifact, it is possible to present a conceptual definition that makes it possible to quantify the influence of the artifact on the observed effect size. For example, the reliability of the dependent variable calibrates the extent to which there is random error of measurement in the measure of the dependent variable. The reliability, and hence the artifact parameter that determines the influence of measurement error on effect size, can be empirically estimated. Journal editors should require authors to furnish those artifact values but often do not. Most of the artifacts cause a systematic attenuation of the correlation; that is, the expected value of the study correlation is lower than the actual correlation by some amount. This attenuation is usually most easily expressed as a product in which the actual correlation is multiplied by an artifact multiplier, usually denoted a.

We denote the actual (unattenuated) population correlation by ρ, and denote the (attenuated) study population correlation by ρ_o. Because we cannot conduct the study without measurement error, this study imperfection systematically biases the actual correlation parameter downward. Thus the study correlation ρ_o is smaller than the actual correlation ρ.

We denote by a_i the artifact value for the study expressed in the form of a multiplier. If the artifact parameter is expressed by a multiplier a_i, then

$$\rho_o = a_i \rho, \tag{15.1}$$

where a_i is some fraction, $0 < a_i < 1$. The size of a_i depends on the artifact: the greater the error, the smaller the value of a_i. In the developments that follow, these artifacts are described as they occur in correlation studies. However, each artifact has a direct analog in experimental studies (for detail, see Schmidt and Hunter 2015, chapters 6–8).

Attenuation artifacts and the corresponding multiplier are as follows:

1. Random error of measurement in dependent variable Y:

   $$a_1 = \sqrt{r_{YY}},$$

   where r_{YY} is the reliability of the measure of Y. Example: $r_{YY} = 0.49$ implies $a_i = 0.70$, $\rho_o = 0.70\rho$, a 30 percent reduction.

2. Random error of measurement in independent variable X:

   $$a_2 = \sqrt{r_{XX}},$$

   where r_{XX} is the reliability of the measure of X. Example: $r_{XX} = 0.81$ implies $a_2 = 0.90$, $\rho_o = 0.90\rho$, a 10 percent reduction.

3. Artificial dichotomization of continuous dependent variable split into proportions p and q:

   $$a_3 = \text{biserial constant} = \phi(c)\big/\sqrt{(pq)},$$

   where $\phi(x) = e^{-x^2/2}\big/\sqrt{2\pi}$ is the unit normal density function and where c is the unit normal distribution cut point corresponding to a split of p. That is, $c = \phi^{-1}(p)$, where $\phi(x)$ is the unit normal cumulative distribution function (Hunter and Schmidt 1990).

   Example: Y is split at the median, $p = q = 0.5$, $a_3 = 0.80$, $\rho_o = 0.80\rho$, a 20 percent reduction.

4. Artificial dichotomization of continuous independent variable split into proportions p and q:

   $$a_4 = \text{biserial constant} = \phi(c)\big/\sqrt{(pq)},$$

   where c is the unit normal distribution cut point corresponding to a split of p and $\phi(x)$ is the unit normal density function. That is, $c = \phi^{-1}(p)$ (Hunter and Schmidt 1990).

   Example: the continuous measure X is split such that $p = 0.9$ and $q = 0.1$. Then $a_4 = 0.60$, $\rho_o = 0.60\rho$, a 40 percent reduction.

5. Imperfect construct validity of the dependent variable Y. Construct validity is the correlation of the dependent variable measure with the actual dependent variable construct:

   $$a_5 = \text{the construct validity of } Y.$$

   Example: supervisor ratings of job performance, $a_5 = 0.72$ (Viswesvaran, Ones, and Schmidt 1996); mean inter-rater reliability of supervisor rating is 0.52; the correlation between the supervisor rating and job performance true scores is $\sqrt{0.52} = 0.72$ (see also Rothstein 1990) $\rho_o = 0.72\rho$, a 28 percent reduction.

6. Imperfect construct validity of the independent variable *X*. Construct validity is defined as in (5):

   a_6 = the construct validity of *X*.

   Example: use of a perceptual speed measure to measure general cognitive ability, $a_6 = 0.65$ (the true score correlation between perceptual speed and general cognitive ability is 0.65).

   $\rho_o = 0.65\rho$, a 35 percent reduction.

7. Range restriction on the independent variable *X*. Range restriction results from systematic exclusion of certain scores on *X* from the sample compared with the relevant (or reference) population:

   a_7 depends on the standard deviation (*SD*) ratio,
   u_X = (*SD*$_X$ *study population*)/
   (*SD*$_X$ *reference population*).

   For example, the average value of u_x among employees for general cognitive ability has been found to be 0.67.

   Range restriction can be either direct (explicit truncation of the *X* distribution) or indirect (partial or incomplete truncation of the *X* distribution, usually resulting from selection on some variable correlated with *X*). Volunteer bias in study subjects, a form of self-selection, can produce indirect range restriction. Most range restriction in real data is indirect (Hunter, Schmidt, and Le 2006). The order in which corrections are made for range restriction and measurement error differs for direct and indirect range restriction. Correcting for direct range restriction when the restriction has actually been indirect results in estimates that are downwardly biased, typically by about 25 percent.

   For both types of range restriction, the size of the multiplier depends on the size of ρ.

   For direct range restriction the formula is:

$$a_7 = u_x \Big/ \sqrt{\left(u_X^2\rho^2 + 1 - \rho^2\right)}.$$

   Example: For $\rho = 0.20$ and $u_X = 0.67$, $a_7 = 0.68$. $\rho_o = 0.68\rho$, a 32 percent reduction.

   The formula for indirect range restriction is more complicated because it involves correlations between the third variable *Z* where selection (explicit truncation) occurred and both *X* and *Y*. Unfortunately, information regarding these correlations is rarely available (Hunter, Schmidt, and Le 2006). One practical method of correcting for indirect range restriction was introduced by John Hunter and Frank Schmidt (2004, chapter 5) and presented in the second edition of this handbook (Schmidt, Le, and Oh 2009). That method relies on the assumption that the effect of *Z* on *Y* is fully mediated by *X*. This assumption is likely to be met in many practical situations (Hunter, Schmidt, and Le 2006; see also Le and Schmidt 2006). Even when this assumption is violated, as long as range restriction is indirect (almost always the case), this method still provides more accurate estimates of the true correlation than the direct range restriction correction method or no correction at all (Le and Schmidt 2006). The formula for this method is

$$a_7 = u_T \Big/ \sqrt{\left(u_T^2\rho^2 + 1 - \rho^2\right)},$$

   where u_T is the ratio of restricted to unrestricted true score standard deviations (computed in Schmidt and Hunter 2015, equation 3.16; or in Hunter, Schmidt, and Le 2006, equation 22):

$$u_T^2 = \left\{u_X^2 - \left(1 - r_{XX_a}\right)\right\} \Big/ r_{XX_a}$$

   Note that in this equation, r_{XX_a} is the reliability of the independent variable X estimated in the *unrestricted* group. Hunter and his colleagues use subscript *a* to denote values estimated in the unrestricted group, and subscript *i* to denote the restricted group (Hunter, Schmidt, and Le 2006). We use the same notation in this chapter.

   Example: For $\rho = 0.20$ and $u_T = 0.56$, $a_7 = 0.57$. $\rho_o = 0.57\rho$, a 43 percent reduction.

   There has been a significant development since the last edition of this handbook regarding a new correction method for indirect range restriction. Specifically, Huy Le and his colleagues introduced a more accurate correction approach that does not require the assumption underlying the earlier method just discussed (Le et al. 2016). The new method, however, requires knowledge of range restriction on *Y*. For situations in which that information cannot be obtained (for example, all personnel selection studies), the Hunter, Schmidt, and Le method provides

reasonably accurate estimates of the effect of indirect range restriction (Hunter, Schmidt, and Le 2006). The basic formula of the new method is:

$$a_7 = \frac{1}{u_T u_P}\left(1 - \frac{\sqrt{(1-u_T^2)(1-u_P^2)}}{\rho}\right)$$

where u_T is the ratio of restricted true score standard deviation of X to its unrestricted true scores standard deviation, as described (Le et al. 2016). The ratio of restricted true score standard deviation of Y to its unrestricted true score standard deviation, u_P, can be computed via the following equation:

$$u_P^2 = \left\{u_Y^2 - (1 - r_{YY_a})\right\} / r_{YY_a}.$$

In this equation, r_{YY_a} is the reliability of Y in the unrestricted group; u_Y is the raw score range restriction ratio on Y.

Example: For $\rho = 0.20$ and $u_T = 0.60$, $u_P = 0.98$, then $a_7 = 0.35$. $\rho_o = 0.35\rho$, a 65 percent reduction.

8. Range restriction on the dependent variable Y. Range restriction results from systematic exclusion of certain scores on Y from the sample in comparison to the relevant population:

   a_8 depends on the SD ratio,
   $u_Y = (SD_Y \textit{ study population}) / (SD_Y \textit{ reference population})$.

   Example: Some workers are fired early for poor performance and are hence underrepresented in the incumbent population. If exactly the bottom 20 percent of performers are fired, this is a condition of direct range restriction. Then from normal curve calculations, $u_Y = 0.83$.

   The size of the multiplier depends on the size of ρ.

$$a_8 = u_Y / \sqrt{(u_Y^2 \rho^2 + 1 - \rho^2)}.$$

   Example: For $\rho = 0.20$ and $u_Y = 0.83$, $a_8 = 0.84$, $\rho_o = 0.84\rho$, a 16 percent reduction.

   Note: Correction of the same correlation for range restriction on both the independent and dependent variables is complicated and requires special formulas (see Schmidt and Hunter 2015, 47–50; Le et al. 2016).

9. Bias in the correlation coefficient. The correlation has a small negative bias:

$$a_9 = 1 - (1 - \rho^2)/(2N - 2).$$

   Comment: For sample sizes of twenty or more, bias is smaller than rounding error. Thus, bias is usually trivial in size, as illustrated in the following example.

   Example: For $\rho = 0.20$ and $N = 68$, $a_9 = 0.9997$, $\rho_o = 0.9997\rho$, a .03 percent reduction.

10. Study-caused variation (covariate-caused confounds).

    Example: Concurrent test validation studies conducted on ability tests evaluate the job performance of incumbent workers who vary in job experience, whereas applicants all start with zero job experience. Job experience correlates with job performance, even holding ability constant.

    Solution: Use partial correlation to remove the effects of unwanted variation in job experience.

    Specific case: Study done in new plant with very low mean job experience (for example, mean = 2 years). Correlation of experience with ability is zero. Correlation of job experience with job performance is 0.50. Comparison of the partial correlation to the zero order correlation shows that $a_{10} = \sqrt{1 - .50^2} = 0.87$. $\rho_o = 0.87\rho$, a 13 percent reduction.

15.1.2.2 Multiple Artifacts Suppose that the study correlation is affected by several artifacts with parameters, a_1, a_2, a_3, · · · . The first artifact reduces the actual correlation from ρ to $\rho_{o1} = a_1\rho$.

The second artifact reduces that correlation to $\rho_{o2} = a_2\rho_{o1} = a_2(a_1\rho) = a_1a_2\rho$.

The third artifact reduces that correlation to $\rho_{o3} = a_3\rho_{o2} = a_3(a_1a_2\rho) = a_1a_2a_3\rho$, and so on.

Thus, the joint effect of the artifacts is to multiply the population correlation by all the multipliers. For m artifacts

$$\rho_{om} = (a_1a_2a_3 \cdots a_m)\rho$$

or

$$\rho_o = A\rho, \qquad (15.2)$$

where A is the compound artifact multiplier equal to the product of the individual artifact multipliers:

$$A = a_1 a_2 a_3 \cdots a_m.$$

15.1.2.3 A Numerical Illustration We now illustrate the impact of some of these artifacts using actual data from a large program of studies of the validity of a personnel selection test. The quantitative impact is large. Observed correlations can be less than half the size of the estimated population correlations based on perfect measures and computed on the relevant (unrestricted reference) population. We give two illustrations: the impact on the average correlation between general cognitive ability and job performance ratings and the impact of variation in artifacts across studies.

A meta-analysis of 425 validation studies conducted by the U.S. Employment Service shows that for medium-complexity jobs, the average applicant population correlation between true scores on general cognitive ability and true scores job performance ratings is 0.73 (Hunter, Schmidt, and Le 2006).[1] We now show how this value is reduced by study artifacts to an observed mean correlation of 0.27.

The ideal study. Ideally, each worker would serve under a population of judges (raters), so that idiosyncrasy of judgment could be eliminated by averaging ratings across judges.

Let the subscript P denote consensus (average) rating of performance by a population of judges, and let the subscript A denote actual cognitive ability.

The correlation r_{AP} would then be computed on an extremely large sample of applicants hired at random from the applicant population. Hence, there would be no unreliability in either measure, no range restriction, and virtually no sampling error. The Hunter, Schmidt, and Le meta-analysis (2006) indicates that the obtained correlation would be 0.73.

The actual study. In the actual study, the independent variable X = score on an imperfect test of general cognitive ability, and dependent variable Y = rating by one immediate supervisor. The correlation r_{XY} is computed on a small sample of range-restricted workers (incumbents) hired by the company based on information available at the time of hire.

Impact of restriction in range. Range restriction biases the correlation downward. All studies by the U.S. Employment Service were conducted in settings in which the General Aptitude Test Battery (GATB) had not been used to select workers, so this range restriction is indirect. When range restriction is indirect, its attenuating effects occur *before* the attenuating effects of measurement error occur (Hunter, Schmidt, and Le 2006). The average extent of restriction in range observed for the GATB was found to be (Hunter 1980):

$$u_X = (SD_X \text{ incumbent population})/(SD_X \text{ applicant population}) = 0.67$$

This observed range restriction u_X translates into a true score u_T value of 0.56 (based on Schmidt and Hunter 2015, equation 3.16; Hunter, Schmidt, and Le 2006, equation 22). This calculation is based on the reliability of the GATB in the unrestricted population, r_{XX_a}, which is 0.81. Further using the equation provided earlier in section 1.2.1(7), we obtain the value of $a_7 = .70$, a 30 percent reduction. Note: in a selection study, the Le et al. (2016) range restriction correction cannot be used, because the range restriction ratio on Y (job performance) is unknown; hence we use the Hunter, Schmidt, and Le (2006) method in our example here.

Impact of measurement error. The value of r_{XX}, the reliability of the GATB, is 0.81 in the unrestricted population, but in the restricted group this value is reduced to 0.58, computed using the following equation: $r_{XX_i} = 1 - (1 - r_{XX_a})/u_X^2$ (Hunter, Schmidt, and Le 2006, equation 27). Consequently, the attenuation multiplier due to measurement error in the independent variable is $a_2 = \sqrt{0.58} = 0.76$. The value of r_{YY}, the reliability of the dependent variable (supervisory ratings of job performance), is 0.50 in the restricted group (Viswesvaran, Ones, and Schmidt 1996). This translates into the attenuation multiplier of 0.71 $\left(a_1 = \sqrt{0.50} = 0.71\right)$.

Taken together, the combined attenuating effect due to indirect range restriction and measurement error in both the independent and dependent variable measures is

$$A = a_1 a_2 a_7 = (0.71)(0.76)(0.70) = 0.38.$$

Hence the expected value of the observed r is

$$r_{XY} = A r_{AP} = a_1 a_2 a_7 r_{AP} = (0.71)(0.76)(0.70)(0.73) = 0.27.^2$$

The total impact of study limitations was to reduce the mean population correlation from 0.73 to 0.27, a reduction of 63 percent.

Now we present numerical examples of the effects of variation in artifacts across studies. Again, here we assume a correlation of 0.73 in the applicant population for perfectly measured variables. Even if there was no true variation in this value across studies (for example, different employers), variation in artifact values would produce substantial variation in observed correlations even in the absence of sampling error.

Indirect range restriction. None of the firms whose data were available for the study used the GATB in hiring. Different firms used a wide variety of hiring methods, including cognitive ability tests other than the GATB. Thus, range restriction on the GATB is indirect. Suppose that the composite hiring dimension used (symbolized as *S*) correlated on average 0.90 with GATB true scores, that is, $\rho_{ST} = 0.90$ (see Hunter, Schmidt, and Le 2006, figure 1). Further assume that some firms are very selective, accepting only the top 5 percent on their composite, $u_S = 0.37$ (calculated based on the procedure presented in Schmidt, Hunter, and Urry 1976). Other firms are relatively lenient, selecting half of the applicants, $u_S = 0.60$. Explicit selection on *S* results in range restrictions on true score *T* of the test (*X*), which effect varies from $u_T = 0.55$ (when $u_S = 0.37$) to $u_T = 0.69$ (when $u_S = 0.60$) (calculations based on Hunter, Schmidt, and Le 2006, equation 18):

$$u_T^2 = \rho_{ST_a}^2 u_S^2 - \rho_{ST_a}^2 + 1.$$

This creates an attenuating effect ranging from $a_7 = 0.69$ (31 percent reduction) to $a_7 = 0.82$ (18 percent reduction). The correlations would then vary from $r_{XY} = 0.51$ to $r_{XY} = 0.60$.

Predictor reliability. Suppose that each study is conducted using either a long or a short ability test. Assume that the reliability for the long test is $r_{XX} = 0.81$ and that for the short test is 0.49 in the unrestricted population. Due to indirect range restriction, reliabilities of the tests in the samples will be reduced. The following equation allows calculation of the restricted reliability (r_{XXi}) from the unrestricted reliability (r_{XXa}) and the range restriction ratio on T (u_T) (Hunter, Schmidt, and Le 2006, equations 25 and 26):

$$r_{XX_i} = \frac{u_T^2 r_{XX_a}}{u_T^2 r_{XX_a} + 1 - r_{XX_a}}.$$

Based on this equation, observed reliabilities for the long test will vary from to 0.56 (when $u_T = 0.55$) to 0.67 (when $u_T = 0.69$); for the short test, reliabilities will be 0.22 (when $u_T = 0.55$) and 0.32 (when $u_T = 0.69$). The corresponding correlations would be as follows:

1. High-range restriction ratio, long test: $r_{XY} = \sqrt{0.67}(0.60) = 0.49$,
2. High-range restriction ratio, short test: $r_{XY} = \sqrt{0.32}(0.60) = 0.33$
3. Low-range restriction ratio, long test: $r_{XY} = \sqrt{0.56}(0.51) = 0.38$,
4. Low-range restriction ratio, short test: $r_{XY} = \sqrt{0.22}(0.51) = 0.24$.

Criterion reliability: one rater versus two raters. Suppose that in some studies one supervisor rates job performance and in other studies there are two raters. Interrater reliability is 0.50 for one rater and (by the Spearman-Browne formula) is 0.67 for ratings based on the average of two raters. Criterion reliability would then be either $r_{YY} = 0.50$ or $r_{YY} = 0.67$. Consequently, the observed correlations would be as follows:

1. High-range restriction ratio, long test, two raters: $r_{XY} = \sqrt{0.67}(0.49) = 0.40$
2. High-range restriction ratio, long test, one rater: $r_{XY} = \sqrt{0.50}(0.49) = 0.35$
3. High-range restriction ratio, short test, two raters: $r_{XY} = \sqrt{0.67}(0.33) = 0.27$
4 High-range restriction ratio, short test, one rater: $r_{XY} = \sqrt{0.50}(0.33) = 0.23$
5. Low-range restriction ratio, long test, two raters: $r_{XY} = \sqrt{0.67}(0.38) = 0.31$
6. Low-range restriction ratio, long test, one rater: $r_{XY} = \sqrt{0.50}(0.38) = 0.27$,
7. Low-range restriction ratio, short test, two raters: $r_{XY} = \sqrt{0.67}(0.24) = 0.20$
8. Low-range restriction ratio, short test, one rater: $r_{XY} = \sqrt{0.50}(0.24) = 0.17$

Thus, variation in artifacts produces variation in study population correlations. Instead of one population correlation of 0.73, we have a distribution of attenuated population correlations: 0.17, 0.20, 0.23, 0.27, 0.27, 0.31, 0.35, and 0.40. Each of these values would be the population correlation underlying a particular study. To that value random sampling error would then be added to yield the correlation observed in the study. In this

example, we have assumed a single underlying population value of 0.73. If there were variation in population correlations prior to the introduction of these artifacts, that variance would be increased because the process illustrated here applies to each value of the population correlation.

15.2 CORRECTING FOR ATTENUATION-INDUCED BIASES

15.2.1 The Population Correlation: Attenuation and Disattenuation

The population correlation can be exactly corrected for the effect of any artifact. The exactness of the correction follows from the absence of sampling error. Because $\rho_o = A\rho$, we can reverse the equation algebraically to obtain

$$\rho = \rho_o / A. \tag{15.3}$$

Correcting the attenuated population correlation produces the value the correlation would have had if it had been possible to conduct the study without the methodological limitations produced by the artifacts. To divide by a fraction is to increase the value. That is, if artifacts reduce the study population correlation, then the corresponding disattenuated (corrected) correlation must be larger than the observed correlation.

15.2.2 The Sample Correlation

The sample correlation can be corrected using the same formula as for the population correlation. This eliminates the systematic error in the sample correlation, but it does not eliminate the sampling error. In fact, sampling error is increased by the correction.

The sample study correlation relates to the (attenuated) population correlation by

$$r_o = \rho_o + e, \tag{15.4}$$

where e is the sampling error in r_o (the observed correlation). To within a close approximation, the average error is zero (Hedges 1989) and the sampling error variance is

$$Var(e) = \left(1 - \rho_o^2\right)^2 / (N - 1). \tag{15.5}$$

The corrected sample correlation is

$$r_c = r_o / A, \tag{15.6}$$

where A is the compound artifact multiplier. The sampling error in the corrected correlation is related to the population correlation by

$$\begin{aligned} r_c = r_o / A &= (\rho_o + e)/A \\ &= (\rho_o / A) + (e/A) \\ &= \rho + e'. \end{aligned} \tag{15.7}$$

That is, the corrected correlation differs from the actual effect size correlation by only sampling error e', where the new sampling error e' is given by

$$e' = e/A. \tag{15.8}$$

Because A is less than 1, the sampling error e' is larger than the sampling error e. This can be seen in the sampling error variance

$$Var(e') = Var(e)/A^2. \tag{15.9}$$

However, because the average error e is essentially zero, the average error e' is also essentially zero (see Schmidt and Hunter 2015, chapter 3).

15.3 META-ANALYSIS OF CORRECTED CORRELATIONS AND SOFTWARE

The meta-analysis methods described in this and the following sections are based on random-effects models (Schmidt, Oh, and Hayes 2009). These procedures are implemented in the Schmidt and Le (2014) Windows-based software and have been shown in simulation studies to be accurate (for example, see Field 2005; Hall and Brannick 2002; Law, Schmidt, and Hunter 1994; Schulze 2004).

If study artifacts are reported for each study, then for each study, we have three numbers.

For study i we have

r_i = the ith study correlation,

A_i = the compound artifact multiplier for study i, and

N_i = the sample size for study i.

We then compute for each study the disattenuated correlation:

r_{ci} = the disattenuated correlation for study i.

(If some of the studies do not provide some of the artifact information, the usual practice is to fill in this missing information with average values from the other studies.)

Two meta-analyses can be computed: one on the biased (attenuated) study correlations (the partial meta-analysis) and one on the corrected (unbiased) correlations.

15.3.1 The Mean Corrected Correlation

Large-sample studies contain more information than small-sample studies and thus should be given more weight (Schmidt and Hunter 1977; Hedges and Olkin 1985). Studies are therefore often weighted by sample size, or the inverse of their sampling error variance, which is nearly equivalent. For corrected correlations, a more complicated weighting formula is recommended that takes into account the other artifact values for the study—for example, the more measurement error there is in the measures of the variables, the less the information there is in the study (Schmidt and Hunter 2015, chapter 3). Thus, a high-reliability study should be given more weight than a low-reliability study (see also Hedges and Olkin 1985, 135–36).

The weight for study i should be

$$w_i = N_i A_i^2, \tag{15.10}$$

where A_i is the compound artifact multiplier for study I (Schmidt and Hunter 2015, chapter 3).

The average correlation can be written

$$Ave(r) = \sum w_i r_i \Big/ \sum w_i, \tag{15.11}$$

where

$w_i = 1$ for the unweighted average,

$w_i = N_i$ for the sample size weighted average, and

$w_i = N_i A_i^2$ for the full artifact weighted average (applied to corrected correlations).

If the number of studies were infinite (so that sampling error would be completely eliminated), the resulting mean would be the same regardless of which weights were used. But for a finite number of studies, meta-analysis does not totally eliminate sampling error; there is still some sampling error left in the mean correlation. Use of the full artifact weights described here minimizes sampling error in the mean corrected correlation.

15.3.2 Corrected Versus Uncorrected Correlations

In some research domains, the artifact values for most individual studies are not presented in those studies. As a result, some published meta-analyses do not correct for artifacts. Failure to correct means that the mean uncorrected correlation will be downwardly biased as an estimate of the actual (unattenuated or construct level) correlation. The amount of bias in a meta-analysis of uncorrected correlations will depend on the extent of error caused by artifacts in the average study. This average extent of systematic error is measured by the average compound multiplier $Ave(A)$.

To a close statistical approximation, the mean corrected correlation $Ave(r_c)$ relates to the mean uncorrected correlation $Ave(r)$ in much the same way as does an individual corrected correlation. Just as for a single study

$$r_c = r/A, \tag{15.12}$$

so to a close approximation we have for a set of studies

$$Ave(r_c) = Ave(r)/Ave(A). \tag{15.13}$$

Thus, to a close approximation, the difference in findings of an analysis that does not correct for artifacts and one that does is the difference between the uncorrected mean correlation $Ave(r)$ and the corrected mean correlation $Ave(r_c)$ (Schmidt and Hunter 2015, chapter 4).

15.3.3 Variance of Corrected Correlations: Procedure

The variance of observed correlations greatly overstates the variance of population correlations. This is true for corrected correlations as well as for uncorrected correlations. From the fact that the corrected correlation is

$$r_{c_i} = \rho_i + e_i'. \tag{15.14}$$

where r_{ci} and ρ_i are the corrected sample and population correlations, respectively, and e'_i is the sampling error, we have the decomposition of variance

$$Var(r_c) = Var(\rho) + Var(e').$$

Thus, by subtraction, we have an estimate of the desired variance

$$Var(\rho) = Var(r_c) - Var(e'). \tag{15.15}$$

The variance of study corrected correlations is the weighted squared deviation of the *i*th correlation from the mean correlation. If we denote the average corrected correlation by $\overline{r}_c$ then

$$\overline{r}_c = Ave(r_c) = \sum w_i r_{c_i} \Big/ \sum w_i, \tag{15.16}$$

$$Var(r_c) = \sum w_i (r_{c_i} - \overline{r}_c)^2 \Big/ \sum w_i. \tag{15.17}$$

The sampling error variance is computed by averaging the sampling error variances of the individual studies. The error variance of the individual study depends on the size of the uncorrected population correlation. To estimate that number, we first compute the average uncorrected correlation $\overline{r}$.

$$\overline{r} = Ave(r) = \sum w_i r_i \Big/ \sum w_i.$$

where the w_i are the sample sizes N_i.

For study *i*, we have

$$Var(e_i) = v_i = (1 - \overline{r}^2)^2 \Big/ (N_i - 1), \tag{15.18}$$

and

$$Var(e'_i) = v'_i = Var(e_i) \Big/ A_i^2 = v_i \Big/ A_i^2.$$

For simplicity, denote the study sampling error variance $Var(e'_i)$ by v'_i. The weighted average error variance for the meta-analysis is the average

$$Var(e') = \sum w_i v'_i \Big/ \sum w_i. \tag{15.19}$$

where the $w_i = N_i A_i^2$.

Procedure. The specific computational procedure involves six steps:

1. Given for each study r_i = uncorrected correlation, A_i = compound artifact multiplier, and N_i = sample size,
2. Compute for each study r_{ci} = corrected correlation, and w_i = the proper weight to be given to r_{c_i}.
3. To estimate the effect of sampling error, compute the average uncorrected correlation $\overline{r}$. This is done using weights $w_i = N_i$.
4. For each study compute the sampling error variance: v'_i = the sampling error variance.
5. The meta-analysis of disattenuated correlations includes four steps:
   (a) Compute the mean corrected correlation using weights $w_i = N_i A_i^2$: Mean corrected correlation = $Ave(r_c)$.
   (b) Compute the variance of corrected correlations using $w_i = N_i A_i^2$: Variance of corrected correlations = $Var(r_c)$.
   (c) Compute the sampling error variance $Var(e')$ by averaging the individual study sampling error variances:

$$Var(e') = Ave(v'_i). \tag{15.20}$$

   (d) Now compute the estimate of the variance of population correlations by subtracting out sampling error:

$$Var(\rho) = Var(r_c) - Var(e').$$

6. The final fundamental estimates (the average and the standard deviation (*SD*) of ρ) are

$$Ave(\rho) = Ave(r_c) \tag{15.21}$$

and

$$SD_\rho = \sqrt{Var(\rho)}. \tag{15.22}$$

As mentioned earlier, software is available for conducting these calculations (Schmidt and Le 2014).[3] A procedure similar to the one described here has also been presented by Nambury Raju and his colleagues (1991).

Simplified examples of application of the approach described earlier to meta-analysis are presented in chapter 3 of *Methods of Meta-Analysis* (Schmidt and Hunter 2015). Numerous meta-analyses of this sort have been published (see, for example, Carlson et al. 1999; Judge et al. 2001; Rothstein et al. 1990).

As an example, in one study, a previously developed weighted biodata form was correlated with promotion or advancement (with years of experience controlled) for 7,334 managers in twenty-four organizations (Carlson et al. 1999). Thus there were twenty-four studies, with a mean N per study of 306. The reliability of the dependent variable—rate of advancement or promotion rate—was estimated at 0.90. The standard deviation (*SD*) of the independent variable was computed in each organization, and the *SD* of applicants (that is, the unrestricted *SD*) was known, allowing each correlation to be corrected for range variation. In this meta-analysis, there was no between-studies variation in correlations due to variation in measurement error in the independent variable-because the same biodata scale was used in all 24 studies. Also, in this meta-analysis, the interest was in the effectiveness of this particular biodata scale in predicting managerial advancement. Hence, mean ρ was not corrected for unreliability in the independent variable.

The results were as follows:

$$Ave(r_i) = 0.48,$$
$$Ave(\rho) = 0.53,$$
$$Var(\rho) = Var(r_c) - Var(e') = 0.00462 - 0.00230 = 0.00232$$

and

$$SD_\rho = \sqrt{Var(\rho)} = 0.048.$$

Thus the mean operational validity of this scale across organization was estimated as 0.53, with a standard deviation of 0.048. After correcting for measurement error, range variation, and sampling error, only a small variability in correlation across organizations is apparent. If we assume a normal distribution for ρ, the value at the 10th percentile is $0.53–1.28 \times 0.048 = 0.47$. Thus, the conclusion is that the correlation is at least 0.47 in 90 percent of these (and comparable) organizations. The value at the 90th percentile is 0.59, yielding an 80 percent credibility interval of 0.47 to 0.59, indicating that an estimated 80 percent of population values of validity lie in this range. Although the computation procedures used are not described in this chapter, confidence intervals can be placed around the mean validity estimate. In this case, the 95 percent confidence interval for the mean is 0.51 to 0.55. Confidence intervals and credibility intervals are different, however, and serve different purposes (Schmidt and Hunter 2015, 228–31). Confidence intervals refer only to the estimate of the mean, whereas credibility intervals are based on the estimated distribution of all of the population correlations. Hence confidence intervals are based on the (estimated) standard error of the mean, and credibility intervals are based on the (estimated) standard deviation of the population correlations.

15.4 ARTIFACT DISTRIBUTION META-ANALYSIS AND SOFTWARE

In most contemporary research domains, the artifact values are not provided in many of the studies. Instead, artifact values are presented only in a subset of studies, usually a different but overlapping subset for each individual artifact. Meta-analysis can be conducted in such domains, although the procedures are more complex.

Simplified examples of application of artifact distribution meta-analysis are presented in chapter 4 of *Methods of Meta-Analysis* (Schmidt and Hunter 2015). Many published meta-analyses have been based on these artifact distribution meta-analysis methods (see Schmidt and Hunter 2015, chapters 1 and 4). A subset of these have been conducted on correlation coefficients representing the validities of various kinds of predictors of job performance—usually tests, but also interviews, ratings of education and job experience, assessment centers, and others. The implications of the findings of these meta-analyses for personnel selection practices have been quite profound and are described in various studies (see Schmidt, Hunter, and Pearlman 1980; Pearlman, Schmidt, and Hunter 1980; Schmidt and Hunter 1981; Schmidt et al. 1985; Schmidt, Hunter, and Raju 1988; Schmidt, Ones, and Hunter 1992; Schmidt and Hunter 1998, 2003; Schmidt, Oh, and Le 2006; Schmidt, Shaffer, and Oh 2008; McDaniel, Schmidt, and Hunter 1988; Le and Schmidt 2006; McDaniel et al. 1994). Artifact distribution meta-analyses have also been conducted in a variety of other research areas, such as role conflict, leadership, effects of goal setting, and work-family conflict. More than two hundred such nonselection meta-analyses have appeared in the literature to date.

The artifact distribution meta-analysis procedures described in this section are implemented in the Schmidt

and Le software (2014), although the methods used in these programs to estimate the standard deviation of the population corrected correlations are slightly different from those discussed later. Although the methods used in the Schmidt-Le programs for this purpose are slightly more accurate (as shown in simulation studies) than those described in the remainder of this chapter, they contain a number of statistical refinements, making them more complex—in fact, too complex to describe easily in a chapter of this sort. These methods are fully described in *Methods of Meta-Analysis* (Schmidt and Hunter 2015, chapter 4). Similar artifact distribution methods for meta-analysis have been presented in other studies (Callender and Osburn 1980; Raju and Burke 1983). In all these methods, the key assumption in considering artifact distributions is independence of artifact values across different artifacts. This assumption is plausible for the known artifacts in research domains that have been examined. The basis for independence is the fact that the resource limitations that produce problems with one artifact, such as range restriction, are generally different and hence independent of those that produce problems with another artifact, such as measurement error in scales used (Schmidt and Hunter 2015, chapter 4). Artifact values are also assumed to be independent of the true score correlation ρ_i. A 1998 computer simulation study found that violation of these independence assumptions has minimal effect on meta-analysis results unless the artifacts values are correlated with the ρ_i, a seemingly unlikely event (Raju et al. 1998).

We use the following notation for the correlations associated with the *i*th study:

ρ_i = the true (unattenuated) study population correlation;

r_{ci} = the study sample corrected correlation that can be computed if artifact information is available for the study so that corrections can be made;

r_{oi} = uncorrected (observed) study sample correlation; and

ρ_{oi} = uncorrected (attenuated) study population correlation.

In the previous section, we assumed that artifact information is available for every (or nearly every) study individually. In such a case, an estimate r_{ci} of the true correlation ρ_i can be computed for each study and meta-analysis can be conducted on these estimates. In this section, we assume that artifact information is missing for many or most studies. However, we assume that the distribution (or at least the mean and variance) of artifact values can be estimated for each artifact. The meta-analysis then proceeds in two steps:

- A "bare bones" or partial meta-analysis is conducted, yielding estimates of the mean and standard deviation of attenuated study population correlations. A bare bones meta-analysis is one that corrects only for sampling error.
- The mean and standard deviation from the bare bones meta-analysis are then corrected for the effects of artifacts other than sampling error.

15.4.1 The Mean of the Corrected Correlations

The attenuated study population correlation ρ_{oi} is related to the actual study population correlation ρ_i by the formula

$$\rho_{oi} = A_i \rho_i,$$

where A_i = the compound artifact multiplier for study i (which is unknown for most studies).

The sample attenuated correlation r_{oi} for each study is related to the attenuated population correlation for that study by

$$r_{oi} = \rho_{oi} + e_{oi},$$

where e_{oi} = the sampling error in study i (which is unknown).

15.4.1.1 Meta-Analysis of Attenuated Correlations The meta-analysis uses the additivity of means to produce

$$Ave(r_{oi}) = Ave(\rho_{oi} + e_{oi}) = Ave(\rho_{oi}) + Ave(e_{oi}).$$

If the number of studies is large, the average sampling error will tend to zero and hence

$$Ave(r_{oi}) = Ave(\rho_{oi}) + 0 = Ave(\rho_{oi}).$$

Thus, the bare-bones estimate of the mean attenuated study population correlation is the expected mean attenuated study sample correlation.

15.4.1.2 Correction of the Mean Correlation The attenuated population correlation for study i is related to the disattenuated correlation for study i by $\rho_{oi} = A_i \rho_i$, where A_i = the compound artifact multiplier for study i.

Thus, the mean attenuated correlation is given by

$$\text{Ave}(\rho_{oi}) = Ave(A_i\rho_i). \quad (15.23)$$

Because we assume that artifact values are independent of the size of the true correlation, the average of the product is the product of the averages:

$$Ave(A_i\rho_i) = Ave(A_i)Ave(\rho_i). \quad (15.24)$$

Hence, the average attenuated correlation is related to the average disattenuated correlation by

$$Ave(\rho_{oi}) = Ave(A_i)Ave(\rho_i),$$

where $Ave(A_i)$ = the *average* compound multiplier across studies.

We need not know all the individual study artifact multipliers, only the average. If the average multiplier is known, then the corrected mean correlation is

$$Ave(\rho_i) = Ave(r_{oi})/Ave(A_i). \quad (15.25)$$

15.4.1.3 The Mean Compound Multiplier To estimate the average compound multiplier, it is sufficient to be able to estimate the average for each single artifact multiplier separately. This follows from the independence of artifacts. To avoid double subscripts, let us denote the separate artifact multipliers by a, b, c, . . . The compound multiplier A is then given by the product $A = abc\ldots$

Because of the independence of artifacts, the average product is the product of averages:

$$Ave(A_i) = Ave(a)Ave(b)Ave(c)\ldots \quad (15.26)$$

Thus, the steps in estimating the compound multiplier are as follows:

1. Consider the separate artifacts.
   (a) Consider the first artifact *a.* For each study that includes a measurement of the artifact magnitude, denote the value a_i. Average those values to produce $Ave(a_i)$ = average of attenuation multiplier for first artifact.
   (b) Consider the second artifact *b*. For each study that includes a measurement of the artifact magnitude, denote the value b_i. Average those values to produce $Ave(b_i)$ = average of attenuation multiplier for second artifact.
   (c) Similarly, consider the other separate artifacts *c*, *d*, and so on, that produce estimates of the averages $Ave(c_i)$, $Ave(d_i)$. . . .

   The accuracy of these averages depends on the assumption that the available artifacts are a reasonably representative sample of all artifacts (see Hunter and Schmidt 2004). Note that even if this assumption is not fully met, results will still tend to be more accurate than those from a meta-analysis that does not correct for artifacts.
2. Compute the product

$$Ave(A_i) = Ave(a)Ave(b)Ave(c)Ave(d)\ldots. \quad (15.27)$$

15.4.2 Correcting the Standard Deviation

The method described in this section for estimating the standard deviation of the population true score correlations is the multiplicative method (Hunter and Schmidt 2004, chapter 4). This approach was first introduced by John Callender and Hobart Osburn and is the least complicated method to present (1980). However, other methods have also been proposed and used. Taylor Series-based methods, for example, have been used under conditions of direct and indirect range restriction (see, respectively, Raju and Burke 1983; Hunter, Schmidt, and Le 2006). Still another procedure is the interactive procedure (Schmidt, Gast-Rosenberg, and Hunter 1980; Schmidt and Hunter 2015, chapter 4). Simulation studies suggest that, with certain refinements, it is slightly more accurate than other procedures (Law, Schmidt, and Hunter 1994; Le and Schmidt 2006). It is therefore explicated most fully in *Methods of Meta-Analysis* (Schmidt and Hunter 2015, chapter 4) and incorporated in the Schmidt and Le software (2014). However, by the usual standards of social science research, all three procedures are quite accurate.

The meta-analysis of uncorrected correlations provides an estimate of the variance of attenuated study population correlations. However, these study population correlations are themselves uncorrected; that is, they are biased downward by the effects of artifacts. Furthermore, the variation in artifact levels across studies causes the study correlations to be attenuated by different amounts in different studies. This produces variation in the size of

the study correlations that could be mistaken for variation due to a real moderator variable, as we saw in our numerical example early in the chapter. Thus, the variance of population study correlations computed from a meta-analysis of uncorrected correlations is affected in two ways. The systematic artifact-induced reduction in the magnitude of the study correlations tends to decrease variability, and at the same time variation in artifact magnitude tends to increase variability across studies. Both sources of influence must be removed to accurately estimate the standard deviation of the disattenuated correlations across studies.

Let us begin with notation. The study correlation free of study artifacts (the disattenuated correlation) is denoted ρ_i, and the compound artifact attenuation factor for study i is denoted A_i. The attenuated study correlation, ρ_{oi} is computed from the disattenuated study correlation by

$$\rho_{oi} = A_i \rho_i. \tag{15.28}$$

The study sample correlation r_{oi} departs form the disattenuated study population correlation ρ_{oi} by sampling error e_i defined by

$$r_{oi} = \rho_{oi} + e_i = A_i \rho_i + e. \tag{15.29}$$

Consider now a bare-bones meta-analysis on the uncorrected correlations. We know that the variance of sample correlations is the variance of population correlations added to the sampling error variance. That is,

$$Var(r_{oi}) = Var(\rho_{oi}) + Var(e_i). \tag{15.30}$$

Because the sampling error variance can be computed by statistical formula, we can subtract it to yield

$$Var(\rho_{oi}) = Var(r_{oi}) - Var(e_i). \tag{15.31}$$

That is, the meta-analysis of uncorrected correlations produces an estimate of the variance of attenuated study population correlations, the actual study correlations after they have been reduced in magnitude by the study imperfections.

At the end of a meta-analysis of uncorrected (attenuated) correlations, we have the variance of attenuated study population correlations $Var(\rho_{oi})$, but we want the variance of actual disattenuated correlations $Var(\rho_i)$. The relationship between them is

$$Var(\rho_{oi}) = Var(A_i \rho_i). \tag{15.32}$$

We assume that A_i and ρ_i are independent. A formula for the variance of this product is given in Hunter and Schmidt (2004, chapter 4). Here we simply use this formula. Let us denote the average disattenuated study correlation by $\bar{\rho}$ and denote the average compound attenuation factor by $\bar{A}$. The variance of the attenuated correlations is given by

$$Var(A_i \rho_i) = \bar{A}^2 Var(\rho_i) + \bar{\rho}^2 Var(A_i) + Var(\rho_i) Var(A_i). \tag{15.33}$$

Because the third term on the right is negligibly small, to a close approximation

$$Var(A_i \rho_i) = \bar{A}^2 Var(\rho_i) + \bar{\rho}^2 Var(A_i). \tag{15.34}$$

We can then rearrange this equation algebraically to obtain the desired equation for the variance of actual study correlations free of artifact effects:

$$Var(\rho_i) = \left[Var(A_i \rho_i) - \bar{\rho}^2 Var(A_i) \right] \Big/ \bar{A}^2. \tag{15.35}$$

That is, starting from the meta-analysis of uncorrected correlations, we have

$$Var(\rho_i) = \left[Var(\rho_{oi}) - \bar{\rho}^2 Var(A_i) \right] \Big/ \bar{A}^2. \tag{15.36}$$

The right-hand side of equation 15.36 has four numbers:

1. $Var(\rho_{oi})$: the population correlation variance from the meta-analysis of uncorrected correlations, estimated using equation 15.31;
2. $\bar{\rho}$: the mean of disattenuated study population correlations, estimated using equation 15.25;
3. $\bar{A}$: the mean compound attenuation factor, estimated from equation 15.26.
4. $Var(A_i)$: the variance of the compound attenuation factor. This quantity has not yet been estimated.

15.4.2.1 Variance of the Artifact Multiplier How do we compute the variance of the compound attenuation

factor, A_i? We are given the distribution of each component attenuation factor, which must be combined to produce the variance of the compound attenuation factor. The key to this computation lies in two facts: (a) that the compound attenuation factor is the product of the component attenuation factors and (b) that the attenuation factors are assumed to be independent. That is, because $A_i = a_i b_i c_i \ldots$, the variance of A_i is

$$Var(A_i) = Var(a_i b_i c_i \ldots). \tag{15.37}$$

The variance of the compound attenuation factor is the variance of the product of independent component attenuation factors (for the formula, see Hunter and Schmidt 2004, 148–49). Here we simply use the result.

For each separate artifact, we have a mean and a standard deviation for that component attenuation factor. From the mean and the standard deviation, we can compute the coefficient of variation, which is the standard deviation divided by the mean. For our purposes here, we need a symbol for the squared coefficient of variation:

$$cv = [SD/Mean]^2. \tag{15.38}$$

For each artifact, we now compute the squared coefficient of variation. For the first artifact attenuation factor a, we compute

$$cv_1 = Var(a)/[Ave(a)]^2. \tag{15.39}$$

For the second artifact attenuation factor b, we compute

$$cv_2 = Var(b)/[Ave(b)]^2. \tag{15.40}$$

For the third artifact attenuation factor c, we compute

$$cv_3 = Var(c)/[Ave(c)]^2, \tag{15.41}$$

and so on. Thus, we compute a squared coefficient of variation for each artifact. These are then summed to form a total

$$CVT = cv_1 + cv_2 + cv_3 + \ldots \tag{15.42}$$

Recalling that $\bar{A}$ denotes the mean compound attenuation factor, we write the formula for the variance of the compound attenuation factor (to a close statistical approximation) as the product

$$Var(A_i) = \bar{A}^2 CVT. \tag{15.43}$$

We now have all the elements needed to estimate the variance in actual study correlations $Var(\rho_i)$. The final formula is

$$\begin{aligned} Var(\rho_i) &= \left[Var(\rho_{oi}) - \bar{\rho}^2 Var(A_i)\right]/\bar{A}^2 \\ &= \left[Var(\rho_{oi}) - \bar{\rho}^2 \bar{A}^2 CVT\right]/\bar{A}^2. \end{aligned} \tag{15.44}$$

The square root of this value is the SD_ρ. Hence we now have two main results of the meta-analysis: the mean of the corrected correlations, from equation 15.25; and the standard deviation of the corrected correlations, from equation 15.44. Using these values, we can again compute credibility intervals around the mean corrected correlation, as illustrated in the previous section. Also, we can compute confidence intervals around the mean corrected correlation (for a description of the methods, see Schmidt and Hunter 2015, 229–31).

For data sets for which the meta-analysis methods described earlier can be applied (that is, correlations can be corrected individually), the artifact distribution methods described in this section can also be applied. When this is done, the results are virtually identical, as expected (see Schmidt and Hunter 2015, chapter 4). Computer simulation studies also indicate that artifact distribution meta-analysis methods are quite accurate (for example, see Le and Schmidt 2006).

15.4.2.2 Decomposition of the Variance Inherent in the derivation in the previous section is a decomposition of the variance of uncorrected (observed) correlations. We now present that decomposition:

$$\begin{aligned} Var(r_{oi}) &= Var(\rho_{oi}) + Var(e_i), \\ Var(\rho_{oi}) &= Var(A_i \rho_i) = \bar{A}^2 Var(\rho_i) + \bar{\rho}^2 Var(A_i), \end{aligned} \tag{15.45}$$

and

$$Var(A_i) = \bar{A}^2 CVT. \tag{15.46}$$

That is,

$$Var(r_{oi}) = \bar{A}^2 Var(\rho_i) + \bar{\rho}^2 \bar{A}^2 CVT + Var(e_i)$$
$$= S1 + S2 + S3. \quad (15.47)$$

where $S1$ is the variance in uncorrected correlations produced by the variation in actual unattenuated effect size correlations, $S2$ is the variance in uncorrected correlations produced by the variation in artifact levels, and $S3$ is the variance in uncorrected correlations produced by sampling error.

In this decomposition, the term $S1$ contains the estimated variance of effect-size correlations. This estimated variance is corrected for those artifacts that were corrected in the meta-analysis. For reasons of feasibility, this does not usually include all the artifacts that affect the study value. (For example, the unsystematic artifact of data errors is rarely correctable.) Thus, $S1$ is an upper-bound estimate of the component of variance in uncorrected correlations due to real variation in the strength of the relationship and not due to artifacts of the study design. To the extent that there are uncorrected artifacts, $S1$ will overestimate the real variation; it may even greatly overestimate that variation. As long as there are uncorrected artifacts, there will be artifactual variation in study correlations produced by variation in those uncorrected artifacts.

15.5 SECOND-ORDER META-ANALYSIS

15.5.1 Need and Purpose of Second-Order Meta-Analysis

Increasingly today, multiple meta-analyses on the same question appear in the literature in diverse fields (for example, psychology, medicine, management), creating the need for methods of synthesizing multiple meta-analyses. As more and more meta-analyses are conducted, this need will only increase in the future. For example, at least a dozen meta-analyses have been undertaken on the relationships between the Big Five personality traits and job performance; some are independent because they were conducted in different countries based on national literatures. These meta-analyses often do not report the same results for the same relationship (for example, self-reports of conscientiousness and supervisor-ratings of job performance), leaving some doubt or ambiguity about the most trustworthy estimate for that relationship. There are three options to choose from if we decide to synthesize results across the first-order meta-analyses conducted on the same relationship (see Schmidt and Oh 2013; Borenstein et al. 2010, 184–86).

The first option is to conduct a full-scale meta-analysis by identifying, coding, pooling all primary studies included in all prior first order meta-analyses. Although ideal, this is typically not a practical solution given that many primary studies (in particular, unpublished studies) included in prior meta-analyses may be unavailable. The second option is to aggregate (that is, average) mean effect sizes across first order meta-analyses of interest while ignoring the between-meta-analysis variance. However, this option does not allow either estimation of the amount of true (that is, non-artifactual) variance between meta-analyses means or estimation of the amount of observed variation across meta-analyses due to second-order sampling error (given that the total number of primary studies in any first-order meta-analysis is less than infinite, the meta-analytic process does not reduce sampling error to zero; the remaining sampling error is called second-order sampling error). The third option is to combine mean effect sizes across meta-analyses of interest while modeling the between-meta-analysis variance. This is the best option when primary studies from all relevant first-order meta-analyses are unavailable and there is a need to estimate the between-meta-analysis variance. This approach overcomes the problems in options 1 and 2. This option is called second-order meta-analysis, also known as overview of reviews, umbrella review, meta-meta-analysis, and meta-analysis of meta-analyses (for example, Cooper and Koenka 2012, 446).

This section presents an introduction to statistical methods for second order meta-analysis that model between-meta-analysis variation (for equation origins, see Schmidt and Oh 2013). More details (including artifact-distribution methods), illustrative examples, and answers to several potential objections to second order meta-analysis are found in a separate study (Schmidt and Oh 2013). Put simply, the statistical methods of second order meta-analysis are a straightforward generalization of first order random effects (RE) meta-analysis methods (Schmidt, Oh, and Hayes 2009) to the synthesis of the meta-analytic mean effect-size estimates across multiple relevant meta-analyses.

Basic equations and principles of first order meta-analysis can be generalized to second order meta-analysis. We introduce two meta-analysis methods. First is a method without corrections for the biases created by measurement error, that

is, bare bones second-order meta-analysis, which corrects only for the effects of second order sampling error. Second is a method that includes corrections for both second order sampling error and the biasing effects of measurement error, that is, construct-level second-order meta-analysis. For convenience of presentation, these methods are illustrated using the metric of the correlation coefficient; analogous equations exist for other effect size indices, such as *d* values.

15.5.2 Bare Bones Second-Order Meta-Analysis

Suppose that *m* independent meta-analyses have been conducted to estimate the same relationship. Equation 15.48 is the fundamental equation when the first-order meta-analyses entering the second-order meta-analysis have corrected only for sampling error:

$$\hat{\sigma}^2_{\bar{\hat{\rho}}_{xy}} = S^2_{\hat{\bar{r}}} - E\left(S^2_{e_{\hat{\bar{r}}_i}}\right), \tag{15.48}$$

where the term on the left side of the equation is the estimate of the population variance of the observed-uncorrected meta-analytic mean correlations ($\hat{\rho}_{xy}$) across the *m* (first-order) meta-analyses after second-order sampling error has been subtracted.

The first term on the right side of equation 15.48 is the weighted variance of the mean meta-analytic correlations across the *m* meta-analyses, computed as follows:

$$S^2_{\hat{\bar{r}}} = \sum_1^m w_i \left(\hat{\bar{r}}_i - \hat{\bar{\bar{r}}}\right)^2 \Big/ \sum_1^m w_i , \tag{15.49a}$$

where

$$\hat{\bar{\bar{r}}} = \sum_1^m w_i \hat{\bar{r}}_i \Big/ \sum_1^m w_i ; \tag{15.49b}$$

and

$$w_i = \left(\frac{S^2_{r_i}}{k_i}\right)^{-1}, \tag{15.49c}$$

and where $S^2_{r_i}$ is the variance of the observed correlations (*r*s) in the *i*th meta-analysis, $\hat{\bar{r}}_i$ is the estimate of the mean effect size for the *i*th meta-analysis, $\hat{\bar{\bar{r}}}_i$ is the estimate of the (weighted) grand mean effect size across the *m* meta-analyses, k_i is the number of primary studies included in the *i*th meta-analysis, and the w_i is the weight applied to the *i*th meta-analysis.

The second term on the right side of equation 15.48 is the expected (weighted average) second-order sampling error variance across the *m* meta-analyses:

$$E\left(S^2_{e_{\hat{\bar{r}}_i}}\right) = \sum_1^m \left(w_i \frac{S^2_{r_i}}{k_i}\right) \Big/ \sum_1^m w_i . \tag{15.49d}$$

Equation 15.49d reduces to equation 15.49e:

$$E\left(S^2_{e_{\hat{\bar{r}}_i}}\right) = m \Big/ \sum_1^m w_i . \tag{15.49e}$$

In sum, each first-order meta-analysis will have reported a meta-analytic mean uncorrected-observed correlation, $\hat{\bar{r}}_i$. The first term on the right in equation 15.48 is the weighted variance of these meta-analytic mean correlations. This computation is shown in equations 15.49a and 15.49b. The weights (w_i) used in equations 15.49a, 15.49b, 15.49d, and 15.49e are as defined in equation 15.49c. Each weight is the inverse of the random-effect (RE) sampling error variance for the meta-analytic mean correlation in the *i*th meta-analysis. The second term on the right in equation 15.49 is the sampling error variance of these meta-analytic mean correlations. Each of the meta-analyses will have reported the variance of the observed mean correlations in that meta-analysis. Dividing each such variance by k_i (the number of studies in that meta-analysis) yields the RE sampling error variance of the meta-analytic mean estimate ($\hat{\bar{r}}_i$) in that meta-analysis (see Schmidt, Oh, and Hayes 2009). The weighted average of these values across the *m* meta-analyses estimates the RE sampling error variance of the mean *r*s as a group, as shown in equations 15.49d and 15.49e. The square root of this value divided by the square root of *m* is the standard error ($SE_{\hat{\bar{\bar{r}}}}$) and can be used to put confidence intervals around the estimate of the (weighted) grand mean ($\hat{\bar{\bar{r}}}_i$; computed in equation 15.49b). Also, using the square root of the value on the left side of equation 19–48 ($\hat{\sigma}_{\bar{\rho}_{xy}}$) one can construct a credibility interval around the grand mean correlation across the *m* meta-analyses, within which a given percentage of the first order population meta-analytic (mean) effect sizes ($\hat{\rho}_{xy}$) is expected to lie (Schmidt and Hunter 2015, chapter 5). If the value on the left side of equation 15.48 is zero, the conclusion is that the mean population correlation values are the same across the meta-analyses. In that case, all the observed variance across the meta-analyses (meta-analytic mean estimates)

is accounted for by second order sampling error, and the conclusion is that there are no moderators. If it is greater than zero, the proportion of variance between-meta-analyses due to second-order sampling error can be computed as the ratio of the second term on the right side of equation 15.48 to the first term on the right side, that is, as follows:

$$\text{ProportionVar} = E\left(S^2_{e_{\hat{\bar{r}}_i}}\right) \Big/ S^2_{\hat{\bar{r}}}, \qquad (15.49f)$$

and 1–ProportionVar denotes the proportion of the variance across first-order meta-analytic (bare bones) mean correlations that is "true" variance (that is, variance not due to second-order sampling error). As such, this number is the reliability of the meta-analytic correlations. considered as a set or vector of values, one for each first-order meta-analysis (Schmidt and Hunter 2015). This follows because reliability is the proportion of total variance that is true variance. This value can be used to produce enhanced accuracy for estimates of these mean (meta-analytic) correlations from the first-order meta-analyses-by regressing them toward the value of the grand mean correlation, that is, the mean across the first order meta-analyses (Schmidt and Oh 2013).

For purposes of detecting moderators across the *m* first-order meta-analytic mean estimates, the absolute amount of true variance across the *m* first-order meta-analytic mean estimates (or even better, its square root, the *SD*) is more important than the relative percentage of variance attributable to second-order sampling error. Meta-analysts should compute and report both estimates. This principle also applies to analyses within individual first order meta-analyses (see Schmidt and Hunter 2015, 425–26).

15.5.3 Psychometric Second-Order Meta-Analysis

Measurement error downwardly biases virtually all relationships examined in psychological and behavioral research. Therefore it is important to include corrections for these biases in psychometric (construct-level) meta-analysis. One approach in psychometric meta-analysis is to correct each correlation individually for the downward bias created by measurement error (for a different approach, based on the artifact-distribution method, see Schmidt and Oh 2013).

When the first-order meta-analyses entering the second-order meta-analysis have corrected each correlation individually for measurement error (and range restriction and dichotomization, if applicable), the fundamental equation for second-order meta-analysis is as follows:

$$\hat{\sigma}^2_{\hat{\bar{\rho}}} = S^2_{\hat{\bar{\rho}}} - E\left(S^2_{e_{\hat{\bar{\rho}}_i}}\right), \qquad (15.50)$$

where the term on the left is the estimate of the actual (not artifactual) variance across the *m* meta-analyses of the population mean disattenuated-corrected correlations ($\hat{\bar{\rho}}$); that is, the variance after variance due to second-order sampling error has been subtracted.

The first term on the right side of equation (15.50) is the variance of the meta-analytic mean individually corrected correlations across the *m* meta-analyses, computed as follows:

$$S^2_{\hat{\bar{\rho}}} = \sum_1^m w_i^* \left(\hat{\bar{\rho}}_i - \hat{\bar{\bar{\rho}}}\right)^2 \Big/ \sum_1^m w_i^*; \text{ where} \qquad (15.51a)$$

$$\hat{\bar{\bar{\rho}}} = \sum_1^m w_i^* \hat{\bar{\rho}}_i \Big/ \sum_1^m w_i^*; \text{ and} \qquad (15.51b)$$

$$w_i^* = \left(\frac{S^2_{r_{ci}}}{k_i}\right)^{-1}, \qquad (15.51c)$$

and where $S^2_{r_{ci}}$ is the weighted variance of the disattenuated (individually corrected) correlations in the *i*th meta-analysis, $\hat{\bar{\rho}}_i$ is the meta-analytic mean disattenuated correlation in that meta-analysis, $\hat{\bar{\bar{\rho}}}$ is the (weighted) grand mean effect size across the *m* meta-analyses, k_i is the number of primary studies included in the *i*th meta-analysis, and the w_i^* is the weight applied to the *i*th meta-analysis.

The second term on the right side of equation (15.50) is the weighted average second-order sampling error variance across the *m* meta-analyses:

$$E\left(S^2_{e_{\hat{\bar{\rho}}_i}}\right) = \sum_1^m w_i^* \left(\frac{S^2_{r_{ci}}}{k_i}\right) \Big/ \sum_1^m w_i^*. \qquad (15.51d)$$

Equation 15.51d reduces to equation (15.51e):

$$E\left(S^2_{e_{\hat{\bar{\rho}}_i}}\right) = m \Big/ \sum_1^m w_i^* \qquad (15.51e)$$

where the w_i^* are as defined in equation (15.51c).

In sum, each first order meta-analysis will have reported an estimate of the meta-analytic mean disattenuated correlation ($\hat{\bar{\rho}}_i$). The first term on the right side of equation (15.50) is the variance of these meta-analytic mean correlations across these first order meta-analyses (meta-analytic mean correlations). This computation is shown in equations (15.51a) and (15.51b). Equation (15.51c) shows the weights that are used in equations (15.51a) and (15.51b). The second term on the right side of equation (15.50) is the expected value of the second-order sampling error variance of these meta-analytic mean correlations. Each meta-analysis will have reported an estimate of the variance of the corrected correlations it included, preferably to four decimal places, for precision. Dividing this value by k (the number of studies in the meta-analysis), yields the RE sampling error variance of the meta-analytic correlation for that meta-analysis. As shown in equations (15.51d) and (15.51e), the weighted mean of these values across the m meta-analyses yields the second order sampling error variance needed in equation (15.50). The square root of this value divided by the square root of m is the standard error ($SE_{\bar{\bar{\rho}}}$) and can be used to put confidence intervals around the grand mean ($\hat{\bar{\bar{\rho}}}$), shown in equation (15.51b).

The term on the left side of equation (15.50) is the estimate of the actual (non-artifactual) variance across meta-analysis of the population mean disattenuated-corrected correlations (the $\hat{\bar{\rho}}_i$), that is, the variance across first-order meta-analytic estimates after removal of variance due to second-order sampling error. Using the square root of this value ($\hat{\sigma}_{\bar{\rho}}$), credibility intervals can be placed around the grand mean computed in equation (15.51b). For example, 80 percent of population mean values are expected to lie within in the 80 percent credibility interval.

If the value on the left side of equation (15.50) is zero, the indicated conclusion is that the mean population correlation values are the same across the multiple meta-analyses. All the variance is accounted for by second-order sampling error. If this value is greater than zero, one can compute the proportion of the between-meta-analyses variance that is explained by second-order sampling error. This is computed as the ratio of the second term on the right side of equation (15.50) to the first term on the right side, that is,

$$\text{ProportionVar} = E\left(S^2_{e_{\hat{\bar{\rho}}_i}}\right) \Big/ S^2_{\hat{\bar{\rho}}}, \qquad (15.51\text{f})$$

and 1–ProportionVar denotes the proportion of the variance across the first order meta-analysis mean population correlation values that is true variance (that is, variance not due to second order sampling error). As such, this number is the reliability of the estimated mean first-order population correlations (Schmidt and Hunter 2015), because reliability is defined as the proportion of total variance that is true variance. This value can be used to refine the estimates of these first-order meta-analysis mean values by regressing them toward the value of the grand mean disattenuated correlation, that is. the mean across the m meta-analyses, computed in equation (15.51b) (for details, see Schmidt and Oh 2013). In addition, when $S^2_{\hat{\bar{\rho}}}$ is zero, the ProportionVar is 100 percent and the reliability of the vector of m first-order meta-analytic mean estimates is zero. This is the same as the situation in which all examinees get the same score on a test, making the reliability of the test zero.

For detecting the presence of moderators across the m first-order meta-analytic mean estimates, the absolute amount of true variance across m first-order meta-analytic mean estimates ($\hat{\sigma}^2_{\bar{\rho}}$) (or even better, its square root, the SD) is more important than the relative percentage of variance attributable to second-order sampling error. Meta-analysts should compute and report both estimates. This principle also applies to moderator analyses conducted within an individual first-order meta-analysis (for further discussion, see Schmidt and Hunter 2015, 425–26).

15.6 SUMMARY

The methods presented in this chapter for correcting meta-analysis results for sampling error and biases in individual study results might appear complicated. More elaborated and extended descriptions of these procedures are available, however (see Hunter and Schmidt 2004; Schmidt and Hunter 2015). Further, Windows-based software is available to apply these methods in meta-analysis (Schmidt and Le 2014). Without these corrections for biases and sampling error, the results of meta-analysis do not estimate the construct-level population relationships that are the relationships of greatest scientific and theoretical interest (Rubin 1990; Hunter and Schmidt 2004; Schmidt and Hunter 2015, chapters 1 and 14). Hence these corrections are essential to developing valid cumulative knowledge about relations between the actual constructs underlying the measures used. This is especially important in light of recent developments concerning the use of meta-analytic results in the testing of causal models. Meta-analytic results are increasingly being used as input

to path analyses and other causal modeling techniques. Path analysis assumes corrections for measurement error and other artifacts. Without appropriate corrections for the artifacts that bias meta-analysis results, these causal modeling procedures will produce erroneous results (Coffman and MacCallum 2005). In addition, second-order meta-analysis can offer unique and useful information that first-order meta-analysis cannot. Second-order meta-analysis is particularly useful in meta-analytic moderator analysis—that is, synthesizing independent first-order meta-analyses on the same relationship conducted in different settings (such as countries, research settings, ethnic or racial groups, time intervals, and so on) and comparing first-order meta-analytic results of the same relationship across settings (for empirical examples, see Schmidt and Oh 2013).

15.7 NOTES

1. Table 3 in Hunter, Schmidt, and Le (2006) shows the value of operational validity for medium complexity job is 0.66, which was obtained from the true score correlation of 0.73.
2. The value here is slightly different from the 0.26 shown in Hunter, Schmidt, and Le (2006) due to rounding of values beyond the second decimal place.
3. The 2014 version of the software (V 2.0) does not include the new correction method for indirect range restriction described in the earlier section (Le et al. 2016); it will be included in the next version.

15.8 REFERENCES

Becker, Betsy J., and Christine Schram. 1994. "Examining Explanatory Models Through Research Synthesis." In *The Handbook of Research Synthesis*, edited by Harris M. Cooper and Larry Hedges. New York: Russell Sage Foundation.

Borenstein, Michael, Larry Hedges, Julian Higgins, and Hannah Rothstein. 2010. "A Basic Introduction to Fixed-Effect and Random-Effects Models for Meta-Analysis." *Research Synthesis Methods* 1(2): 97–111.

Callender, John, and Hobart Osburn. 1980. "Development and Test of A New Model for Validity Generalization." *Journal of Applied Psychology* 65(5): 543–58.

Carlson, Kevin, Steve Scullen, Frank Schmidt, Hannah Rothstein, and Frank Erwin. 1999. "Generalizable Biographical Data Validity: Is Multi-Organizational Development and Keying Necessary?" *Personnel Psychology* 52(3): 731–56.

Coffman, Donna, and Robert MacCallum. 2005. "Using Parcels to Convert Path Analysis Models into Latent Variable Models." *Multivariate Behavioral Research* 40(2): 235–59.

Colquitt, Jason, Jeffrey LePine, and Raymond Noe. 2002. "Toward an Integrative Theory of Training Motivation: A Meta-Analytic Path Analysis of 20 Years of Research." *Journal of Applied Psychology* 85(5): 678–707.

Cook, Thomas, Harris M. Cooper, David Cordray, Heidi Hartman, Larry V. Hedges, Richard Light, Thomas Louis, and Fredrick Mosteller. 1992. *Meta-Analysis for Explanation: A Casebook*. New York: Russell Sage Foundation.

Cooper, Harris M., and Alison Koenka. 2012. "The Overview of Reviews: Unique Challenges and Opportunities When Research Syntheses Are the Principal Elements of New Integrative Scholarship." *American Psychologist* 67(6): 446–62.

Field, Andy. 2005. "Is the Meta-Analysis of Correlation Coefficients Accurate When Population Correlations Vary?" *Psychological Methods* 10(4): 444–67.

Gulliksen, Harold. 1986. "The Increasing Importance of Mathematics in Psychological Research (Part 3)." *The Score* 9(1): 1–5.

Hall, Steven M., and Michael T. Brannick. 2002. "Comparison of Two Random-Effects Methods of Meta-Analysis." *Journal of Applied Psychology* 87(2): 377–89.

Hedges, Larry V. 1989. "An Unbiased Correction for Sampling Error in Validity Generalization Studies." *Journal of Applied Psychology* 74(3): 469–77.

Hedges, Larry V., and Ingram Olkin. 1985. *Statistical Methods for Meta-Analysis*. Orlando, Fl.: Academic Press.

Hunter, John. 1980. *Validity Generalization for 12,000 Jobs: An Application of Synthetic Validity and Validity Generalization to the General Aptitude Test Battery (GATB)*. Washington: U.S. Department of Labor.

Hunter, John, and Frank Schmidt. 1990. "Dichotomization of Continuous Variables: The Implications for Meta-Analysis." *Journal of Applied Psychology* 75(3): 334–49.

———. 2004. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings,* 2nd ed. Thousand Oaks, Calif.: Sage Publications.

Hunter, John, Frank Schmidt, and Huy Le. 2006. "Implications of Direct and Indirect Range Restriction for Meta-Analysis Methods and Findings." *Journal of Applied Psychology* 91(3): 594–612.

Judge, Timothy, Carl Thoresen, Joyce Bono, and Gregory Patton. 2001. "The Job Satisfaction-Job Performance Relationship: A Qualitative and Quantitative Review." *Psychological Bulletin* 127(3): 376–401.

Law, Kenneth, Frank Schmidt, and John Hunter. 1994. "Nonlinearity of Range Corrections in Meta-Analysis: A Test of an Improved Procedure." *Journal of Applied Psychology* 79(3): 425–38.

Le, Huy, In-Sue Oh, Frank Schmidt, and Colin Wooldridge. 2016. "Correction for Range Restriction in Meta-Analysis Revisited: Improvements and Implications for Organizational Research." *Personnel Psychology* 69(4): 975–1008.

Le, Huy, and Frank Schmidt. 2006. "Correcting for Indirect Range Restriction in Meta-Analysis: Testing a New Analytic Procedure." *Psychological Methods* 11(4): 416–38.

McDaniel, Michael, Frank Schmidt, and John Hunter. 1988. "A Meta-Analysis of the Validity of Methods for Rating Training and Experience in Personnel Selection." *Personnel Psychology* 41(2): 283–314.

McDaniel, Michael, Debra Whetzel, Frank Schmidt, and Steven Maurer. 1994. "The Validity of Employment Interviews: A Comprehensive Review and Meta-Analysis." *Journal of Applied Psychology* 79(4): 599–616.

Pearlman, Kenneth, Frank Schmidt, and John Hunter 1980. "Validity Generalization Results for Tests Used to Predict Job Proficiency and Training Criteria in Clerical Occupations." *Journal of Applied Psychology* 65: 373–407.

Raju, Nambury, Tobin Anselmi, Jodi Goodman, and Adrian Thomas. 1998. "The Effects of Correlated Artifacts and True Validity on the Accuracy of Parameter Estimation in Validity Generalization." *Personnel Psychology* 51(2): 453–65.

Raju, Nambury, and Michael Burke. 1983. "Two New Procedures for Studying Validity Generalization." *Journal of Applied Psychology* 68(4): 382–95.

Raju, Nambury, Michael Burke, Jacques Normand, and George Langlois. 1991. "A New Meta-Analysis Approach." *Journal of Applied Psychology* 76(3): 432–46.

Rothstein, Hannah. 1990. "Interrater Reliability of Job Performance Ratings: Growth to Asymptote with Increasing Opportunity to Observe." *Journal of Applied Psychology* 75(3): 322–27.

Rothstein, Hannah, Frank Schmidt, Frank Erwin, William Owens, and Paul Sparks. 1990. "Biographical Data in Employment Selection: Can Validities Be Made Generalizable?" *Journal of Applied Psychology* 75(2): 175–84.

Rubin, David. 1990. "A New Perspective on Meta-Analysis." In *The Future of Meta-Analysis*, edited by Kenneth Wachter and Miron Straf. New York: Russell Sage Foundation.

Schmidt, Frank. 1992. "What Do Data Really Mean? Research Findings, Meta-Analysis, and Cumulative Knowledge in Psychology." *American Psychologist* 47(10): 1173–81.

Schmidt, Frank, Ilene Gast-Rosenberg, and John Hunter. 1980. "Validity Generalization Results for Computer Programmers." *Journal of Applied Psychology* 65(6): 643–61.

Schmidt, Frank, and John Hunter. 1977. "Development of a General Solution to the Problem of Validity Generalization." *Journal of Applied Psychology* 62(5): 529–40.

———. 1981. "Employment Testing: Old Theories and New Research Findings." *American Psychologist* 36(10): 1128–37.

———. 1998. "The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings." *Psychological Bulletin* 124(2): 262–74.

———. 2003. "History, Development, Evolution, and Impact of Validity Generalization and Meta-Analysis Methods 1975–2001." In *Validity Generalization: A Critical Review*, edited by Kevin Murphy. Mahwah, N.J.: Lawrence Erlbaum.

———. 2015. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*, 3rd ed. Thousand Oaks, Calif.: Sage Publications.

Schmidt, Frank, John Hunter, and Kenneth Pearlman. 1980. "Task Difference and Validity of Aptitude Tests in Selection: A Red Herring." *Journal of Applied Psychology* 66(2): 166–85.

Schmidt, Frank, John Hunter, Kenneth Pearlman, and Hannah Hirsh. 1985. "Forty Questions About Validity Generalization and Meta-Analysis." *Personnel Psychology* 38(4): 697–798.

Schmidt, Frank, John Hunter, and Nambury Raju. 1988. "Validity Generalization and Situational Specificity: A Second Look at the 75% Rule and the Fisher *z* Transformation." *Journal of Applied Psychology* 73(4): 665–72.

Schmidt, Frank, John Hunter, and Vern Urry. 1976. "Statistical Power in Criterion-Related Validation Studies." *Journal of Applied Psychology* 61(4): 473–85.

Schmidt, Frank, and Huy Le. 2014. *Software for the Hunter-Schmidt Meta-Analysis Methods*. Department of Management and Organizations. University of Iowa, Iowa City, Iowa.

Schmidt, Frank, Huy Le, and In-Sue Oh. 2009. "Correcting for the Distorting Effects of Study Artifcats in Meta-Analysis." In *The Handbook of Research Synthesis and Meta-Analysis,* 2nd edition, edited by Harris Cooper, Larry V. Hedges, and Jeffrey C. Valentine. New York: Russell Sage Foundation.

Schmidt, Frank, and In-Sue Oh. 2013. "Methods for Second Order Meta-Analysis and Illustrative Applications." *Organizational Behavior and Human Decision Processes* 121(2): 204–18.

Schmidt, Frank, In-Sue Oh, and Theodore Hayes. 2009. "Fixed vs. Random Models in Meta-Analysis: Model Properties and Comparison of Differences in Results." *British Journal of Mathematical and Statistical Psychology* 62(Pt. 1): 97–128.

Schmidt, Frank, In-Sue Oh, and Huy Le. 2006. "Increasing the Accuracy of Corrections for Range Restriction: Implications for Selection Procedure Validities and Other Research Results." *Personnel Psychology* 59(2): 281–305.

Schmidt, Frank, Deniz Ones, and John Hunter. 1992. "Personnel Selection." *Annual Review of Psychology* 43(1): 627–70.

Schmidt, Frank, Jonathon Shaffer, and In-Sue Oh. 2008. "Increased Accuracy for Range Restriction Corrections: Implications for the Role of Personality and General Mental Ability in Job and Training Performance." *Personnel Psychology* 61(4): 827–68.

Schulze, Ralf. 2004. *Meta-Analysis: A Comparison of Approaches*. Cambridge, Mass.: Hogrefe and Huber.

Tukey, John. 1960. "A Survey of Sampling From Contaminated Distributions." In *Contributions to Probability and Statistics*, edited by Ingram Olkin. Stanford, Calif.: Stanford University Press.

Viswesvaran, Chockalingam, Deniz Ones, and Frank Schmidt. 1996. "Comparative Analysis of the Reliability of Job Performance Ratings." *Journal of Applied Psychology* 81(5): 557–74.

16

MODEL-BASED META-ANALYSIS AND RELATED APPROACHES

BETSY JANE BECKER
Florida State University

ARIEL M. ALOE
University of Iowa

CONTENTS

16.1 WHAT ARE MODEL-BASED AND PARTIAL-EFFECTS RESEARCH SYNTHESES?

In this chapter, we describe model-based meta-analysis and related approaches to meta-analysis that examine models and questions more complex than those addressed in meta-analyses of bivariate correlations. Model-based (and model-driven) meta-analysis and the related concept of linked meta-analysis are described, and illustrated with examples. We also make connections to methods for summarizing correlational indices of partial relationships and regression results.

16.1.1 Model-Based Meta-Analysis

Decades ago, Gene Glass coined the term *meta-analysis* to capture the idea of analyzing series of related experiments (1976). Many early meta-analyses concerned treatments, but it soon became clear that questions of association could also be examined by using meta-analytic methods to summarize correlational studies. Many early correlational meta-analyses looked at straightforward questions, such as whether two variables were related. Sometimes one variable was defined as an outcome and the other as a predictor, but even so most meta-analyses examined bivariate correlations (for example, Apling 1981; Kavale 1980; Viana 1982).

Model-based meta-analysis techniques address more complex interrelations at the within-study level, including the prediction of outcomes based on *sets* of precursor variables (do A, B, and C relate to D?), chains of connections among predictors and outcomes (does A lead to B *and* B lead to C?), and questions about whether certain variables are mediators of relationships. We use the term *model* to mean "a set of postulated interrelationships among constructs or variables" at the participant level, within each study (Becker and Schram 1994, 358). Model-based meta-analyses aim to examine such interrelations cohesively, by way of a unified analysis on correlational data (typically r matrices) from a collection of studies. For example, an early meta-analysis on the prediction of science achievement examined the separate relationships of science affect and science ability to achievement (Steinkamp and Maehr 1980). Their joint impacts on achievement were examined in Betsy Becker's model-based synthesis on the topic (1992a).

Syntheses of partial effect indices can accomplish some but not all of the things that model-based meta-analyses can do; most relevant here is that an analysis of partial effects would likely examine any set of relationships in a piecewise manner, as discussed later in this chapter. Last, many meta-analyses posit models for explaining variation in effect sizes such as standardized mean differences or correlations, using predictors at the study level. These between-studies models do not address participant-level relations, thus would not fall under our definition of model-based meta-analysis.

Model-based meta-analysis has been called by several different names—model-based and model-driven meta-analysis are terms Becker has used (2001, 2009; Becker and Schram 1994). Model-*driven* meta-analysis can be distinguished from model-based meta-analysis by the fact that it is guided from the start by a theoretical or conceptual model, rather than, say, being empirically derived using any constructs that appear with a certain outcome in the literature. Studies are included in a model-driven meta-analysis only if they measure variables that are part of the relevant theory or a priori conceptualization of the problem. Model-based meta-analysis is broader, in that it also includes meta-analyses of models derived simply based on collections of empirically observed relations.

A related term—*linked meta-analysis*—was coined by Mark Lipsey (1997). Lipsey envisioned connecting multiple separate but related meta-analyses for the purpose of policy development. He argued that linked meta-analyses might address developmental changes, and could involve both individual and social level inputs (potentially based on completely different sets of studies), thus they go a step beyond the ideas we present here. Linked meta-analysis has been used to look at relationships between measures taken at two or more time points in Lipsey and James Derzon's 1998 work on the prediction of serious delinquency, but the approach has not been widely applied.

The term *meta-analytic structural equation modeling* (MASEM) was used by Mike Cheung and Wai Chan for their proposed use of SEM analyses as a data-analytic

approach to conducting model-based meta-analyses (2005). Most model-based meta-analyses consider linkages among manifest variables; in the terminology of SEM, this is the structural portion of the model, or the path model. Becker's original analytic approach to model-based analysis aimed to estimate such path models (1992b). Latent components and measurement models can be incorporated in model-based meta-analyses given access to raw data, such as in an individual participant (or patient) data analysis (Cooper and Patall 2009; Stewart and Clarke 1995), or by using the MASEM approach (Cheung 2015).

In short, model-based meta-analyses are more extensive and more complicated than typical meta-analyses of single effect indices, but consequently they can yield benefits that go beyond those of a more typical synthesis. Certainly, with the added complexity come additional caveats; they are addressed in this chapter.

16.1.2 Syntheses of Partial Effect Sizes

On occasion, a researcher may be interested in only one or two specific partial relationships, not a full complicated model. Indices such as part and partial correlations describe relationships between two variables that would have occurred had a third variable (or more) been adjusted for or partialed out. Thus, they represent more complex relationships than do bivariate *r*s. Meta-analyses of various partial correlational indices have been conducted. Syntheses of semi-partial correlations (Aloe and Becker 2009) and partial correlations (Mathur et al. 2016), combinations of different kinds of correlations (Perry, Sibley and Duckitt 2013), and even combinations of variance-explained measures from full and reduced regression models (Yang, Aloe, and Feeley 2014) appear in the literature.

Meta-analyses of partial effect-size indices focus on partial relationships without the need for a model of the connections among multiple component variables. For example, Ariel Aloe and Becker (2009) summarized studies that related teacher verbal ability to measures of student achievement. Many studies reported on this relationship by way of multiple regression models that had partialed out such other features as teacher or student socioeconomic status or prior student achievement (though the latter was rarely included). Aloe and Becker used the semi-partial correlation to represent these effects.

One concern about (and limitation of) the approach of synthesizing partial effects is that studies usually vary in the variables that are controlled, so similar partial effect sizes may not be available. For example Maya Mathur and her colleagues found that some studies had controlled for age, but others for both age and sex when studying the relationship of perceived psychological stress to telomere length (2016). The addition of control variables can change the size of the resulting partial or semi-partial *r*. If the extra partialed variables are extraneous (do not affect the relationship of interest) their presence is less of a concern because they will not affect the size of the partial effect index (Thompson, Aloe, and Becker 2018).

When identical or highly similar partial effects are available, each study could contribute a single partial effect size (or small set of effects) to the review, and standard univariate (or multivariate) meta-analysis techniques would be used to summarize the data. If several partial correlations were extracted from each study, multivariate analyses that deal explicitly with the dependence of the (partial) effect sizes or other approaches based on the use of robust standard errors could be employed (on multivariate analyses, see Raudenbush, Becker, and Kalaian 1988; Riley 2009; on robust standard errors, Tipton 2015).

16.1.3 Examples of Model-Based and Partial-Effects Meta-Analyses

Suppose a meta-analyst wanted to understand the prediction of metabolic control from three psychological factors—anxiety, depression, and coping strategies—considered jointly. This question, among others, was investigated by Sharon Brown and her colleagues (Brown et al. 2015; Brown et al. 2016; Brown and Hedges 1994). Figure 16.1 shows a model for this scenario.

We denote the measures of the psychological variables as X_1 through X_3 and the outcome, three-month average blood glucose levels (glycated hemoglobin, also denoted A1c), as Y. Lower A1c values reflect good outcomes. This model might be represented in a single study via the raw regression model $\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4$, or via a standardized-regression-equation model. The predictors in this model act jointly to produce the outcome Y, so a meta-analysis that looked at each bivariate X-Y connection separately (for example, by summarizing the pairwise correlations) would not be a true model-based meta-analysis.

The meta-analyst has at least two options for gathering information on a model such as the one shown in figure 16.1. Model-based meta-analysis of correlations provides one way to summarize studies that inform us about the joint relations in this model, or about the whole model.

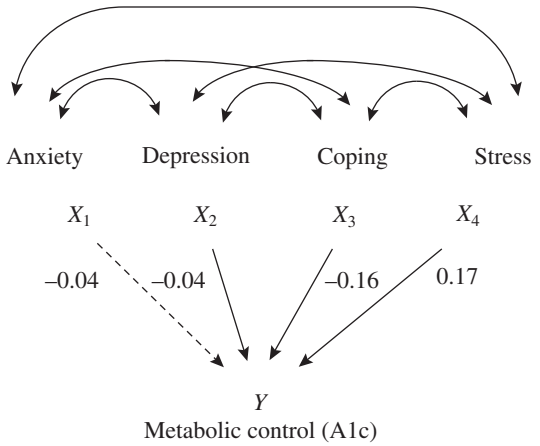

Figure 16.1 Prediction of Metabolic Control from Four Psychological Variables

SOURCE: Authors' tabulation.

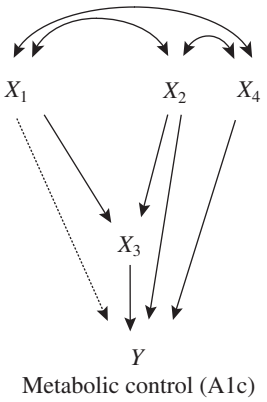

Figure 16.2 Coping Skill (X_3) Mediates the Effect of Anxiety (X_1) and Depression (X_2) on Metabolic Control

SOURCE: Authors' tabulation.

Sets of correlations (preferably complete matrices among the *X*s and *Y*) from the series of studies could be used to derive an overall mean *r* matrix, which then would be used to estimate the model above. The full process of analysis is described later in the chapter.

Another approach is to combine the partial regression slopes provided in a set of primary studies. If all studies have estimated the same "target" equation using identical measures X_1 through X_4, the raw slopes (b_is) can be combined (Becker and Wu 2007). However, Becker and Meng-Jia Wu noted that in many research domains the metrics of the measures used are quite diverse, preventing the synthesis of raw regression coefficients (2007). Studies may include the same predictor variables (constructs), but measure them using different scales (causing the raw slopes to be on different scales as well). If the primary studies report standard deviations of the outcome and the predictors, their raw regression coefficients can be standardized, and the standardized slopes can be combined.

Suppose now that the meta-analyst expects that the *X*s relate to each other as well, and wants to examine whether anxiety and depression (X_1 and X_2) affect A1c (*Y*) by way of the variable coping skills (X_3), as well as directly. Figure 16.2 shows the pathways involved. The direct paths from anxiety, depression, and stress (X_4)—all negative characteristics—to A1c suggest that persons with higher levels of these three variables would also have high levels of A1c (thus these *X*s would have positive slopes). The supportive variable of coping skill should lower levels of A1c, so coping should relate negatively to A1c. If coping skill (X_3) mediates the effects of anxiety on A1c, it would both be predicted by anxiety (with a negative slope), and relate (negatively) to A1c. The presence of large negative coefficients for these two paths would mean that people who cope well may be able to offset the effects of their anxiety, and thus would have lower levels of A1c than those who do not cope well. Complete mediation would be revealed if including both coping and anxiety in the model revealed a zero coefficient for anxiety's direct path to A1c (shown by a dashed line in figure 16.2). Models like this one were examined by Brown and her colleagues using a set of studies of diabetic outcomes (Brown et al. 2015; Brown et al. 2016).

As was true for the model shown in figure 16.1, model-based meta-analysis provides a way to combine correlations and estimate the pathways in figure 16.2. In this case, a summary of partial regression coefficients is less likely to be possible, because it is hard to find sets of studies that have all examined the two component models in figure 16.2 (that is, the models with X_3 and *Y* as outcomes). As the sets of pathways among predictors become more numerous and complex, direct combinations of slope coefficients become increasingly more difficult to achieve.

A model-driven meta-analysis of a larger model is found in Mary Whiteside and Becker's review of factors relevant to child-custody decisions, including features of

the co-parent relationship, mother-child and father-child relationships, and frequency of father-child contact (2000). They examined nine predictors of five possible outcomes for young children in divorcing families. Only three predictor-outcome relationships were significant, but many of the predictors mediated the impact of other co-parenting variables such as between-parents cooperation and hostility, and the frequency of father visitation on child outcomes.

Other applications of model-based meta-analysis include Becker's examination of science-achievement outcomes, and Lynette Craft and her colleagues' meta-analysis of the role of anxiety in predicting sport performance (Becker 1992a; Craft et al. 2003). We use data from the latter study in our examples.

Meta-analyses of partial effect sizes can be found in the literature as well (see, for example, Aloe and Becker 2009; Bowman 2010; Denson 2009). For instance, Jeffrey Valentine, David DuBois, and Harris Cooper summarized studies of the relation between self-beliefs and academic achievement, adjusting for prior achievement (2004). Nicholas Bowman used regression slopes and partial correlations to examine how college diversity experiences may affect cognitive development, controlling for other college experiences the students may have had (2010). Partial-effects syntheses are becoming more common now that methods for summarizing such effects have been developed.

16.2 WHAT CAN WE LEARN FROM MODELS?

Unlike meta-analyses that focus on single correlation coefficients, model-based meta-analyses address how *sets* of predictors relate to an outcome (or outcomes) of interest. We may examine partial relations, because we will model or statistically control for additional variables, and it is also possible to examine indirect relations and mediator effects. We provide examples of how these kinds of effects have been examined in existing model-based meta-analyses and syntheses of partial effects.

16.2.1 Partial Effects Can Be Examined

Most correlational meta-analyses take one predictor at a time, and ask, "Does the predictor X relate to the outcome Y?" However, real-world relationships are usually not so simple. Consider the issue of the effectiveness of diabetic patients at controlling their blood sugar levels. Many factors play roles in this process, including the person's diet, activity level, and attitudes. The severity of the person's disease may be important, so researchers may want to control or adjust for that. This control can be exerted through design variations (such as by selecting only patients whose disease is at a particular level of severity) or by statistical means, such as by blocking on and analyzing an indicator of disease severity or adding a variable such as the Diabetes Complications Severity Index to a model predicting the outcome of interest (Young et al. 2008).

Assuming that correlations of key predictors and outcomes with control variables are available in the primary studies, model-based meta-analyses can incorporate control variables into a more complex model than is possible in a traditional univariate meta-analysis. An example from Becker's work with Sharon Brown and colleagues on the diabetes model-based meta-analysis shows this benefit. The study involved a variety of predictors of diabetic control, measured by one of three outcomes, including hemoglobin A1c (Brown et al. 2016). The data set included correlations among anxiety (X_1), depression (X_2), coping skills (X_3), and as well as stress (X_4), and the larger data set included correlations among self-efficacy, health beliefs, and measures of adherence to diet, physical activity, and medication regimes, weight and body mass, among others.

We consider the roles of the four psychological factors as an example. Under the random-effects model, each predictor correlated significantly with A1c on average, but only coping and stress showed nontrivial correlations with A1c. Even those were small: for coping $\overline{r}_{\cdot 3Y} = -.18$ ($SE = .012$, with $k = 21$ studies); for stress $\overline{r}_{\cdot 4Y} = .17$ ($SE = .007$, $k = 66$). The correlation of anxiety with A1c was significant but negligible at $\overline{r}_{\cdot 1Y} = .02$ ($SE = .008$, $k = 35$), and depression showed a very low correlation, with $\overline{r}_{\cdot 2Y} = .07$ ($SE = .004$, $k = 116$).

When modeling all four components together as shown in figure 16.1, coping ($b = -0.16$), stress ($b = 0.17$), and depression ($b = -0.04$) remained significantly related to A1c, but anxiety ($b = -0.04$) was no longer significant. In addition, the sign of the slope for depression became negative and its standardized slope was near zero, suggesting effects of collinearity or suppression with the other predictors. As anxiety correlated on average 0.53 with anxiety and 0.45 with stress, this is a possibility. When multiple variables are controlled for, interrelations among the variables may come into play, and must be considered when estimating complex models in meta-analyses.

Partial-effects meta-analyses also enable the meta-analyst to examine partial relationships, but in a more

focused way. An example of a partial-effects meta-analysis comes from Jeffrey Valentine and his colleagues. When designing their meta-analysis,Valentine and colleagues realized the importance of controlling for previous level of academic achievement (2004). Consequently, their inclusion criteria explicitly stated that primary studies must have adjusted for pretest scores to be included in their meta-analysis. They found a small but consistent positive effect of self-beliefs on achievement once prior achievement was controlled for. Just a single partial relationship was investigated. This example also illustrates that the data required for the synthesis of partial effect sizes is typically different from that needed for model-driven meta-analysis; we elaborate on this point later on.

16.2.2 Indirect Effects Can Be Examined

A second major benefit of the use of model-based meta-analyses is the ability to examine indirect effects, where the effects of a predictor on an outcome manifest via a third intervening variable. The analysis of indirect relationships is a key aspect in primary-study analyses using structural equation models (see, for example, Kaplan 2000), as well as in meta-analytic path models and MASEMs. Complex models have been posited for a variety of outcomes in primary studies in many fields, and it makes sense that a meta-analyst might want to examine such models in a meta-analysis. Model-based meta-analyses allow us to examine indirect relationships among the predictors in the theoretical models.

A very powerful example of this benefit comes from the model-driven synthesis of predictors of child outcomes in divorcing families (Whiteside and Becker 2000). An important consideration in child-custody decisions is the extent of parental visitation for the noncustodial parent. When the mother is granted custody, decisions must be made about the extent of father visitation and contact with the child. Curiously and counterintuitively, earlier narrative syntheses showed only a weak influence for the extent of father visitation on child-adjustment outcomes such as internalizing and externalizing behaviors (see for example, Hodges 1991). Indeed, when Whiteside and Becker examined the bivariate associations of measures of father-child contact with adjustment outcomes, the quality of the father-child relationship was the only significant correlate of child outcome variables. Good father-child relationships related to higher levels of child cognitive skills (a good outcome), and lower levels of externalizing symptoms and internalizing symptoms (also good outcomes, as high internalizing and externalizing symptoms reflect problems for the child). Measures of father visitation, and pre-separation and current levels of father involvement, did not relate significantly to child outcomes.

However, Whiteside and Becker found that when father-child contact variables were examined in the context of a more realistic multivariate model for child outcomes, extent of father contact showed a significant indirect relationship, through the variable of father-child relationship quality (2000). Figure 16.3 shows that both pre-separation father involvement and extent of father visitation showed positive indirect impacts on child outcomes, suggesting that *when a child has a good relationship with their father,* more father contact has a positive impact on psychological outcomes. Children who had good relationships with their father and more involvement with their father before parental separation showed more positive outcomes. These results would have been overlooked if Whiteside and Becker had not examined the indirect relationships specified in the model-driven synthesis.

Because it has the potential to model indirect effects, model-based meta-analysis also allows for tests of mediating effects. One example of an application where mediators were critical examined the roles of risk and protective factors in substance abuse (Collins, Johnson, and Becker 2007). In a synthesis of community-based interventions for substance abuse, risk factors such as friends' use of drugs mediated the impact of preventive interventions on a variety of substance-abuse outcomes. David Collins and his colleagues used a variation of the regression approach proposed by Charles Judd and David Kenny in a meta-analytic context (Collins, Johnson, and Becker 2007; Judd and Kenny 1981). One could also test for mediation in other ways, using variations on the approaches David MacKinnon and his colleagues described (2002), or using model-comparison tests under the more traditional SEM framework, as Cheung has proposed (2015). Connections can also be made to logical analyses based on directed acyclic graphs, which allow for the exploration of causal inferences (Pearl 2009).

16.2.3 Models Can Be Compared

Model-based meta-analyses also allow us to compare models based on moderator variables. For instance, a

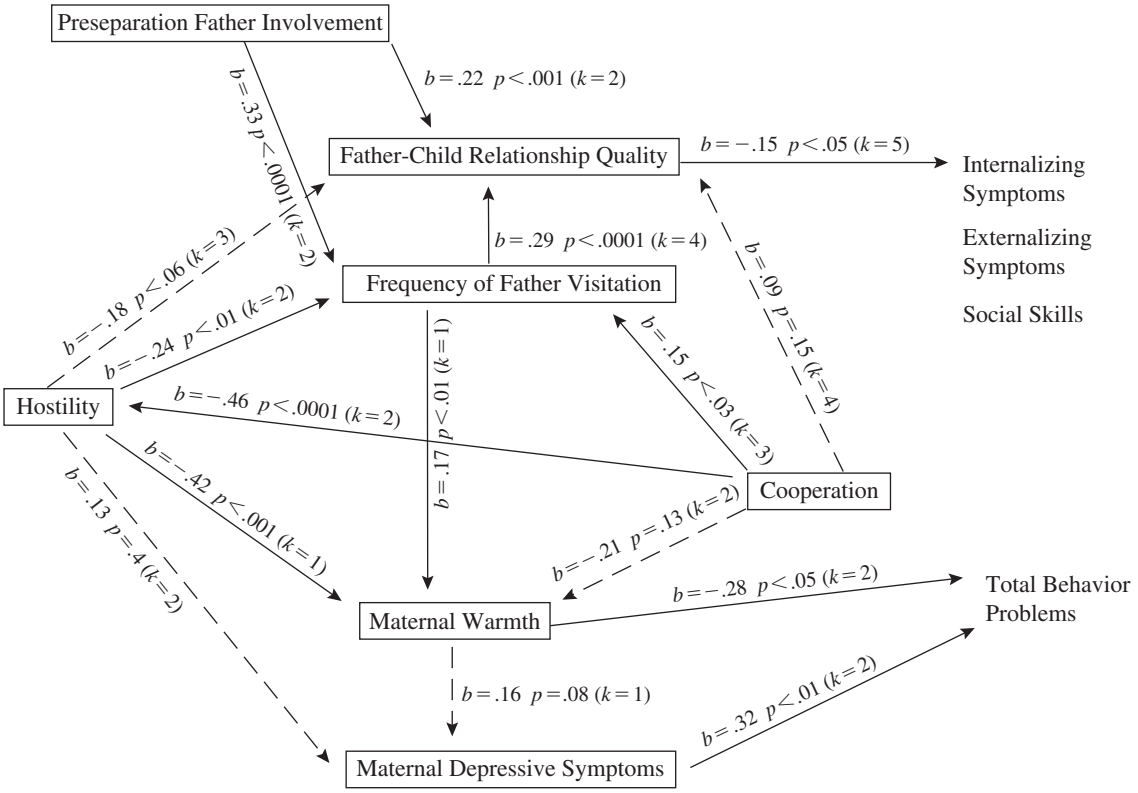

Figure 16.3 Model of Child Outcomes in Divorcing Families

SOURCE: Whiteside and Becker 2000.

meta-analyst may want to ask whether the same individual-level predictive relationships hold for two or more groups. Subgroup analyses can address such questions. For example, Whiteside and Becker found that the same predictors were important for child outcomes when comparable models were fit for young boys and young girls in divorcing families (2000). Craft and her colleagues examined a variety of aspects of the athletes and sports in their synthesis of studies of anxiety and sport performance (2003). Team sports showed significantly lower correlations than individual sports, and nearly all comparisons were significant among elite athletes, European club athletes, college athletes, and college physical education students. Type of skill (open versus closed) and time of administration of the anxiety scale (in terms of minutes prior to performance) showed significant differences as well.

Complications can arise when moderator variables are examined. If key variables are not examined or reported in some subsets of studies, it may be impossible to estimate the same models for subsets of studies as for the whole collection of studies. For categorical moderators, studies may not have analyzed their data separately by subgroup, or may have analyzed differences in levels of the outcome by the moderator but not reported separate correlations for each subgroup. Also analyses of subsamples will often be less precise because of the smaller numbers of studies (although between-studies variation may counteract this effect if the subsets' results are more homogeneous). For Whiteside and Becker, only five of their seventeen studies had reported results separately by gender, and four correlations of interest had not been studied for those samples (2000). Thus some models estimated for the full set of studies could not be examined for

the gender-specific samples, and the power was lower for the analyses that could be done.

16.2.4 We Can Learn What Has Not Been Studied

Designing a meta-analysis with a specific model in mind enables the meta-analyst to identify relationships that have not been studied at all or that have not been studied extensively. In a meta-analysis of one bivariate relationship, absence of data would cause the synthesis to be abandoned, because without correlations the given relationship cannot be studied. In the context of a model, the absence of data on a particular relationship will restrict the complexity of some analyses (for example, certain paths may not be estimable), but other analyses will still be possible.

Most critically, such missing data can reveal areas for future research. In their synthesis on child outcomes in divorcing families, Whiteside and Becker found no correlations for maternal restrictiveness with two other predictors—co-parent hostility and maternal depressive symptoms (2000). Thus two pairs of predictors (restrictiveness and hostility, and restrictiveness and maternal depression) could not be used together to predict any outcome. In their synthesis on diabetic outcomes, Brown and her colleagues were interested in the predictive power of five kinds of patient adherence—to medicine regimes, diet restrictions, physical activity plans, blood monitoring, and keeping appointments with their doctor (2016). However, appointment-keeping adherence is relatively new in the literature, and only eighteen of the thirty-five possible correlations involving appointment keeping with other variables were observed. Of these, eleven arose from one study. Appointment keeping thus could not appear in models that included the seventeen variables it had not been studied with, and no information was presented about variation in the strength of the correlations for the eleven relationships found in the single study. This variable is a candidate for further study, given that it did play a role in A1c for the few studies that had examined it.

Often it is desirable to quantify the amount of evidence available for each relationship. Counting the correlations that contribute to each path is one way to identify the amount of evidence available about each relationship. Certainly, it is not common for all paths to be represented by the same numbers of correlations. In the Brown et al. data, adherence to diet and physical activity were each studied with A1c in more than sixty studies; in contrast, each occurred with free blood glucose in fewer than ten studies (2016).

16.2.5 Extensive Models Can Be Built and Multiple Operations Examined

Model-based syntheses may seek to test models that were never examined in any one study, or that are so complex that the costs (both monetary and in terms of participant time) to collect all of the variables in a single primary study would be prohibitive.

Brown and her colleagues carefully described the selection of variables for their synthesis of diabetes outcomes (2015). Their broader theoretical diabetes-care model outlined thirteen constellations of variables, each of which contained several potential constructs. They focused first on six of these model components, targeting relationships among thirty variables, and in so doing retrieved 4,145 correlations from their set of 775 studies.

Model-based meta-analysis also allows the meta-analyst to examine whether multiple operationalizations of constructs lead to similar results (showing robustness of the model under study), or to define potentially finer-grained constructs that may show different associations with other variables. Such differences could lead to more differentiated models and better understandings of the constellation of variables under review.

16.2.6 Limitations of Model-Based Meta-Analysis

Model-based meta-analysis has some limitations. Of particular concern is the issue of missing data. When a particular relationship has not been studied, one cannot estimate any correlation between the variables of interest. This may provide future directions for research, but it also means that certain configurations of variables cannot be studied in the models in the meta-analysis. Even if every relationship has been studied at least once, most studies will probably not examine all relationships.

Some advances have been made in dealing with missing data in the context of correlation matrices (Furlow and Beretvas 2005), and the generalized least squares methods we describe later on easily handle studies in which not all correlations are reported. Wu and Becker investigated a method based on factored likelihoods and the sweep operator that handles missing data and performs well in many conditions (2013). However, for their method to work, the correlations observed in any study must be a subset of what is observed in other studies. So if r_1, r_2, and r_3 are reported in one study and r_1 and r_2 appear in another study the latter data are nested within the set of three rs, but if a third study reported r_1, r_3, and

r_4 its data would not be nested with either of the first two studies, and the method could not be applied.

A second concern when data are missing is that the results of a model-based meta-analysis may be so dispersed as to appear fractionated. Specifically, correlations relevant to one relationship might arise from a set of studies completely different from those providing results for other relationships. Thus the connection between variables A and C could be based on samples (or populations) different from those examining variables B and C, thus the overall results may not apply well to any particular population. This becomes more likely as the numbers of studies and of relationships in the review increase. Many studies in the Brown et al. diabetes dataset included only two or three of the thirty variables of interest, and across all their studies the number of correlations reported by a single study ranged from one to fifty-two (2016).

Last, as in all situations where data are not reported, the meta-analyst needs to assess the possibility that publication or reporting bias has played a role in the absence of correlations. Some primary-study authors may report only correlations of predictors with the outcome of interest, or may be restricted in what they can report by editorial practices that aim for shorter journal articles. To date nothing has been written about publication bias in the case of model-based meta-analysis, but available techniques, such as funnel plots (Egger et al. 1997), can be applied to *r*s for each element of the accumulated correlation matrix or tests of asymmetry.

16.3 HOW CAN WE CONDUCT MODEL-BASED AND PARTIAL-EFFECTS SYNTHESES?

Many of the tasks required in conducting a model-based meta-analysis are essentially the same as those required in doing a typical meta-analysis. Both Becker and Christine Schram and Brown and her colleagues provide details about these tasks (Becker and Schram 1994; Brown et al. 2015). We therefore touch on only certain key points here. We also cover points common to the synthesis of partial effect sizes.

16.3.1 Problem Formulation, Searching, and Inclusion Criteria

The process of meta-analysis always begins with problem formulation, which involves setting clear rules about the characteristics of studies that will be included. For model-based meta-analysis, the meta-analyst often begins with a model like the ones shown in figures 16.1 through 16.3—with critical components specified. For some searches particularly important relationships or population features may serve as inclusion criteria. Whiteside and Becker required that every study in their synthesis reported a correlation involving one of their child outcomes with either father-child contact or the co-parenting relationship, measured as hostility or cooperation (2000). These two constructs were chosen because they can be influenced by the courts in divorce custody decisions, which was critical to the review.

Using detailed inclusion criteria requires complex intersection searches; strategies often involve systematic pairings of relevant terms. The more stringent and specific the inclusion rules, the more limited the set of pertinent studies will be. For instance the search rule "(hostility or cooperation) and (anxiety or internalizing behaviors or depression or externalizing behaviors)" identified studies that meet one of the conditions required in the Whiteside and Becker synthesis (2000). None of Whiteside and Becker's studies reported on all correlations among the fourteen variables of interest (eleven of which are shown in figure 16.3).

Brown and her colleagues searched for studies with participants who had type 2 diabetes mellitus, that examined either hemoglobin A1c, free blood glucose, or body mass index as an outcome, and that measured at least one of the following predictors: "psychological factors (stress, depression, anxiety, coping), motivational factors (self-efficacy), or behavioral factors (adherence to diet, physical activity, medications, glucose self-monitoring, or appointment keeping)" (2016, 5). It was important to allow for this flexibility, because most studies presented correlations of just a few predictors with the outcomes. With more elaborate models, it is unlikely that all studies will report correlations among all variables, thus requiring that individual studies include all variables of interest might reduce the collection of available studies to the null set. Even though the Brown et al. analysis involved 775 studies, the maximum number of correlations observed for any relationship was 116.

As in other meta-analyses diversity in the included studies, and the number of studies deemed relevant, can be controlled by having well-planned inclusion and exclusion criteria. Populations can be carefully defined using a framework such as MUTOS, an acronym for methods, units, treatments, observations, and settings (Aloe and Becker 2008; Becker and Aloe 2008), or in the medical realm PICOS, an acronym for patients, interventions,

comparators, outcomes, and study design (Richardson et al. 1995).

16.3.2 Data Extraction

In model-based meta-analyses, we estimate an average correlation matrix and its variance-covariance matrix. The meta-analyst must therefore collect correlations, preferably correlation matrices, for as many of the variables of interest as can be found. Requiring that studies report all correlations among all variables can restrict the collection of studies greatly. The general rule is to extract from each study that meets other inclusion criteria any information that links model components together.

In partial-effects meta-analyses, a full correlation matrix is usually not needed because partial effect sizes and their respective variances can be estimated from reported regression results (Aloe and Becker 2012; Aloe and Thompson 2013; Becker and Wu 2007). However, partial effect sizes can be computed from correlation matrices as well. Several partial effect-size indices are described shortly.

One variable the meta-analyst *must* record is the sample size. When primary studies report correlation matrices computed using pairwise deletion, the *r*s for different relations in a matrix may be based on different sample sizes. In such cases, the smallest sample size reported across all relationships (all correlations) represents a conservative choice for *n*. One may also use some average value (for example, the harmonic mean *n*) but that will be less conservative than the smallest *n*.

For partial effect sizes, it is also important to code the number and kind of variables that are controlled or partialed out. Aloe and Becker recommended coding dichotomous indicators for key control variables that have been omitted from the regression model from which a semi-partial correlation is extracted (2012). This enables the meta-analyst to assess the biasing impact of not controlling those important variables. Clearly, this recommendation also holds for other partial effect sizes (such as regression slopes and partial correlations).

For both model-based and partial-effect meta-analyses, the meta-analyst may want to ask questions about moderator variables, and will thus code predictors specific to the substantive problem at hand. Participant characteristics, information about settings, and features of the measures (for example, self versus other report) would be selected in line with the key research questions.

16.3.2.1 Effect Data for Model-Based Meta-Analyses

Each of these strategies—model-based and partial meta-analyses—relies on the extraction of effect indices to represent within-study relationships of interest. Single correlations, full correlation matrices, or subsets of relevant *r*s can be extracted from individual studies for use in model-based analyses. In theory, model-based meta-analyses can be based on collections of structural equation models or factor analyses (see, for example, Cho 2015; Becker 1996) or regression models from primary studies. However, such indices are partial effect sizes. Because exact replications are rare, the effects from these multivariate within-study models may not be commensurable across studies, making it harder to use this approach to build a model-based meta-analysis.

Thus, for most model-based meta-analyses, an average bivariate correlation matrix among all relevant variables will be estimated. Ideally one would compute the average matrix from many large, representative studies that include all the variables of interest, measured with a high degree of specificity. In reality, the meta-analyst will encounter a diverse set of studies, each of which examines subsets of the variables of interest. For example, Brown and her colleagues targeted thirty variables that represented six main constructs in a model of diabetic outcomes (2015). A 30×30 matrix contains 435 unique correlations, but the most correlations any one study reported was fifty-two, and two-thirds of the studies reported no more than five *r* values.

For our examples, we consider data from the Craft et al. meta-analysis on the prediction of sport performance from three anxiety-related factors. The first step in the model-based meta-analysis was to gather estimates of the correlations in the matrix

$$\mathbf{R} = \begin{bmatrix} 1 & r_{Y1} & r_{Y2} & r_{Y3} \\ r_{1Y} & 1 & r_{12} & r_{13} \\ r_{2Y} & r_{21} & 1 & r_{23} \\ r_{3Y} & r_{31} & r_{32} & 1 \end{bmatrix},$$

where r_{ab} is the correlation between the anxiety scales X_a and X_b for a and $b = 1$ to 3, and r_{aY} is the correlation of X_a with Y (sport performance), for $a = 1$ to 3. We use the subscript i to represent the ith study, and $\mathbf{R}_i$ represents the square form of the correlation matrix from study i. The index p represents the total number of variables in the matrix (here $p = 4$). Therefore any study may have up to $p^* = p(p-1)/2$ unique correlations. For this example $p^* = 6$.

The matrix $\mathbf{R}_i$ will contain $m_i < p^*$ unique elements if one or more variables are not examined in study *i,* or if some correlations are not reported. The rows and columns of $\mathbf{R}_i$ can be arranged in any order desired, though it is convenient to place the most important or ultimate outcome in either the first or last position. In the anxiety and sport performance example, correlations involving sport outcomes appear in the first row and column of **R**. From the k $\mathbf{R}_i$ matrices the meta-analyst then estimates an average correlation matrix, and functions of that mean matrix give the path coefficients for our models.

16.3.2.2 Effect Data for Partial-Effects Meta-Analyses When regression models are reported by primary studies (in the absence of correlation matrices) the meta-analyst can usually estimate the semi-partial correlation, the partial correlation, or the slope of a standardized regression for each study, all of which are in a standard metric. These indices are most appropriate if one particular predictor is of greatest interest, because the indices do not easily allow for the creation of a full system of equations. For example, to examine the model shown in figure 16.2 using a partial-effects meta-analysis would take six partial effect sizes. Two partials would represent the model that relates X_1 and X_2 to X_3, and the other four effects would relate X_1 through X_4 to Y. However, not all studies using regression analyses will report similar models. And, despite calls to do so, many authors fail to report the descriptives needed for computing the partial indices of choice (Gozutok, Alghamdi, and Becker 2018).

Aloe and Thompson (2013) show that the partial correlation can be obtained from the regression-slope test as

$$r_{\mathrm{P}} = \frac{t_f}{\sqrt{t_f - df}},$$

where t_f is the t statistic associated with the null hypothesis that the slope for X_f is zero (H_0: $\beta_f = 0$), and df represents the degrees of freedom (the sample size minus the number of predictors in the model, minus one).

The semi-partial correlation can be obtained as

$$r_{\mathrm{SP}} = \frac{t_f \sqrt{1 - R^2}}{\sqrt{df}},$$

where R^2 is the variance explained by the model (Aloe and Becker 2012). Each of these indices has its own variance, and can be analyzed using standard meta-analytic techniques.

Another index, the standardized slope b^*, is often reported in regression studies. It can also be computed from the raw slope (b) as $b^* = b\,(S_{\mathrm{X}}/S_{\mathrm{Y}})$, as long as standard deviations of the outcome (S_{Y}) and the predictor of interest (S_{X}) are available. We advise against treating standardized slopes as comparable to correlations, and suggest summarizing them separately using their own inverse variances as weights instead (Aloe 2014, 2015; Kim 2011). For details on the synthesis of these three indices, including their asymptotic variances, see work by Aloe and Thompson 2013).

16.3.3 Data Management

Because one step in a model-based meta-analysis is to compute a correlation matrix, it is best to lay out the desired contents of that matrix early in the data-collection process so that extracted correlations can be recorded systematically. Correlations may be stored in the form of a typical data record—that is, in a line or row of a data matrix—or in the form of a square correlation matrix. For models with few variables it may be easiest to record the data on a form that looks like the "real" matrix, then enter the data from that form into a file. Database interfaces that look like the desired matrix can be created. Certain computational approaches to correlation-matrix data require that values be read in a square form, but values can be easily translated from row to square form using most computer packages.

Data-collection forms are often created so that the first correlations to be entered are those involving the primary outcome. These are the most frequently reported correlations and can easily be entered first, to be followed by missing-data codes if correlations among predictors are not reported. When a correlation does not appear, the data form may be left blank, but it may be preferable to note the absence of values with a code such as NR (not reported), NA (not available, preferred by the R language), or some numerical code. If primary-study authors have omitted values that were not significant, one could enter NS so that omitted nonsignificant results can be counted, giving a crude index of the possibility of publication bias. Often, though, it is not possible to determine why a result was not reported.

For an extensive model-based meta-analysis with many variables, forms that resemble the full correlation

matrix can be quite unwieldy. Using them would likely contribute to errors of transcription. Brown's team opted to enter the correlation for each relationship into a spreadsheet, along with its associated sample size and a dummy indicator of whether it was a Pearson correlation or was computed from other information such as a contingency table (Brown et al. 2015, 2016).

The meta-analyst may also record values of other variables of interest as moderators, or simply as descriptors of the primary studies. There are no particular requirements for those additional variables, and each relationship in a matrix or each variable may have its own moderator variables (for example, type of outcome would be relevant to *X-Y* correlations, but not to interpredictor correlations).

Thus for each study the data will include, at a minimum, a study identification code, the sample size, and up to p^* (here, six) correlations. Table 16.1 shows the data for ten studies drawn from the Craft et al. meta-analysis (2003). The record for a study shows a value of 9.00 when a correlation is missing. The fifth line of table 16.1 shows the data from a study that is missing all interpredictor correlations. The table also includes a variable representing whether the sport examined was a team (T) or individual (I) sport.

One last set of information is needed for the generalized least squares approach to synthesizing correlation matrices—a set of indicator variables (typically stored as a matrix) that identify which correlations are reported in each study. The methods described in this chapter represent the study data as vectors of correlations, and the correlations reported in each study are coded using matrices of dummy variables. If data were stored in the form of one record per correlation, then a set of p^* dummy variables would indicate which relationships appeared in each study. We illustrate this process shortly.

Consider the studies with IDs 3 and 6 in table 16.1. Study 3 estimates all six correlations among a performance outcome and the subscales of the Competitive State Anxiety Index (CSAI) (Martens, Vealey, and Burton 1990). The results of study 3 appear as a 6×1 vector r_3 together with the 6×6 indicator (identity) matrix $\mathbf{X}_3$, associated with the model $\mathbf{r}_3 = \mathbf{X}_3 \boldsymbol{\rho} + \mathbf{e}_3$. (The

Table 16.1 Relationships Between CSAI Subscales and Sport Performance

			Variable Names and Corresponding Correlations					
			Cognitive/ Performance	Somatic/ Performance	Self-Confidence/ Performance	Cognitive/ Somatic	Cognitive/ Self-Confidence	Somatic/ Self-Confidence
ID	n_i	Type of Sport	C1 r_{i1Y}	C2 r_{i2Y}	C3 r_{i3Y}	C4 r_{i12}	C5 r_{i13}	C6 r_{i23}
1	142	I	–.55	–.48	.66	.47	–.38	–.46
3	37	I	.53	–.12	.03	.52	–.48	–.40
6	16	T	.44	.46	9.00	.67	9.00	9.00
10	14	I	–.39	–.17	.19	.21	–.54	–.43
17	45	I	.10	.31	–.17	9.00	9.00	9.00
22	100	I	.23	.08	.51	.45	–.29	–.44
26	51	T	–.52	–.43	.16	.57	–.18	–.26
28	128	T	.14	.02	.13	.56	–.53	–.27
36	70	T	–.01	–.16	.42	.62	–.46	–.54
38	30	I	–.27	–.13	.15	.63	–.68	–.71

SOURCE: Authors' tabulation based on Craft et al. 2003.

matrix $\mathbf{X}$ does not contain the raw data on $X_1 - X_3$.) Specifically,

$$\mathbf{r}_3 = \begin{bmatrix} r_{31} \\ r_{32} \\ r_{33} \\ r_{34} \\ r_{35} \\ r_{36} \end{bmatrix} = \begin{bmatrix} .53 \\ -.12 \\ .03 \\ .52 \\ -.48 \\ -.40 \end{bmatrix}, \mathbf{X}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \text{and } \boldsymbol{\rho} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \rho_4 \\ \rho_5 \\ \rho_6 \end{bmatrix},$$

where $\mathbf{e}_3$ represents the deviation of the observed values in $\mathbf{r}_3$ from the unknown population values in $\boldsymbol{\rho}$. Here $\mathbf{X}_3$ is an identity matrix because study 3 has all six correlations and the columns of $\mathbf{X}$ represent each of the correlations being studied.

In contrast, study 6 did not measure self-confidence, and reports only three correlations: for the first, second, and fourth relationships in the correlation matrix. Thus the data from study 6 are represented by

$$\mathbf{r}_6 = \begin{bmatrix} r_{61} \\ r_{62} \\ r_{64} \end{bmatrix} = \begin{bmatrix} .44 \\ .46 \\ .67 \end{bmatrix} \quad \text{and} \quad \mathbf{X}_6 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

The indicator matrix still contains six columns to allow for the six correlations we plan to summarize across all studies. However, three of the columns in $\mathbf{X}$ for study 6 contain only zeroes. The elements of the $\mathbf{X}$ matrices can be entered as data or can be created from the correlation data and the missing-data codes using software such as R or SAS.

16.3.4 Analysis and Interpretation

A main goal of all approaches to model-based meta-analysis is to discover patterns in the targeted relationships and to find potential explanations for important variability in the correlations of interest across studies. Another objective is to explicitly identify missing (unstudied) or less-well-studied relationships; these may suggest fruitful areas for future research.

The estimation and data-analysis methods used to accomplish model-based meta-analysis have been referred to as estimating synthetic linear models (Becker 1992b, 1995), meta-analytic structural equation modeling or MASEM (Cheung 2014; Cheung and Chan 2005), MA-SEM (Furlow and Beretvas 2005), and two-stage structural equation modeling (TSSEM) (Cheung and Chan 2005). Interest in these strategies is still strong years after the methods were first proposed; a 2016 special issue of *Research Synthesis Methods* included five articles and two commentaries on various aspects of the topic. Before describing the steps in the process of data analysis, we present some additional notation.

16.3.4.1 Distribution of the Correlation Vector r

Regardless of approach, the meta-analyst begins a model-based analysis by estimating an average correlation matrix across studies. Most estimators are based on asymptotic assumptions, and most require the variances for the correlations to be summarized. Multivariate approaches also need their covariances.

16.3.4.1.1 Fixed Effects. We first define notation for the correlation vector for the matrix $\mathbf{R}_i$ by listing the unique elements of $\mathbf{R}_i$ in a vector, here, $\mathbf{r}_i = (r_{iY1}, r_{iY2}, r_{iY3}, r_{i12}, r_{i13}, r_{i23})'$. We refer to the entries (r_{iY1}, r_{iY2}, r_{iY3}, r_{i12}, r_{i13}, r_{i23}) as r_{i1} through r_{ip^*}, or more generally as $r_{i\alpha}$, for $\alpha = 1$ to p^*, so that the elements of $\mathbf{r}_i$ are indexed by just two subscripts. In our example $\mathbf{r}_i = (r_{i1}, r_{i2}, \ldots, r_{i6})'$.

Ingram Olkin and Minoru Siotani show that, in large samples of size n_i, the correlation vector $\mathbf{r}_i$ is approximately normally distributed with mean vector $\boldsymbol{\rho}_i$ and variance-covariance matrix $\boldsymbol{\Sigma}_i$, where the elements of $\boldsymbol{\Sigma}_i$ are defined by $\sigma_{i\alpha\gamma}$ (1976). Formulas for the variances and covariances are also provided (238). To distinguish among the p variables involved in the p^* correlations in each study, we use full subscripting because the covariance of r_{ist} with r_{iuv} involves all of the cross-correlations (r_{ist}, r_{isu}, r_{isv}, r_{itu}, r_{itv}, and r_{iuv}). Specifically the large-sample variance of $r_{i\alpha} = r_{ist}$ is

$$\sigma_{i\alpha\alpha} = \mathrm{Var}(r_{i\alpha}) = \mathrm{Var}(r_{ist}) = \left(1 - \rho_{ist}^2\right)^2 / n_i. \quad (16.1)$$

The covariance $\sigma_{i\alpha\gamma}$ can be expressed by noting that if the correlation between the sth and tth variables in study i is $r_{i\alpha} = r_{ist}$, and $r_{i\gamma} = r_{iuv}$, with ρ_{ist} and ρ_{iuv} the corresponding population values, then

$$\begin{aligned} \sigma_{i\alpha\gamma} &= \mathrm{Cov}\left(r_{i\alpha}, r_{i\gamma}\right) = \mathrm{Cov}\left(r_{ist}, r_{iuv}\right) \\ &= \left[0.5\,\rho_{ist}\rho_{iuv}\left(\rho_{isu}^2 + \rho_{isv}^2 + \rho_{itu}^2 + \rho_{itv}^2\right) + \rho_{isu}\rho_{itv} + \rho_{isv}\rho_{itu}\right. \\ &\quad \left. - \left(\rho_{ist}\rho_{isu}\rho_{isv} + \rho_{its}\rho_{itu}\rho_{itv} + \rho_{ius}\rho_{iut}\rho_{iuv} + \rho_{ivs}\rho_{ivt}\rho_{ivu}\right)\right] / n_i \end{aligned} \quad (16.2)$$

for *s, t, u*, and $v = 1$ to p. Sometimes the denominators in (16.1) and (16.2) are shown as $(n_i - 1)$ rather than n_i.

Although these variances and covariances are treated as known, in practice they must be computed. Often $\sigma_{i\alpha\alpha}$ and $\sigma_{i\alpha\gamma}$ are estimated by substituting the corresponding sample estimates for the parameters in (16.1) and (16.2). However, substituting individual study correlations into these formulas is often not the best approach (Becker and Fahrbach 1994; Furlow and Beretvas 2005; Hafdahl 2007, 2008). In short, we recommend computing these variances and covariances by substituting values of the mean correlations across studies for the population values. We denote the estimates of (16.1) and (16.2) as $s_{i\alpha\alpha}$ and $s_{i\alpha\gamma}$ respectively. The estimated variance-covariance matrix for study i with elements estimated using (16.1) and (16.2) is denoted $\mathbf{S}_i$.

Another issue is whether to analyze raw correlations or transform them using Sir Ronald Fisher's z transformation, which is $z = 0.5 \log [(1 + r)/(1 - r)]$, and has the attractive feature of being a variance-stabilizing transformation (1921). The variance of $z_{i\alpha}$ is usually taken to be $\psi_{i\alpha\alpha} = n_i^{-1}$ or $(n_i - 3)^{-1}$, neither of which involve the parameter $\rho_{i\alpha}$. The covariance between two z values, $\psi_{i\alpha\gamma} = \text{Cov}(z_{i\alpha}, z_{i\gamma}) = \sigma_{i\alpha\gamma}/[(1 - \rho_{i\alpha}^2)(1 - \rho_{i\gamma}^2)]$, is more complicated than the covariance in 16.2. Various work supports the use of the Fisher transformation both in terms of estimating the mean correlation matrix, and improving the behavior of associated test statistics (Becker and Fahrbach 1994; Furlow and Beretvas 2005; Hafdahl 2007). However, results in the z metric, particularly between-studies variances, are more challenging to interpret, and require that point estimates be transformed back to the r metric. This is problematic when working with functions of the mean matrix and its variance, such as the path coefficients estimated for models. The problem occurs because there is no easy way to transform the between-studies variance of Fisher's z back to the r scale. Thus we present analyses based on the use of raw correlations.

The formulas for variances and covariances in (16.1) and (16.2) apply under the fixed-effects model, that is, when one can assume that all k correlation matrices arise from a single population. If that is unreasonable either on theoretical grounds or because tests of homogeneity are significant, then one should augment the within-study uncertainty quantified in $\mathbf{S}_i$ by adding between-studies variation. A number of approaches exist for estimating between-studies variation in the univariate context (see, for example, chapter 12; Sidik and Jonkman 2005). Estimators for the multivariate context have also been proposed (see Becker and Schram 1994; Jackson, White, and Thompson 2010; van Houwelingen, Arends, and Stijnen 2003). Ian White provides the Stata command *mvmeta* for multivariate meta-analysis (2009); *mvmeta* implements the maximum likelihood estimator of Hans van Houwelingen and his colleagues (2003). In the R environment, several other approaches are available in Wolfgang Viechtbauer's *metafor* package (2010).

Our own assessment based on experience with a variety of data sets is that though it is relatively simple to estimate between-studies variances, estimation of between-studies *covariance* components can be problematic (for example, producing between-studies correlations beyond ±1), particularly with small numbers of studies (see also Riley et al. 2007). If all studies have complete data, a multivariate hierarchical modeling approach (Kalaian and Raudenbush 1996) may solve these problems. However, complete data are rare when summarizing correlation matrices. Thus it is sometimes advisable to add only the variance components when adopting a random-effects model, to constrain covariances to more reasonable in-range values if the covariance terms do not appear reasonable, or to use a more sophisticated variance estimator rather than a simpler one.

Example. We calculate the covariance matrix using data from study 3 in the anxiety and sport example. This study reported all 6 correlations, thus has a 6×6 covariance matrix. The covariance of $\mathbf{r}_3$, computed using $n = 37$ and the sample-size weighted mean correlations across all studies, specifically $\bar{\mathbf{r}}_n = (-.074\ -.127\ .323\ .523\ -.416\ -.414)'$, is

$$\text{Cov}(\mathbf{r}_3) = \begin{bmatrix} .027 & .014 & -.010 & -.002 & .007 & .003 \\ .014 & .026 & -.009 & -.001 & .003 & .007 \\ -.010 & -.009 & .022 & .000 & -.000 & -.001 \\ -.002 & -.001 & .000 & .014 & -.005 & -.005 \\ .007 & .003 & -.000 & -.005 & .018 & .008 \\ .003 & .007 & -.001 & -.005 & .008 & .019 \end{bmatrix}.$$

For studies that report fewer than p^* correlations the $\text{Cov}(\mathbf{r}_i)$ matrix will be reduced in size accordingly. In our data set studies 6 and 17, with only three correlations each, have 3×3 covariance matrices.

Once correlations or other partial effect sizes have been extracted from multiple studies, the meta-analyst can estimate both the average strength of the relationship

between variables and the degree to which the correlations vary across studies (Becker 1992b). Estimating such variation is consistent with adopting the random-effects model (DerSimonian and Laird 1986; Hedges 1983; Hedges and Vevea 1998). In the random-effects case, estimates of variation across studies are incorporated into the uncertainty of the mean correlations. Adopting random-effects also allows estimates of the average correlations to be generalized to a broader range of situations, at the cost of wider confidence intervals for the parameters. Tests of the appropriateness of the fixed-effects model for the entire correlation matrix can be computed if desired (see Becker 1992b; Cheung and Chan 2005).

16.3.4.2 Estimating the Mean Correlation Matrix Under Fixed Effects Data from series of correlation matrices are inherently multivariate. Generalized least squares (GLS) methods for multivariate meta-analysis data were proposed by Stephen Raudenbush, Becker, and Hripsime Kalaian (1988; see also Berkey, Anderson, and Hoaglin 1996; Jackson, White, and Thompson 2010; Riley 2009). The GLS approach has also been applied to correlation matrices (Becker 1992b, 1995; Becker and Schram 1994). Becker and Schram also presented likelihood-based methods. Bayesian methods can be used to estimate the mean matrix as well (Cheung and Chan 2005; Prevost et al. 2007).

Under fixed effects, the GLS estimator of the mean is

$$\hat{\boldsymbol{\rho}} = \left(\mathbf{X}'\mathbf{S}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{S}^{-1}\mathbf{r}, \tag{16.3}$$

where $\mathbf{X}$ is a stack of k $p^* \times p^*$ indicator matrices, $\mathbf{r}$ is the vector of correlations, and $\mathbf{S}$ is a blockwise diagonal matrix with the matrices $\mathbf{S}_i$ on its diagonal. The variance of $\hat{\boldsymbol{\rho}}$ is

$$\text{Var}\left(\hat{\boldsymbol{\rho}}\right) = \left(\mathbf{X}'\mathbf{S}^{-1}\mathbf{X}\right)^{-1}. \tag{16.4}$$

These GLS estimators have been extensively studied, along with other estimation methods. Restricted maximum likelihood (REML) methods, which for some estimators provide estimates identical to GLS values, are widely used and can be obtained via the R package *metafor*, using the *rma.mv* function (Viechtbauer 2010).

Example. Analyses for our examples are performed using independent R code (our package called *metaRmat* is available from the authors) as well as the *metafor* package. For the ten studies in our example, the mean correlation matrix obtained from GLS analysis of the correlations under fixed effects (in square format) is

$$\hat{\boldsymbol{\rho}} = \begin{bmatrix} 1 & -.074 & -.127 & .316 \\ -.074 & 1 & .523 & -.415 \\ -.127 & .523 & 1 & -.405 \\ .316 & -.415 & -.405 & 1 \end{bmatrix}.$$

Identical results are obtained using maximum likelihood estimation. The first row and column of the matrix contain the correlations of the CSAI subscales with sport behavior. The last entry in row one, which represents the average correlation for self-confidence with sport behavior, is the largest *X-Y* correlation. However these values should not be interpreted because the fixed-effects model is not appropriate for these data. Nonetheless for later comparison we present the covariance matrix of the means under fixed effects:

$$\text{Cov}(\hat{\boldsymbol{\rho}}) = \begin{bmatrix} 0.00156 & 0.00080 & -0.00057 & -0.00012 & 0.00040 & 0.00020 \\ 0.00080 & 0.00153 & -0.00055 & -0.00005 & 0.00018 & 0.00039 \\ -.0.00057 & -0.00055 & 0.00129 & 0.00002 & -0.00000 & -0.00007 \\ -.0.00012 & -0.00005 & 0.00002 & 0.00090 & -0.00032 & -0.00032 \\ 0.00040 & 0.00018 & -0.00000 & -0.00032 & 0.00118 & 0.00053 \\ 0.00020 & 0.00039 & -0.00007 & -0.00032 & 0.00053 & 0.00119 \end{bmatrix}. \tag{16.5}$$

16.3.4.3 Test of Homogeneity, with H_{01}: $\rho_1 = \ldots = \rho_k$ In a univariate meta-analysis, it can be useful to test whether all studies appear to be drawn from a single population (Hedges 1982). One may choose any model without testing, however, on the basis of theoretical reasons. The meta-analyst can test whether all k correlation matrices appear to be equal or homogeneous, or can do p^* tests of whether *each* of the relationships in the matrix arises from one population. A test of the hypothesis of homogeneity of the k population correlation matrices is a test of H_{01}: $\rho_1 = \ldots = \rho_k$, specifically

$$Q_E = \mathbf{r}'\left[\mathbf{S}^{-1} - \mathbf{S}^{-1}\mathbf{X}'\left(\mathbf{X}'\mathbf{S}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{S}^{-1}\right]\mathbf{r}. \tag{16.6}$$

When H_{01} is true, Q_E has approximately a chi-square distribution with $(k-1)p^*$ degrees of freedom if all studies report all correlation values. (See however results that suggest such chi-square tests are misused, as well as overly likely to reject H_0 for small samples; Hoaglin 2016.) If some correlations are missing, the degrees of freedom equal the total number of observed correlations

across studies minus p^*. Here the *df* are 9(6) – 6 = 48 because of the three correlations missing from each of two studies. This test is also provided by *metafor*.

Cheung and Chan investigated the chi-square test of homogeneity based on raw correlations (2005). They find, as Becker and Fahrbach did (1994), that the test rejected the null hypothesis at rates above the nominal rate, making it conservative. Cheung and Chan also studied a Bonferroni adjusted test—the BA1 (Bonferroni at least 1) rule—based on the individual tests of homogeneity for the p^* unique relationships in the correlation matrix (Cheung 2000). The BA1 test rejects H_{01}: $\rho_i = \rho$ if at least one of the p^* individual homogeneity tests is significant at level α/p^*. For a 4 by 4 matrix of means with $p^* = 6$ correlations, the adjusted significance level for the $\alpha = 0.05$ test would be 0.05/6 = 0.0083.

Example. The ten studies in our example include fifty-four correlations ($k \times p^* = 10 \times 6 = 60$ were possible, but six are missing) and the value of $Q_E = 198.66$ is significant with $df = 48$ and $p < .001$, suggesting that the matrices do not arise from a single population. According to the BA1 rule, at least one of the individual homogeneity tests for the six relationships in the matrix must be significant at the .0083 level. All of the homogeneity tests for the correlations of the three CSAI subscales with sport performance are highly significant, with *p*-values well below 0.0001 (and below the BA1 cutoff); the three sets of correlations among the subscales are each homogeneous. (The six Q values are 80.2 [df = 9], 45.1 [9], 52.0 [8], 7.5 [8], 13.0 [7], and 11.6 [7].) Thus we also reject H_0 according to the BA1 test. We explore the between-studies variances for the correlations of the three CSAI subscales with sport performance in the next section.

16.3.4.4 Estimating Between-Studies Variation If a random-effects model seems sensible, either on principle or because the hypothesis of homogeneity has been rejected, the meta-analyst must estimate the between-studies variance τ^2_α (with $\alpha = 1$ to p^*) for each set of correlations. When analyses are conducted on raw correlations, variances and possibly covariances are estimated in the *r* metric and the estimate $\hat{\mathbf{T}}$ is added to each $\mathbf{S}_i$ matrix to get $\hat{\boldsymbol{\Sigma}}_i^{RE} = \mathbf{S}_i + \hat{\mathbf{T}}$. The blockwise diagonal matrix $\hat{\boldsymbol{\Sigma}}^{RE}$ would then be used in place of $\mathbf{S}$ in equations (16.3) and (16.4) to obtain random-effects estimates, given in (16.7) and (16.8). If covariance components of the $\hat{\mathbf{T}}$ matrix are problematic then the diagonal matrix $\hat{\mathbf{T}}^D = \text{diag}(\hat{\tau}_{11}, \hat{\tau}_{22}, \ldots, \hat{\tau}_{p^* p^*})$ can be used in its place. The meta-analyst would add $\hat{\mathbf{T}}^D$ rather than $\hat{\mathbf{T}}$ to each of the within-study covariance matrices $\mathbf{S}_i$.

Example. An important issue when reporting results from random-effects models is the approach implemented to estimate between-studies variation. We selected the REML estimation option for our R code, given past experiences with the method-of-moments estimator, as well as the research on variance estimation. The values on and above the diagonal in $\hat{\mathbf{T}}$ are the variances and covariances, and correlations obtained by calculating $\hat{\tau}_{\alpha\beta} \Big/ \sqrt{\hat{\tau}_{\alpha\alpha}\hat{\tau}_{\beta\beta}}$ are below the diagonal. Based on the homogeneity tests for these data, we expect at least one of the $\tau_{\alpha\alpha}$ estimates to be nonzero for the correlations of the three anxiety scales with sport performance (the first three diagonal entries). This pattern is seen in the REML estimate

$$\hat{\mathbf{T}} = \begin{bmatrix} .126 & .077 & -.040 & -.010 & .013 & .016 \\ .88 & .060 & -.061 & -.006 & -.019 & .013 \\ -.45 & -.99 & .062 & -.008 & .005 & .001 \\ -.64 & -.50 & -.71 & .002 & -.003 & .002 \\ .34 & -.72 & .20 & -.64 & .011 & -.001 \\ .58 & .68 & .06 & .59 & -.11 & .006 \end{bmatrix}.$$

Inspection of the correlations associated with $\hat{\mathbf{T}}$ shows one value very near to –1; it is, for $\hat{\tau}_{23}$, associated with the correlations of performance with cognitive anxiety and performance with somatic anxiety.

The square root of each diagonal element of $\hat{\mathbf{T}}$ is the estimated standard deviation of the population of correlations for the relationship. As expected, the first three diagonal elements are larger than the later three. For instance, the ρ_{1i} values for correlations of cognitive anxiety with sport behavior have a standard deviation of $\hat{\tau}_1 = \sqrt{0.126} = 0.35$. If the true average correlation were zero, then roughly 95 percent of the population correlations would fall between –0.69 and 0.69. This is quite a wide spread in population correlation values. The range obtained for the correlations of somatic anxiety with performance (the ρ_{2i}) is a little narrower: +0.48. These ranges are analogous to the plausible values range computed for hierarchical linear model parameters at level 2 (see, for example, Raudenbush and Bryk 2002). In practice, we would center this interval on an appropriate mean, thus we next introduce the mean under the random-effects model.

16.3.4.5 Random-Effects Mean Correlation If the meta-analyst believes that a random-effects model is

more appropriate for the data, the value of $\hat{\boldsymbol{\Sigma}}^{RE}$—the variance including between-studies differences—would be used in place of **S** in (16.3). The random-effects mean $\hat{\boldsymbol{\rho}}^{RE}$ is thus

$$\hat{\boldsymbol{\rho}}^{RE} = \left(\mathbf{X}'\left[\hat{\boldsymbol{\Sigma}}^{RE}\right]^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\left[\hat{\boldsymbol{\Sigma}}^{RE}\right]^{-1}\mathbf{r}, \qquad (16.7)$$

where **X** is still the stack of k $p^* \times p^*$ indicator matrices, **r** is the vector of correlations, and $\hat{\boldsymbol{\Sigma}}^{RE}$ is the blockwise diagonal matrix with the matrices $\hat{\boldsymbol{\Sigma}}_i^{RE}$ on its diagonal. The variance of $\hat{\boldsymbol{\rho}}^{RE}$ is

$$\mathrm{Var}\left(\hat{\boldsymbol{\rho}}^{RE}\right) = \left(\mathbf{X}'\left[\hat{\boldsymbol{\Sigma}}^{RE}\right]^{-1}\mathbf{X}\right)^{-1}. \qquad (16.8)$$

Example. In our data the first three correlations showed significant variation, and now under the random-effects model their means are slightly smaller than under the fixed-effects assumptions (the values were –.07, –.13, and .32). The random-effects mean obtained with the REML **T** estimate (shown in square form) is

$$\hat{\boldsymbol{\rho}}^{RE} = \begin{bmatrix} 1 & -.034 & -.071 & .233 \\ -.034 & 1 & .544 & -.453 \\ -.071 & .544 & 1 & -.397 \\ .233 & -.453 & -.397 & 1 \end{bmatrix},$$

with covariance

$$\mathrm{Cov}\left(\hat{\boldsymbol{\rho}}^{RE}\right) = \begin{bmatrix} 0.0150 & 0.0089 & -0.0048 & 0.0005 & -0.0016 & 0.0008 \\ 0.0089 & 0.0081 & -0.0049 & 0.0007 & -0.0019 & 0.0004 \\ -0.0048 & -0.0049 & 0.0085 & -0.0013 & 0.0020 & -0.0016 \\ 0.0005 & 0.0007 & -0.0013 & 0.0011 & -0.0007 & -0.0001 \\ -0.0016 & -0.0019 & 0.0020 & -0.0007 & 0.0027 & 0.0005 \\ 0.0008 & 0.0004 & -0.0016 & -0.0001 & 0.0005 & 0.0020 \end{bmatrix}.$$

Many of the values in the first three rows and columns of $\mathrm{Cov}(\hat{\boldsymbol{\rho}}^{RE})$ are considerably larger than the fixed-effects variances and covariances in equation (16.5). Values for the latter three columns and rows remain small because these relationships had very small $\hat{\tau}_{\alpha\beta}$ values.

16.3.4.6 Test of No Association, with H_{02}: $\rho = 0$ Meta-analysts nearly always want to know whether any correlations in the matrix ρ are nonzero, that is, to test H_{02}: $\rho = \mathbf{0}$. Rejecting H_{02} implies that at least one element of the matrix ρ is nonzero, but it does not mean that all elements are nonzero. GLS theory provides such a test for the correlation vector, and maximum-likelihood-based tests are also available and are given in *metafor*'s *rma.mv* routine. The GLS based test under random effects uses the statistic

$$Q_B = \hat{\boldsymbol{\rho}}^{RE\prime}\left(\mathbf{X}'\left[\hat{\boldsymbol{\Sigma}}^{RE}\right]^{-1}\mathbf{X}\right)^{-1}\hat{\boldsymbol{\rho}}^{RE}. \qquad (16.9)$$

When the null hypothesis is true, Q_B has approximately a chi-square distribution with p^* degrees of freedom. If the random-effects model is not called for, $\hat{\boldsymbol{\Sigma}}^{RE}$ can be replaced with **S**, or used as is, because it will be close to **S** under fixed effects (often within rounding error).

Becker and Fahrbach found that under fixed effects the rejection rate of Q_B for $\alpha = .05$ was a bit high, but never more than .07, and was usually within .01 of $\alpha = .05$ (1994). When the within-study sample size was 250, the rate even dropped below .05 in some cases. Usually the test Q_B computed under random-effects conditions will be smaller than the analogous value using a fixed-effects mean and variance, thus effects are less likely to be judged significant in the context of the random-effects model. A similar test for a subset of correlations can be done by selecting a submatrix of m values from $\hat{\boldsymbol{\rho}}^{RE}$ and using the corresponding submatrices of **X** and $\hat{\boldsymbol{\Sigma}}^{RE}$ to compute Q_B for the subset of correlations.

Example. The value of Q_B based on the REML estimates is quite large under both the fixed and random-effects models. Under the more appropriate random-effects model, $Q_B = 403.9$ with $df = 6$, suggesting that at least one population correlation differs from zero. Individual random-effects tests of the significance of the first two correlations do not reject the null ($z = –0.28$ for the cognitive anxiety-performance correlation and $z = –0.79$ for somatic anxiety with performance). However, $z = 2.53$ for the correlation of self-confidence and performance, which reaches significance, as do all tests for the interpredictor correlations ($p < .05$).

16.3.4.7 Estimating Linear Models Once the meta-analyst has a mean correlation matrix, the next step in model-based meta-analysis is to estimate a function, series of regression equations, or structural model using the matrix of average correlations among the predictors and outcome, along with its covariance matrix. The model-based estimation procedure produces a set of standardized

slopes, and standard errors for those slopes. Individual slopes can be examined to determine whether each differs from zero (that is, the meta-analyst can test whether each $\beta_j^* = 0$), which indicates a statistically significant contribution of the tested predictor to the relevant outcome, adjusting for other included variables. An overall or omnibus test of whether all standardized slopes equal zero is also available, thus the meta-analyst can test the hypothesis that $\beta_j^* = 0$ for all j.

Because standardized regression models express predictor-outcome relationships in terms of standard-deviation units of change in the variables, comparisons of the relative importance of different predictors are possible. As in primary studies, the slopes themselves are interrelated, and appropriate comparisons of the relative importance of slopes account for the covariation between slopes (Becker 2000; Becker and Schram 1994). These comparisons can be accomplished by testing differences between coefficients for the samples of interest (Becker 1992b), or via comparisons of models that constrain the model coefficients to be equal or not (Cheung 2014).

Synthetic partial or semi-partial correlations can also be obtained from the mean by applying the same formulas used with primary-study data to $\hat{\boldsymbol{\rho}}^{RE}$ (or a fixed-effects mean); to obtain appropriate standard errors the variance of the vector $\hat{\boldsymbol{\rho}}^{RE}$ must be taken into account. Aloe and Roberto Toro Rodriguez show that when all effects arise from studies that report on the same variables, computation of these coefficients using the mean correlation matrix produces on average the same point estimates and standard errors as those obtained from univariate synthesis of the analogous partial coefficients (2018).

In many model-driven meta-analyses, the goal is to estimate path models such as those shown in figures 16.1 through 16.3. We can use $\hat{\boldsymbol{\rho}}$ or $\hat{\boldsymbol{\rho}}^{RE}$ (with an appropriate variance) to estimate a model under standard assumptions. We begin by arranging the mean correlations from $\hat{\boldsymbol{\rho}}$ or $\hat{\boldsymbol{\rho}}^{RE}$ in a square form, which we denote $\bar{\mathbf{R}}$. The matrix $\bar{\mathbf{R}}$ is partitioned so that the mean correlations of the predictors with the outcome of interest are in submatrix $\bar{\mathbf{R}}_{XY}$, and the intercorrelations among the predictors appear in $\bar{\mathbf{R}}_{XX}$. Different subsets of $\bar{\mathbf{R}}$ are used depending on the model(s) to be estimated.

Working with the 4×4 matrix of means from the Craft et al. data, the matrix is arranged so that correlations involving sport performance are in the first row and column. We partition the matrix as follows to estimate a model with performance as the outcome. $\bar{\mathbf{R}}_{XY}$ is a vector of three values $(-.034, -.071, .233)'$ and $\bar{\mathbf{R}}_{XX}$ is the lower square matrix:

$$\bar{\mathbf{R}} = \left[\begin{array}{c|ccc} 1 & -.034 & -.071 & .233 \\ \hline -.034 & 1 & .544 & -.453 \\ -.071 & .544 & 1 & -.397 \\ .233 & -.453 & -.397 & 1 \end{array}\right] = \left[\begin{array}{c|c} 1 & \bar{\mathbf{R}}_{XY} \\ \hline \bar{\mathbf{R}}_{XY} & \bar{\mathbf{R}}_{XX} \end{array}\right].$$

Then the product $\mathbf{b}^* = \bar{\mathbf{R}}_{xx}^{-1} \bar{\mathbf{R}}_{XY}$ is computed to obtain the standardized regression slopes in the structural model. Becker and her colleagues provide further details about the estimation of variances and tests associated with such models (Becker 1992b, 1995; Becker and Schram 1994). The variance of $\mathbf{b}^*$ is obtained as a function of the variance of the vector form of $\bar{\mathbf{R}}$ (here, $\text{Var}(\hat{\boldsymbol{\rho}}^{RE})$), or a submatrix of that variance if not all variables are used in the estimated model.

Multilevel structural equation modeling programs can be used to obtain many of these same results using multigroup SEM analyses (Cheung and Chan 2005). The MASEM and TSSEM approaches also provide tests of model quality as well as model-comparison tests that enable the reviewer to evaluate the importance of sets of paths in the model. Both fixed-effects and random-effects approaches exist (Cheung and Cheung 2016).

Example. A path model was estimated for the example data on anxiety and sport behavior, under random-effects assumptions. Figure 16.4 shows that only self-confidence is a significant direct predictor of sport behavior under the random-effects model. (Dashed lines represent paths that were tested but found not significant.) Moreover, the estimates suggest that both cognitive and somatic aspects of anxiety appear to have indirect effects, via self-confidence. Both are significant predictors of self-confidence, even under the random-effects model. The REML estimate of the contribution of somatic anxiety is slightly lower and is not significant.

Presenting more information than just the path coefficients and some indication of their significance on path diagrams makes the images very "busy." Thus it is not a good idea to add standard errors to the display, much as in primary-study applications of SEM. Brown and her colleagues tabled the slope coefficients, their standard errors (SEs), and p-values for z tests computed as the slope-to-SE ratio (2016). This information can also be presented in text as the slopes are discussed. For example, the path from self-confidence to performance shows that

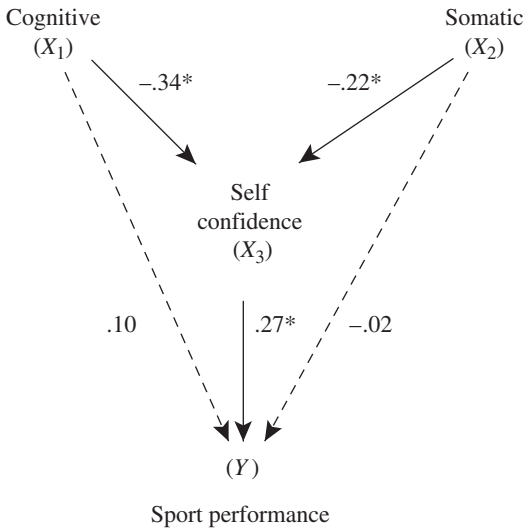

Figure 16.4 Random-Effects Path Coefficients for Prediction of Sport Performance

SOURCE: Authors' tabulation.

a one standard-deviation shift in self-confidence has just over a quarter of a standard deviation impact on sport performance (with a slope of 0.27, $SE = 0.105$, $z = 2.59$, and $p = .001$). Other slopes could be discussed in a similar fashion. Last, the variance explained in performance is estimated using an approach parallel to that used in primary studies (Cooley and Lohnes 1971, 53). The three anxiety subscales explain only 6.08 percent of variation in sport performance, suggesting that considerable variability in performance is not accounted for.

16.3.4.8 Moderator Analyses A last step one might take in the analysis of a model via meta-analysis is to examine categorical moderator variables. To do so, separate models are estimated for the groups of interest and compared; if using Cheung's estimation approach, constraints can be applied then tested to examine whether equality of subgroup slopes holds.

For the anxiety data, we examine the moderator "type of sport activity"—specifically, we have four studies of team sports (type T in table 16.1) versus six of individual sports (type = I). Figure 16.5 shows the slopes from team sports (in the left panel) and individual sports (to the right). Again, dashed lines represent paths with coefficients that do not differ from zero.

Because the data sets are relatively small, coefficients of the same magnitude as found for the full set of studies now do not reach significance, and on the whole, the anxiety measures do not seem to predict sport performance well for team sports. For individual sports, somatic anxiety relates to self-confidence when controlling for cognitive anxiety, whereas the path does not reach significance for team sports. We discuss only one specific comparison for illustrative purposes.

Example. The contribution of self-confidence to performance appears significant for individual sports but not

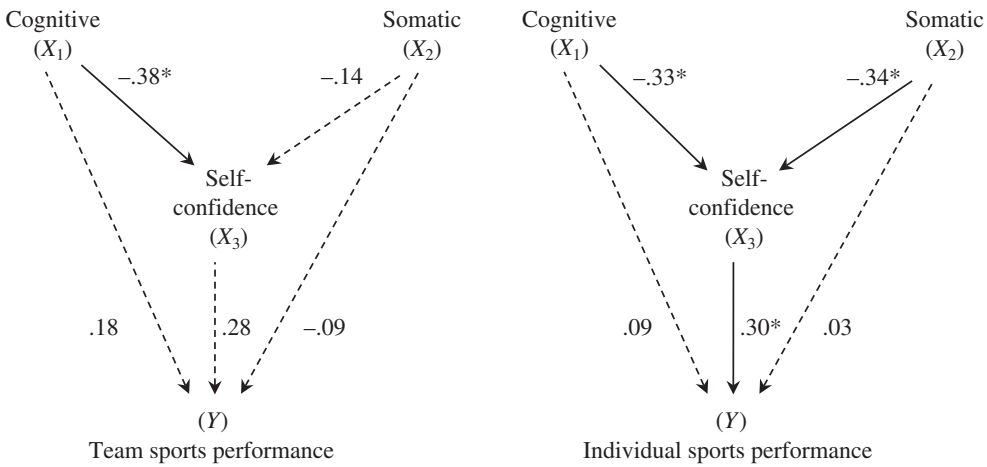

Figure 16.5 Random-Effects Path Coefficients for Prediction by Type of Sport

SOURCE: Authors' tabulation.

for team sports, in spite of being close in value. Because the studies of the two types of sport are independent, we can construct a test of differences in the slopes. The variance-covariance matrices of the slopes for the direct effects on performance in each subgroup are below; the last diagonal element in each is needed for the test:

$$\begin{bmatrix} 0.0813 & 0.0112 & 0.0155 \\ 0.0112 & 0.0082 & 0.0106 \\ 0.0515 & 0.0106 & 0.0431 \end{bmatrix} \quad \begin{bmatrix} 0.0399 & -0.0073 & 0.0170 \\ -0.0073 & 0.0163 & 0.0037 \\ 0.0170 & 0.0037 & 0.0208 \end{bmatrix}$$

Team Sports Performance Individual Sports Performance

In large samples the difference in slopes is approximately normal, thus we use

$$z = \frac{\left(b_T^* - b_I^*\right)}{\sqrt{Var\left(b_T^*\right) + Var\left(b_I^*\right)}} = \frac{(.28 - .30)}{\sqrt{.0431 + .0208}}$$

$$= \frac{-.02}{.2528} = -0.079.$$

This can be compared with critical values of the standard normal distribution. For a two-sided test at the .05 level this difference clearly does not reach significance, so we fail to reject the hypothesis that $\beta_T^* = \beta_I^*$ and conclude that self-confidence is equally important to performance for individual and team sports.

16.3.4.9 Synthetic Partial Correlations In cases where a complex model is not of interest, it may make sense to examine just a single partial correlation. This can be done using a standard univariate meta-analysis of partial correlations, or by estimating what we term a *synthetic partial correlation* matrix from the summary correlation matrix. Using an appropriate estimator of the mean to create the square matrix $\bar{\mathbf{R}}$ for the full set of studies, we obtain the partial correlation matrix via

$$\bar{\mathbf{R}}_{\text{partial}} = -\left(diag\left(\bar{\mathbf{R}}^{-1}\right)\right)^{-\frac{1}{2}} \bar{\mathbf{R}}^{-1} \left(diag\left(\bar{\mathbf{R}}^{-1}\right)\right)^{-\frac{1}{2}}.$$

Estimating the partial correlation matrix in this way can produce negative diagonal elements, which are typically ignored. For our example, using the mean $\hat{\boldsymbol{\rho}}^{RE}$ for $\bar{\mathbf{R}}$ produces a synthetic partial correlation matrix (where the correlation for each pair of variables adjusts for all others) with values

$$\begin{bmatrix} -1 & .080 & -.014 & .237 \\ .080 & -1 & .445 & -.317 \\ -.014 & .445 & -1 & -.193 \\ .237 & -.317 & -.193 & -1 \end{bmatrix}.$$

Thus, in this matrix the synthetic partial correlation of self-confidence and performance adjusting both for cognitive anxiety and somatic anxiety is 0.237, with *SE* = .091. The *SE* is obtained from the variance-covariance matrix of the synthetic partial correlation. It is based on the multivariate delta method and is given in Aloe and Toro Rodriguez (2018). Its values for our example are

$$\begin{bmatrix} 0.0137 & -0.0007 & 0.0053 & -0.0004 & -0.0045 & 0.0016 \\ -0.0007 & 0.0038 & 0.0007 & -0.0007 & -0.0002 & -0.0016 \\ 0.0053 & 0.0007 & 0.0083 & -0.0017 & 0.0005 & -0.0025 \\ -0.0004 & -0.0007 & -0.0017 & 0.0018 & 0.0000 & 0.0016 \\ -0.0045 & -0.0002 & 0.0005 & 0.0000 & 0.0051 & -0.0027 \\ 0.0016 & -0.0016 & -0.0025 & 0.0016 & -0.0027 & 0.0043 \end{bmatrix}.$$

Alternately, the meta-analyst may estimate the desired partial correlation directly for each study, and then synthesize those partial correlations directly. This requires that each study provide the necessary elements to compute the partial *r*. For our data set studies 6 and 17 do not present all the necessary elements to estimate the partial correlation of self-confidence and performance adjusting both for cognitive anxiety and somatic anxiety. The partial correlation values for the 8 samples with sufficient data for computation range from –0.07 to 0.65 and are shown in table 16.2.

A univariate analysis of these values under random effects (see chapter 12) obtained using *metafor* produces a mean partial *r* of 0.328 (*SE* = 0.090). This mean is slightly higher than that from the synthetic partial correlation analysis, and its *SE* is similar to the synthetic partial's *SE* of .091. This is explained in part by the fact that study 16, which contributed data to the synthetic-partial-correlation computation above but not to the direct synthesis of partial *r* values, had the only negative bivariate correlation between self-confidence and performance. The partial correlations are heterogeneous ($Q(7) = 36.88$, $p < .001$) with between-studies variance of 0.047. Thus 95 percent of the true partial correlations are likely to lie between –0.10 and 0.75, a very broad range of population values.

Table 16.2 Self-Confidence and Sports Performance, Adjusting for Cognitive and Somatic Anxiety

ID	*n*	Partial Correlation of Self-Confidence with Performance	Variance
1	142	.536	0.0037
3	37	.332	0.0240
10	14	–.070	0.0990
22	100	.654	0.0034
26	51	.044	0.0212
28	128	.247	0.0071
36	70	.434	0.0100
38	30	–.024	0.0384

SOURCE: Authors' calculations based on Craft et al. 2003.

16.4 SUMMARY AND FUTURE POSSIBILITIES

Although the steps and data required to implement a model-based meta-analysis are definitely more complex and involved than those for other meta-analyses, if the questions of interest are suitable, the benefits of the model-based analyses are well worth the effort. In this chapter, we illustrate the strengths of model-based analyses for examining such complexities as partial relationships, mediating variables, and indirect effects. The small data set used for the examples did not allow for an illustration of the ways model-based meta-analysis can identify areas in need of further exploration, but some of the examples drawn from the literature make clear that even when data are sparse, much can still be learned. For example, Whiteside and Becker discovered that nine of the ninety-one correlations among the fourteen variables in their studies of children of divorce had not been examined (2000). Of key importance was the absence of data on the relations of amount of father visitation, pre-separation father involvement, and current father involvement to cognitive outcomes. This made it impossible to study the roles of these three potentially important variables in child cognitive outcomes.

Using mean correlation matrices to obtain regressions across studies solves the problem of having different scales of measurement—a factor that often limits whether regressions can be directly combined. Becker and Wu compared regression models derived from pooled correlation matrices with the slopes one would obtain from a pooled sample of data, such as an individual-participant-data analysis (2007). They showed that under fixed-effects models, when estimates of mean squared error are available, a summary of slopes from identical models will be equivalent to the same regression model computed from the pooled primary data. However, meta-analytic summaries of slopes may differ from pooled-analysis results when the models from which slopes are drawn are not identical across studies.

One key benefit of estimating partial correlations and slopes from the mean correlation matrix is that covariances among the synthetic partial *r* values or slopes can be obtained. On the other hand, a clear benefit of summarizing partial effects estimated directly from each study is that the meta-analyst can analyze these effects using univariate techniques, estimating between-studies variances in the metric of the partial correlation or slope; these are not available when partial effect values are computed from the mean correlation matrix.

Is the model-based meta-analysis approach of estimating a regression model or synthetic partial correlation from a summary of correlation matrices preferable to a direct synthesis of regression slopes or partial correlations? Aloe and Toro Rodriguez's results suggest that either approach can be applied to obtain mean partial correlations, if appropriate data are available (2018). Their results may apply to other indices such as regression slopes and semi-partial correlations; however this topic needs further investigation.

Technically, for raw slopes or standardized slopes to be comparable across studies, the same set of predictors should appear in all of the regression models to be synthesized. This is almost never the case, as researchers build upon and add to models already appearing in the literature, making exact replications very rare (Makel, Plucker, and Hegarty 2012).

Other complications arise when summarizing sets of regression models, such as the need to consider the scales of both the predictors and outcomes in a synthesis of raw regression slopes (Becker and Wu 2007). If slopes can be transformed to a common scale, or if slopes from standardized regression models are available (Kim 2011), then they can be combined. However, this is often not possible in the social sciences where no "true" scale exists. Because correlations are scale free, this issue is skirted in regressions and path models estimated via model-based meta-analysis. Model-based methods and summaries of correlation matrices give the meta-analyst sensible ways to address complicated questions about interrelationships among variables via research synthesis.

16.5 REFERENCES

Aloe, Ariel M. 2014. "An Empirical Investigation of Partial Effect Sizes for Meta-Analysis of Correlational Data." *Journal of General Psychology* 141: 47–64.

———. 2015. "Inaccuracy of Regression Results in Replacing Bivariate Correlations." *Research Synthesis Methods* 6(1): 21–27.

Aloe, Ariel M., and Betsy J. Becker. 2008. *Modeling Heterogeneity in Meta-Analysis: Generalizing Using Cronbach's (M)UTOS Framework and Meta-Analytic Data.* Paper presented at the annual meeting of the American Educational Research Association, New York.

———. 2009. "Teacher Verbal Ability and School Outcomes: Where Is the Evidence?" *Educational Researcher* 38(8): 612–24.

———. 2012. "A Partial Effect Size for Regression Models." *Journal of Educational and Behavioral Statistics* 37(2): 278–97.

Aloe, Ariel M., and Roberto Toro Rodriguez. 2018. "Synthesis of Multivariate and Univariate Partial Effect Sizes." Unpublished manuscript, University of Iowa.

Aloe, Ariel M., and Christopher G. Thompson. 2013. "The Synthesis of Partial Effect Sizes." *Journal of the Society for Social Work and Research* 4(4): 390–405.

Apling, Richard N. 1981. "Combining the Results of Correlational Studies: Theoretical Considerations and Practical Approaches." Ph.D. diss., Harvard University.

Becker, Betsy J. 1992a. "Models of Science Achievement: Factors Affecting Male and Female Performance in School Science." In *Meta-analysis for Explanation: A Casebook*, edited by Thomas D. Cook, Harris M. Cooper, David S. Cordray, Larry V. Hedges, Heidi Hartmann, Richard J. Light, Thomas A. Louis, and Frederick Mosteller. New York: Russell Sage Foundation.

———. 1992b. "Using Results from Replicated Studies to Estimate Linear Models." *Journal of Educational Statistics* 17(4): 341–62.

———. 1995. "Corrections to 'Using Results from Replicated Studies to Estimate Linear Models.'" *Journal of Educational Statistics* 20(1): 100–102.

———. 2000. "Multivariate Meta-analysis." In *Handbook of Applied Multivariate Statistics and Mathematical Modeling*, edited by Howard E. A. Tinsley and Steven D. Brown. San Diego, Calif.: Academic Press.

———. 2001. "Examining Theoretical Models Through Research Synthesis: The Benefits of Model-Driven Meta-Analysis." *Evaluation and the Health Professions* 24(2): 190–217.

———. 2009. "Model-Based Meta-Analysis." In *The Handbook of Research Synthesis and Meta-Analysis*, edited by Harris M. Cooper, Larry V. Hedges, and Jeffrey C. Valentine. New York: Russell Sage Foundation.

Becker, Betsy J., and Ariel M. Aloe. 2008. *A Framework for Generalization in Meta-analysis: Medical and Social-science Examples*. Keynote presentation at the 16th annual meeting, Merck-Temple Conference on Biostatistics, Philadelphia, PA.

Becker, Betsy J., and Kyle R. Fahrbach. 1994. "A Comparison of Approaches to the Synthesis of Correlation Matrices." Paper presented at the annual meeting of the American Educational Research Association. New Orleans, La.

Becker, Betsy J., and Christine M. Schram. 1994. "Examining Explanatory Models Through Research Synthesis." In *The Handbook of Research Synthesis*, edited by Harris M. Cooper and Larry V. Hedges. New York: Russell Sage Foundation.

Becker, Betsy J., and Meng-Jia Wu. 2007. "The Synthesis of Regression Slopes in Meta-analysis." *Statistical Science* 22(3): 414–29.

Becker, Gilbert. 1996. "The Meta-Analysis of Factor Analyses: An Illustration Based on the Cumulation of Correlation Matrices." *Psychological Methods* 1(4): 341–53.

Berkey, Catherine S., John J. Anderson, and David C. Hoaglin. 1996. "Multiple-Outcome Meta-Analysis of Clinical Trials." *Statistics in Medicine* 15(5): 537–57.

Bowman, Nicholas A. 2010. "College Diversity Experiences and Cognitive Development: A Meta-Analysis." *Review of Educational Research* 80(1): 4–33.

Brown, Sharon A., Betsy J. Becker, Alexandra A. Garcia, Adama Brown, and Gilbert Ramírez. 2015. "Model-Driven Meta-Analyses for Informing Health Care: A Diabetes Meta-Analysis as an Exemplar." *Western Journal of Nursing Research* 37: 517–535.

Brown, Sharon A., Alexandra A. García, Adama Brown, Betsy J. Becker, Vicki S. Conn, Gilbert Ramírez, Mary A. Winter, Lisa L. Sumlin, Theresa J. Garcia, and Heather E. Cuevas. 2016. "Biobehavioral Determinants of Glycemic Control in Type 2 Diabetes: A Systematic Review and Meta-Analysis." *Patient Education and Counseling* 99(10): 1558–67.

Brown, Sharon A., and Larry V. Hedges. 1994. "Predicting Metabolic Control in Diabetes: A Pilot Study Using Meta-Analysis to Estimate a Linear Model." *Nursing Research* 43(6): 362–368.

Cheung, Mike W.-L. 2014. "Fixed- and Random-Effects Meta-Analytic Structural Equation Modeling: Examples and Analyses in R." *Behavior Research Methods* 46(1): 29–40.

———. 2015. "Meta-Analysis: A Structural Equation Modeling Approach." Chichester: John Wiley & Sons.

Cheung, Mike W.-L., and Wai Chan. 2005. "Meta-Analytic Structural Equation Modeling: A Two-Stage Approach." *Psychological Methods* 10(1): 40–64.

Cheung, Mike W.-L., and Shu Fai Cheung. 2016. "Random-Effects Models for Meta-Analytic Structural Equation Modeling: Review, Issues, and Illustrations." *Research Synthesis Methods* 7(2): 140–55.

Cheung, Shu Fai. 2000. "Examining Solutions to Two Practical Issues in Meta-Analysis: Dependent Correlations and Missing Data in Correlation Matrices." Ph.D. diss., Chinese University of Hong Kong.

Cho, Kyunghwa. 2015. "Meta-Analysis of Factor Analyses: Comparison of Univariate and Multivariate Approaches Using Correlation Matrix and Factor Loadings." Ph.D. diss., Florida State University.

Collins, David, Knowlton W. Johnson, and Betsy J. Becker. 2007. "A Meta-Analysis of Effects of Community Coalitions Implementing Science-Based Substance Abuse Prevention Interventions." *Substance Use and Misuse* 42(6): 985–1007.

Cooley, William W., and Paul. R. Lohnes. 1971. *Multivariate Data Analysis*. New York: John Wiley & Sons.

Cooper, Harris M., and Erika A. Patall. 2009. "The Relative Benefits of Meta-Analysis Conducted with Individual Participant Data Versus Aggregated Data." *Psychological Methods* 14(2): 165–176.

Craft, Lynette L., T. Michelle Magyar, Betsy J. Becker, and Debra L. Feltz. 2003. "The Relationship Between the Competitive State Anxiety Index-2 and Athletic Performance: A Meta-Analysis." *Journal of Sport and Exercise Psychology* 25(1): 44–65.

Denson, Nida. 2009. "Do Curricular and Cocurricular Diversity Activities Influence Racial Bias? A Meta-Analysis." *Review of Educational Research* 79(2): 805–38.

DerSimonian, Rebecca, and Nan Laird. 1986. "Meta-Analysis in Clinical Trials." *Controlled Clinical Trials* 7(3): 177–88.

Egger, Matthias, George Davey Smith, Martin Schneider, and Christoph Minder. 1997. "Bias in Meta-Analysis Detected by a Simple, Graphical Test." *British Medical Journal* 315 (September): 629–34.

Fisher, Ronald A. 1921. "On the 'Probable Error' of a Coefficient of Correlation Deduced from a Small Sample." *Metron* 1(1): 1–32.

Furlow, Carolyn F., and S. Natasha Beretvas. 2005. "Meta-Analytic Methods of Pooling Correlation Matrices for Structural Equation Modeling Under Different Patterns of Missing Data." *Psychological Methods* 10(2): 227–54.

Glass, Gene V. 1976. "Primary, Secondary, and Meta-Analysis of Research." *Educational Researcher* 5(1): 3–8.

Gozutok, Ahmet S., Abdullah Alghamdi, and Betsy J. Becker. 2018. "Worrisome News: Many Authors of Single-level and Multilevel Regression Studies still do not Follow the Reporting Standards." Unpublished manuscript, Florida State University.

Hafdahl, Adam R. 2007. "Combining Correlation Matrices: Simulation Analysis of Improved Fixed-Effects Methods." *Journal of Educational and Behavioral Statistics* 32(2): 180–205.

———. 2008. "Combining Heterogeneous Correlation Matrices: Simulation Analysis of Fixed-Effects Methods." *Journal of Educational and Behavioral Statistics* 33(4): 507–33.

Hedges, Larry V. 1982. "Estimation of Effect Size from a Series of Independent Experiments." *Psychological Bulletin* 92(2): 490–99.

———. 1983. "A Random Effects Model for Effect Sizes." *Psychological Bulletin* 93(2): 388–95. DOI:10.1037/0033-2909.93.2.388.

Hedges, Larry V., and Jack L. Vevea. 1998. "Fixed- and Random-Effects Models in Meta-Analysis." *Psychological Methods* 3(4): 486–504.

Hoaglin, David C. 2016. "Misunderstandings About Q and 'Cochran's Q Test' in Meta-Analysis." *Research Synthesis Methods* 35(4): 485–95.

Hodges, William F. 1991. *Interventions for Children of Divorce*, 3rd ed. New York: John Wiley & Sons.

Huang, Chiungjung, and Jyun-Hong Chen. 2015. "Meta-Analysis of the Factor Structures of the Beck Depression Inventory–II." *Assessment* 22(4): 459–72.

Jackson, Dan, Ian R. White, and Simon G. Thompson. 2010. "Extending DerSimonian and Laird's Methodology to Perform Multivariate Random Effects Meta-Analyses." *Statistics in Medicine* 29(12): 1282–97.

Judd, Charles M., and David A. Kenny. 1981. "Process Evaluation: Estimating Mediation in Treatment Evaluations." *Evaluation Review* 5: 602–19.

Kalaian, Hripsime A., and Stephen W. Raudenbush. 1996. "A Multivariate Mixed Linear Model for Meta-Analysis." *Psychological Methods* 1(3): 227–35.

Kaplan, David. 2000. *Structural Equation Modeling: Foundations and Extensions*. Newbury Park, Calif.: Sage Publications.

Kavale, Kenneth A. 1980. "Auditory-Visual Integration and Its Relationship to Reading Achievement: A Meta-Analysis." *Perceptual and Motor Skills* 51(3, Pt 1): 947–55.

Kim, Rae Seon. 2011. "Standardized Regression Coefficients as Indices of Effect Sizes in Meta-Analysis." Ph.D. diss., Florida State University.

Lipsey, Mark W. 1997. "Using Linked Meta-Analysis to Build Policy Models." In *Meta-Analysis of Drug Abuse Prevention Programs (NIDA Research Monograph 170)*, edited by William J. Bukoski. Rockville, Md.: National Institute of Drug Abuse.

Lipsey, Mark W., and James H. Derzon. 1998. "Predictors of Violent or Serious Delinquency in Adolescence and Early Adulthood: A Synthesis of Longitudinal Research." In *Serious and Violent Juvenile Offenders: Risk Factors and Successful Interventions*, edited by Rolf Loeber and David P. Farrington. Thousand Oaks, Calif.: Sage Publications.

MacKinnon, David P., Chondra M. Lockwood, Jeanne M. Hoffman, Stephen G. West, and Virgil Sheets. 2002. "A Comparison of Methods to Test Mediation and Other Intervening Variable Effects." *Psychological Methods* 7(1): 83–104.

Makel, Matthew C., Jonathan A. Plucker, and Boyd Hegarty. 2012. "Replications in Psychology Research: How Often Do They Really Occur?" *Perspectives on Psychological Science* 7(6): 537–42.

Martens, Rainer, Robin S. Vealey, and Damon Burton, eds. 1990. *Competitive Anxiety in Sport*. Champaign, Ill.: Human Kinetics.

Mathur, Maya B., Elissa Epel, Shelley Kind, Manisha Desai, Christine G. Parks, Dale P. Sandler and Nayer Khazeni. 2016. "Perceived Stress and Telomere Length: A Systematic Review, Meta-analysis, and Methodologic Considerations for Advancing the Field." *Brain, Behavior, and Immunity* 54 (May):158–69.

Olkin, Ingram, and Minoru Siotani. 1976. "Asymptotic Distribution of Functions of a Correlation Matrix." In *Essays in Probability and Statistics*, edited by Sadeo Ikeda et al. Tokyo: Shinko Tsusho Co.

Pearl, Judea. 2009. "Causal Inference in Statistics: An Overview." *Statistics Surveys* 3: 96–146.

Perry, Ryan, Chris G. Sibley, and John Duckitt. 2013. "Dangerous and Competitive Worldviews: A Meta-Analysis of Their Associations with Social Dominance Orientation and Right-Wing Authoritarianism." *Journal of Research in Personality* 47(1): 116–27.

Prevost, A. Toby, Dan Mason, Simon Griffin, Ann-Louise Kinmonth, Stephen Sutton, and David Spiegelhalter. 2007. "Allowing for Correlations Between Correlations in Random-Effects Meta-Analysis of Correlation Matrices." *Psychological Methods* 12(4): 434–50.

Raudenbush, Stephen W., Betsy J. Becker, and Hripsime Kalaian. 1988. "Modeling Multivariate Effect Sizes." *Psychological Bulletin* 103(1): 111–20.

Raudenbush, Stephen W., and Anthony S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, Calif.: Sage Publications.

Richardson, W. Scott, Mark W. Wilson, Jim Nishikawa, Robert S. A. Hayward. 1995. "The Well-Built Clinical Question: A Key to Evidence-Based Decisions." *ACP Journal Club* 123(3): A12-3.

Riley, Richard. 2009. "Multivariate Meta-Analysis: The Effect of Ignoring Within-Study Correlation." *Journal of the Royal Statistical Society, Series A* 172(4): 789–811.

Riley, Richard D., Abrams, Keith R., Lambert, Paul C., Sutton, Alexander J., Thompson, John R. 2007. "Bivariate Random Effects Meta-Analysis and the Estimation of Between-Study Correlation." *BMC Medical Research Methodology* 7 (January): 3. DOI: 10.1186/1471-2288-7-3.

Sidik, Kurek, and Jeffrey N. Jonkman. 2005. "Simple Heterogeneity Variance Estimation for Meta-Analysis." *Applied Statistics* 54(2): 367–84.

Steinkamp, Marjorie W., and Martin L. Maehr. 1980. "Affect, Ability, and Science Achievement: A Quantitative Synthesis of Correlational Research." *Review of Educational Research* 53(3): 369–96.

Stewart, Lesley A., and Mike J. Clarke. 1995. "Practical Methodology of Meta-Analyses (Overviews) Using Updated Individual Patient Data." *Statistics in Medicine* 14(19): 2057–79.

Thompson, Christopher G., Ariel M. Aloe, and Betsy J. Becker. 2018. "Synthesizing Estimators from Regression Models of Different Sizes." Unpublished manuscript, Texas A&M University.

Tipton, Elizabeth. 2015. "Small Sample Adjustments for Robust Variance Estimation with Meta-regression." *Psychological Methods* 20(3): 375–93.

Valentine, Jeffrey C., David L. DuBois, and Harris M. Cooper. 2004. "The Relation Between Self-Beliefs and Academic Achievement: A Meta-Analytic Review." *Educational Psychologist* 39(2): 111–33. DOI: 10.1207/s15326985ep3902_3.

van Houwelingen, Hans C., Lidia R. Arends, and Theo Stijnen. 2003. "Advanced Methods in Meta-Analysis: Multivariate Approach and Meta-Regression." *Statistics in Medicine* 21(4): 589–624.

Viana, Marlos A. G. 1982. "Combined Estimators for the Correlation Coefficient." *Communications in Statistics—Theory and Methods* 11(13): 1483–504.

Viechtbauer, Wolfgang. 2010. "Conducting Meta-Analyses in R with the *metafor* Package." *Journal of Statistical Soft-*

ware 36(1): 1–48. Accessed December 12, 2018. http://www.jstatsoft.org/v36/i03/.

White, Ian R. 2009. "Multivariate Random-Effects Meta-Analysis." *Stata Journal* 9(1): 40–56.

Whiteside, Mary F., and Betsy J. Becker. 2000. "The Young Child's Post-Divorce Adjustment: A Meta-Analysis with Implications for Parenting Arrangements." *Journal of Family Psychology* 14(1): 1–22.

Wu, Meng-Jia, and Betsy J. Becker. 2013. "Synthesizing Regression Results: A Factored Likelihood Method." *Research Synthesis Methods* 4(2): 127–43.

Yang, Janet Z., Ariel M. Aloe, and Thomas H. Feeley. 2014. "Risk Information Seeking and Processing Model: A Meta-Analysis." *Journal of Communication* 64(1): 20–41.

Young, Bessie Ann, Elizabeth Lin, Michael Von Korff, Greg Simon, Paul Ciechanowski, Evette J. Ludman, Siobhan Everson-Stewart, Leslie Kinder, Malia Oliver, Edward J. Boyko, and Wayne J. Katon. 2008. "Diabetes Complications Severity Index and Risk of Mortality, Hospitalization, and Healthcare Utilization." *American Journal of Managed Care* 14(1): 15–23.

PART VI

DATA DIAGNOSTICS

17

MISSING DATA IN META-ANALYSIS

TERRI D. PIGOTT
Loyola University Chicago

CONTENTS

17.1 INTRODUCTION

This chapter presents strategies for examining the sensitivity of meta-analysis results to missing data on predictors in linear models of effect size, or meta-regression. Researchers using meta-analysis will invariably find that studies in a research synthesis differ in the types and quality of the information reported. The chapter accordingly discusses the types of missing data that occur in a research synthesis, and statistical methods researchers can use to explore the sensitivity of their results to missing data. The goal of methods for handling missing data in any statistical analysis is not to recover the values for the missing observations, but instead to examine the extent to which missing data may potentially affect the conclusion drawn in the analysis.

Researchers should attempt to recover any missing data as a first strategy in any meta-analysis by contacting the authors of the primary study included in the meta-analysis. When recovery of the missing information is unsuccessful, the next strategy is to use statistical methods for missing data to check the sensitivity of results to the presence of missing data. The robustness of meta-analysis results can be assessed by providing evidence about the potential risk of bias of results under different assumptions about the missing data. The main goal of statistical methods for missing data is to make valid inferences about a population of interest (Graham and Schafer 1999). This chapter introduces methods for testing the sensitivity of meta-analysis results when missing data occur.

The focus is on missing data in the meta-analysis of study-level data, particularly when missing data occurs in predictors of effect-size models. Although researchers have developed missing data methods for individual participant data meta-analysis (IPD), it is not covered in this chapter (Burgess et al. 2013; Jolani et al. 2015; Quartagno and Carpenter 2016). Missing data also occurs in more complex meta-analysis models such as network meta-analysis but these emerging methods will also not be addressed here.

17.2 TYPES OF MISSING DATA

Researchers conducting systematic reviews encounter missing data in three ways in a meta-analysis: in the form of missing studies, missing effect sizes (or information needed to compute effect sizes), and missing study descriptors that could be used in an effect-size model. As will be detailed later, understanding both the type of missing data and the potential reasons for the missingness are important for choosing strategies for handling missing data. Although the reasons for missing observations on any of these three areas vary, each type of missing data presents difficulties for the analysis.

17.2.1 Missing Studies

A number of mechanisms might lead to studies missing in a research synthesis. Researchers in both medicine and in the social sciences have documented the bias in the published literature toward statistically significant results (see, for example, Rosenthal 1979; Hemminki 1980; Smith 1980; Begg and Berlin 1988; Rothstein, Sutton, and Borenstein 2005). Chapter 18 in this volume discusses the identification of publication bias in a meta-analysis and methods for examining the sensitivity of results to the presence of publication bias. Another reason studies may be missing in a synthesis is accessibility. Some studies may be unpublished reports that are not identifiable through commonly used search strategies. For example, Matthias Egger and George Davey (1998) and Peter Jüni and his colleagues (2002) both demonstrate that studies published in languages other than English are not only more difficult to identify, but might also produce results different from studies in English.

Researchers undertaking a comprehensive synthesis expend significant effort identifying and obtaining unpublished studies to maintain the representative nature of the sample for the synthesis. Strategies for preventing, assessing and adjusting for publication bias are examined, as mentioned, in the next chapter (see also Rothstein, Sutton, and Borenstein 2005). The focus here is therefore on missing data within studies, not on how to handle publication bias.

17.2.2 Missing Effect Sizes

A common problem in research syntheses is missing information for computing an effect size. Researchers commonly encounter studies that are missing the descriptive statistics needed to compute an effect size, such as the group means and standard deviations required for the computation of a standardized mean difference. In this scenario, the author of the primary study did not report enough information to compute an effect size. An-Wen Chan and colleagues refer to these studies as having incompletely reported outcomes (2004). Studies may

also only partially report outcomes, such as including only the value of the effect size, the sample size, or *p*-value, but not the values of the summary information needed to compute the effect size directly. Studies with qualitatively reported outcomes may include only the *p*-value, with or without the sample size. In some cases, an estimated effect size could be computed based on the *p*-value and sample sizes, or the value of the test statistic. David Wilson's effect-size calculator can be used to compute effect sizes from a variety of statistics (2016).

Effect sizes might also be missing from a study because a study has selectively reported outcomes. The issue of selective outcome reporting has been well documented in the medical literature (Hutton and Williamson 2000; Vedula et al. 2009; Kirkham et al. 2010), and in education (Pigott et al. 2013). Primary researchers may have a range of reasons for not reporting all the outcomes collected in a study. In some cases, authors may not have space in the article to provide information about all outcomes, and thus may report only on outcomes considered central. In other cases, they may omit outcomes because statistical tests of those outcomes were not statistically significant. The issue of outcome reporting bias is of great interest in the medical and social science literature at present given concerns about the overall quality of research and subsequent policy decisions (Ioannidis 2005). Effect sizes are often missing from primary studies because the statistical tests for these outcomes are not statistically significant.

A problem related to missing effect sizes is missing outcome data within studies. If an individual patient does not have a measure for the target outcome, then that patient cannot provide any information about the efficacy of the treatment, leading to a potentially biased estimate of the effect size. Studies have explored methods for missing outcome data in clinical trials, methods that could lead reviewers to compute a more accurate effect size for a study included in a meta-analysis (Higgins, White, and Wood 2008; Jackson et al. 2014). Another strategy for handling missing outcome data in a study is the use of pattern mixture models (Little 1993; Andridge and Little 2011). Dimitris Mavridis and his colleagues provide an overview of methods for addressing missing outcome data in meta-analysis (2014). Mavridis and his colleagues later build on Roderick Little's work on pattern-mixture models, developing techniques for missing outcome data in study-level and network meta-analysis (Mavridis et al. 2015; Little 1993). Studies may also fail to report information about the sample variances needed to compute an effect size. Missing variances of measures collected in a study leads directly to missing effect sizes. Amit Chowdhry, Robert Dworkin, and Michael McDermott discuss a method to handle missing sampling variances within studies and test its performance through simulation (2015). More research is needed on whether these techniques developed for meta-analyses in medicine could apply to the social sciences where the meta-analyses include a larger number of studies, and where reporting standards are less rigorous.

When effect sizes are missing from a study, most reviewers drop these studies from the analysis. As discussed, researchers in medicine are developing a number of strategies for handling the problem of missing effect sizes, and these methods could prove useful beyond the medical literature. The remainder of this chapter focuses on missing data on predictors used for examining variation in effect sizes across studies.

17.2.3 Missing Predictor Variables

A major task in the data evaluation phase of a meta-analysis involves coding aspects of a study's design and methods. Studies may be missing potential predictors that might be used as moderators in an effect-size model because the primary authors did not collect this information or did not report it. Missing descriptor variables are an inherent problem in meta-analysis given that not every study author will collect or report the same information. For example, Seokyung Hahn and colleagues were interested in studying whether the effect of prophylactic malaria treatment for pregnant women differs as a function of whether a woman is pregnant for the first time (2000). However, not all studies reported on whether study participants were experiencing their first pregnancy and thus they could not complete this analysis.

Missing data on study descriptor variables arise because of differences among research synthesists and primary researchers in the importance placed on particular information. Synthesists may have hypotheses about how variation in study methods and procedures might relate to variation in effect size, whereas primary authors are concerned about the design and implementations of a single study. One way that missing descriptor variables in a study might occur is due to disciplinary practices in a given field. Robert Orwin and David Cordray use the term *macrolevel reporting* to refer to practices in a given research area that influence how constructs are defined and reported (1985). For example, Selcuk Sirin finds

several measures of socioeconomic status reported in his meta-analysis of the relationship between academic achievement and socioeconomic status (SES) (2005). These measures include parent scores on Hollingshead's occupational status scale, parental income, parental education level, free lunch eligibility in school and composites of several of these measures. Differences in the types of SES measures reported could derive from traditional ways disciplines have reported SES. Education researchers may tend to have more access to free lunch eligibility status when gathering data, whereas economists may use large data sets with income data reported.

Primary authors also differ from each other in writing style and thoroughness of reporting. Orwin and Cordray refer to individual differences of primary authors as *micro-level reporting* quality (1985). In this case, whether a given descriptor is reported across a set of studies may be random given that it depends on individual writing preferences and practices. Primary authors may also be constrained by a particular journal's publication practices, and thus do not report on information a synthesis considers important. Individual journal reporting constraints would likely result in descriptors missing randomly from a study.

Study descriptors may also be missing from primary study reports because of selective reporting. As in outcome reporting bias, primary study authors may selectively report on study descriptors because of the actual values of those descriptors. For example, a primary author might not report the gender of the participant sample if an outcome does not differ significantly by gender. It is also hard to imagine that a primary author might report on the racial and ethnic background of study participants when the sample is more homogeneous than the author intended. Synthesists often find that measures of attrition from a trial are missing, and many might assume that the amount of attrition could have influenced the decision to include the values in the report. When information about descriptor variables are missing because of their values, this selective reporting cannot be considered random, and leads to problems discussed later in this chapter. Also, potential study-level predictors of effect-size models may be reported as if they were completely observed within the study. For example, studies might report the average income level of participants based only on those study participants that provided their income level. Reviewers should consider coding the percentage of missing data on study-level descriptor variables to understand the impact of missing data at both the study level and at the level of the meta-analysis.

No matter what the reasons for missing data, researchers using meta-analysis also need to examine closely the pattern of the missing predictors. Meta-analysts often have a number of effect-size models they wish to fit in the presence of heterogeneity, and these models include a number of predictors. Missing data could affect each of these variables separately, and may dramatically reduce the number of cases that include complete data on two or more of these variables. Each model fit in a meta-analysis may draw on a different sample of cases from the originally identified studies.

17.3 REASONS FOR MISSING DATA

Roderick Little and Donald Rubin remains the seminal work on statistical analysis with missing data (2002). The strategies they discuss rely on specific assumptions the data analyst is willing to make about the reasons for missing data, and about the joint distribution of all variables in the data. Rubin had introduced a framework for describing the response mechanism, the reasons for why data may be missing in a given research study in 1976. This framework describes three assumptions about the nature of the missing data: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). These assumptions are described in the context of typical meta-analytic data in the following sections. The potential mechanism for the missing data is critical for choosing an analysis strategy when missing data occur, as we see later in the chapter.

17.3.1 Missing Completely at Random

Data are missing completely at random when the observed data can be considered a random sample of the originally collected data. For example, primary study authors might differ on whether they report on particular details of a study. Educational researchers could differ on whether they report the exact enrollment of the schools in their sample. If school size is not an important factor in a given study, there may be no reason to suspect that the average size of schools in a given primary study is missing because of the value of the school. Thus a research synthesis might postulate that average school size could be missing randomly from the set of studies identified for the synthesis. When data are MCAR, we assume that the probability of the values being missing are not related to their unobserved values or to values of any other observed

or unobserved variables in the data set. When data are MCAR, the studies with completely observed data can be considered a random sample of the studies originally identified for the synthesis.

17.3.2 Missing at Random

We refer to data as missing at random when the probability of missing a value on a variable is not related to the missing value, but may be related to other observed variables in the data. For example, Selcuk Sirin reports on a number of moderator analyses in his meta-analysis of the relationship between SES and academic achievement (2005). One moderator examined is type of components of measured SES. These include parental education, parental occupation, parental income, and eligibility for free or reduced lunch. A second source is source of information on SES, whether parent, student, or a secondary source. Imagine if all studies report the source of information on SES but not all report the component used. It is likely that the source of information on SES is highly related to the component of SES measured in the study. Students are much less likely to report income, but may be more likely to report on parental education or occupation. Secondary sources of income are usually derived from eligibility for free or reduced lunch. In this case, the component of SES is missing at random because we can assume that the likelihood of observing any given component of SES depends on the completely observed value, source of SES information. Note that MAR refers to what is formally called the response mechanism. This is generally unobservable, unless a researcher can gather direct information from participants about why they did not respond to a given survey question, or from primary authors about why they did not collect a particular variable.

17.3.3 Not Missing at Random

We refer to data as not missing at random when the probability of observing a given value for a variable is related to the missing value itself. For example, effect sizes are not missing at random when they are not reported in a study because they are not statistically significant (see Chan et al. 2004; Williamson et al. 2005). In this example, data are not missing at random due to a censoring mechanism, which results in certain values having a higher probability of being missing than other values. Meta-analysts have developed methods to handle censored effect-size data in the special case of publication bias (for example, Hedges and Vevea 1996; Vevea and Woods 2005; chapter 18, this volume).

Missing study descriptors could also be NMAR. Primary authors whose research participants are racially homogeneous may not report fully on the distribution of ethnicities among participants as a way to provide a more positive picture of their study. As with the effect-size example, the actual ethnic distribution of the samples are related to their probability of being missing. When potential predictors in effect-size models are missing because of reporting bias, the response mechanism is NMAR.

Synthesists rarely have direct evidence about the reasons for missing data and need to assume data are MCAR, MAR, or NMAR when they are conducting sensitivity analyses to examine the robustness of meta-analysis results to missing data. In the rest of the chapter, I discuss the options available to meta-analysts when faced with various forms of missing data.

17.4 COMMONLY USED METHODS

Before Little and Rubin, most researchers used one of three strategies to handle missing data: using only those cases with all variables completely observed (listwise deletion), using available cases that have particular pairs of variables observed (pairwise deletion), or replacing missing values in a given variable with a single value such as the mean for the complete cases (single-value imputation) (2002). These methods will not uniformly produce results that are defensible.

17.4.1 Complete Case Analysis

In complete case analysis, the researcher uses only those cases with all variables fully observed. This procedure, also known as listwise deletion, is usually the default procedure for many statistical computer packages. When some cases are missing values of a particular variable, only cases observing all the variables in the analysis are used. When the missing data are MCAR, the complete cases can be considered a random sample from the originally identified set of cases. Thus, a synthesist can make the assumption that values are in fact missing completely at random, using only complete cases will produce unbiased results.

Complete case analysis for models of effect size is likely one of the most common methods used when missing data on predictors occurs. Research synthesists typically use only those studies reporting on a given predictor

when estimating a model of effect-size variation. But in meta-analysis as in other statistical analyses, using only complete cases can seriously limit the number of observations available for the analysis. Losing cases decreases the power of the analysis and ignores the information contained in the incomplete cases (Kim and Curry 1977; Little and Rubin 2002). When data are NMAR or MAR, complete case analysis yields biased estimates because the complete cases cannot be considered representative of the original sample. If studies are missing predictors for a given model and these predictors are missing either because of their values or due to the values of other observed predictors in the model, using only complete cases will lead to bias in the estimates of the model. For example, say the authors are interested in how the efficacy of an intervention differs for high- versus low-income students. If income information tends to be missing for studies with higher-income students, then using only complete cases will lead to a biased estimate of the relationship between treatment efficacy and students' income level. If the probability of observing income is related to another observed variable, such as achievement, complete case analysis in a model with only income-predicting effect size will also lead to a biased estimate.

17.4.1.1 Complete Case Analysis Example Table 17.1 presents a subset of studies from a meta-analysis examining the effects of oral anticoagulant therapy for patients with coronary artery disease (adapted from Sonia Anand and Salim Yusuf 1999). To illustrate the use of missing data methods with MCAR and MAR data, I used two methods for deleting the age of the study in the total data set. For MCAR data, I randomly deleted ten values (30 percent) of study age from the data. For MAR data, I randomly deleted the value of study age from ten of the studies that reported providing a high dose of oral anticoagulants to patients. The last two columns of table 17.1 indicate the cases that are missing study age under the conditions of MCAR and MAR. This simple example used throughout the chapter is suggestive of issues that may arise with missing data; more rigorous simulation studies are needed to understand how these methods work in meta-analysis.

Table 17.2 compares the complete case analysis results of a random-effects meta-regression for examining the variation in the log-odds ratio across studies as a function of the magnitude of the oral anticoagulant dose and the age of the study. Two dummy variables were coded for dose—high and moderate. Although the model as a whole is not significant, we can examine the value of the estimates for each coefficient. The complete case analysis with MCAR data provides results consistent in effect direction with the results from the original data set. The coefficients for high and moderate doses are smaller than the original data. The complete case results with MAR data, however, differ in both direction and magnitude. Unless data can be considered MCAR, complete case analysis will produce biased estimates (Schafer 1997; Enders 2010).

17.4.2 Available Case Analysis

An available case analysis, also called pairwise analysis, estimates parameters using as much data as possible. An available case analysis uses all the complete cases for the estimates of the means, and for bivariate statistics, such as correlations, uses all possible pairs of observations. In table 17.1, we would use all the cases to estimate the correlation between the log-odds ratio and dose, but only those complete cases for the correlation between the log-odds ratio and age. If there were an additional variable in the data with missing data, there could be another set of cases that we would use to estimate this variable's correlation with the log-odds ratio.

This simple example illustrates the drawback of using available case analysis: each correlation in the variance-covariance matrix estimated using available cases could be based on different subsets of the original data set. If data are MCAR, these subsets are individually representative of the original data, and available case analysis provides unbiased estimates. If the data are MAR, however, these subsets are not individually representative of the original data and will produce biased estimates.

Much of the early research on methods for missing data focuses on the performance of available case analysis versus complete case analysis (for example, Glasser 1964; Haitovsky 1968; Kim and Curry 1977). Kyle Fahrbach examines the research on available case analysis and concludes that such methods provide more efficient estimators than complete case analysis when correlations between two independent variables are moderate, that is, around 0.6 (2001). This view, however, is not shared by all who have examined this literature (see, for example, Allison 2002). One statistical problem that could arise from the use of available cases under any form of missing data is a nonpositive definite variance-covariance matrix, that is, a variance-covariance matrix that cannot be inverted to obtain the estimates of slopes for a regression model. One reason for this problem is that different subsets of studies are used to compute the elements of the

Table 17.1 Oral Anticoagulant Therapy Data

ID	Log-Odds Ratio	Variance of Log-Odds Ratio	Intensity of Dose	Year of Publication	Age of Study	MCAR Data	MAR Data
1	3.02	2.21	High	1960	39	0	0
2	–1.84	0.8	High	1960	39	1	0
3	0.24	0.16	High	1961	38	0	1
4	0.15	0.07	High	1961	38	1	0
5	0.47	0.07	High	1964	35	1	0
6	0.38	0.28	High	1964	35	1	1
7	–0.38	0.18	High	1966	33	1	0
8	–0.47	0.22	High	1967	32	0	1
9	–0.25	0.07	High	1967	32	1	0
10	0.22	0.1	High	1969	30	0	0
11	0.84	0.08	High	1969	30	0	0
12	0.35	0.04	High	1980	19	1	0
13	0.33	0.02	High	1990	9	0	1
14	0.12	0.01	High	1994	5	1	1
15	–1.81	0.82	High	1969	30	1	1
16	0.43	0.02	High	1974	25	1	1
17	0.18	0.06	High	1980	19	1	1
18	0.39	0.07	High	1980	19	0	1
19	0.9	0.41	High	1993	6	1	1
20	0.65	0.29	High	1996	3	0	0
21	1.42	2.74	High	1990	9	1	1
22	0.14	4.06	Moderate	1990	9	0	1
23	0.04	1.36	Moderate	1982	17	1	1
24	0.35	0.35	Moderate	1981	18	1	1
25	0.08	0.03	Moderate	1982	17	1	1
26	0.06	0.03	Moderate	1969	30	1	1
27	0.43	0.07	Moderate	1964	35	1	1
28	–1.16	0.93	Moderate	1986	13	1	1
29	0.75	0.98	Moderate	1982	17	1	1
30	0.81	0.6	Moderate	1998	1	1	1
31	0.04	0.82	Moderate	1994	5	1	1
32	0.35	0.70	Low	1998	1	0	1
33	0.34	0.06	Low	1997	2	1	1
34	0.15	0.02	Low	1997	2	1	1

SOURCE: Author's tabulation based on Anand and Yusuf 1999.
NOTE: In last two columns, 0 = study age is missing in this scenario.

Table 17.2 Complete Case Results

	Full Data		MCAR Data		MAR Data	
Coefficient	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	0.213	0.142	0.214	0.153	0.193	0.122
High dose	0.060	0.180	0.014	0.209	–0.008	0.152
Moderate dose	–0.033	0.216	–0.012	0.239	–0.149	0.206
Age of study	–0.001	0.005	–0.002	0.006	0.004	0.006

SOURCE: Author's compilation.

variance-covariance matrix. Further, Paul Allison points out that a more difficult problem in the application of available case analysis concerns the computation of standard errors of available case estimates (2002). At issue is the correct sample size when computing standard errors, given that each parameter could be estimated with a different subset of data. Some of the standard errors could be based on the full data set, and others on the subset of studies that observe a particular variable or pair of variables. Most statistical computing packages provide the option of available case analysis or listwise deletion, but how standard errors are computed differs widely.

Although available case analysis appears sensible given that it uses all available data, consensus is scant in the literature about the conditions where available case analysis outperforms complete case analysis when data are MCAR. A version of available case analysis occurs when research synthesists fit a number of different models to the data, and thus inadvertently use different sets of cases for each model fit. For example, if we were exploring different models for the oral anticoagulant data, we might first fit a model with dose, and then add study age. If we were not careful about the missing data, we might try to compare these two models despite the fact that they use different sets of cases. Harris Cooper calls this strategy a shifting units of analysis approach, a common method in meta-analysis (2017). Reviewers need to be aware of the different samples used to fit effect-size models.

17.4.3 Single-Value Imputation Methods

When values are missing in a meta-analysis (or any statistical analysis), many researchers replace the missing value with a reasonable value, such as the mean for the cases that observed the variable. Little and Rubin refer to this strategy as single-value imputation (2002). In meta-analysis, researchers may fill in the complete case mean for a missing predictor value. Another strategy uses regression with the complete cases to estimate predicted values for missing observations given the observed values in a particular case. Single-value imputation methods do not provide accurate standard errors for any statistical analysis because the filled-in data set is treated as if it had no incomplete cases. The sample sizes for all analyses will be that for the original data set without accounting for the uncertainty caused by missing values. The problems with single-value imputation are illustrated in the following section.

17.4.3.1 Imputing the Complete Case Mean Replacing the missing values in a variable with the complete case mean of the variable is also referred to as unconditional mean imputation. When we substitute a single value for all the missing values, we decrease the variation in that variable. The estimated variance thus does not reflect the true uncertainty in the variable. Instead, the smaller variance wrongly indicates more certainty about the value than is warranted. Whenever we fill in the mean of a missing predictor in a meta-analysis model, the variance associated with that filled-in variable will be decreased. Table 17.3 compares the results of using mean imputation for the missing values of study age for the MCAR and MAR data. Note that different mean values are used to fill in the missing values for the MCAR and MAR data as the complete case mean for these two scenarios differ. For the MCAR data, the missing values were filled in with a mean study age of 20.08, whereas the missing values for the MAR data used a value of 16.42. (In comparison, the mean study age of the full data is 20.35). The estimates and their standard errors for the MCAR data are consistent with the full data values, wrongly reflecting the amount of certainty in the estimates given that 30 percent of the studies are missing study age. The estimates in the MAR data differ in magnitude and direction. They also have standard errors that do not reflect the actual uncertainty present due to missing values. Under the assumption of MCAR, imputing the complete case mean might provide parameter estimates close to the true values, but the standard errors will not be reflective of the uncertainty caused by missing data. It is likely that filling in a mean value for a missing predictor could lead a researcher to find statistical significance when it is not warranted.

17.4.3.2 Single-Value Imputation with Conditional Means A single-value imputation method that provides less biased results with missing data was first suggested a half century ago by S. F. Buck (1960). Each missing

Table 17.3 Mean Imputation Results

	MCAR		MAR	
Coefficient	Estimate	SE	Estimate	SE
Intercept	0.215	0.144	0.197	0.128
High dose	0.063	0.189	0.001	0.160
Moderate dose	–0.029	0.226	–0.132	0.216
Age of study	–0.001	0.006	0.003	0.006

SOURCE: Author's compilation.

value is replaced not with the complete case mean but instead with the predicted value from a regression model using the variables observed in that particular case as predictors and the missing variable as the outcome. This method is also referred to as conditional mean imputation or regression imputation. For each pattern of missing data, the cases with complete data on the variables in the pattern are used to estimate regressions using the observed variables to predict the missing values. The result is that each missing value is replaced by a predicted value from a regression using the values of the observed variables in that case. When data are MCAR, each of the subsets used to estimate prediction equations are representative of the original sample. This method results in more variation than unconditional mean imputation because the missing values are replaced with those that depend on the regression equation. However, the standard errors using Buck's method are too small because the missing values are replaced with predicted values that lie directly on the regression line used to impute the values. In other words, Buck's method yields values predicted exactly by the regression equation without error.

Little and Rubin present the form of the bias for Buck's method and suggest corrections to the estimated variances to account for the bias (2002). If we have two variables, Y_1 and Y_2 and Y_2 has missing observations, then the form of the bias using Buck's method to fill in values for Y_2 is given by

$$(n - n^{(2)})(n-1)^{-1}\sigma_{22.1}$$

where n is the sample size, $n^{(2)}$ is the number of cases that observe Y_2, and $\sigma_{22.1}$ is the residual variance from the regression of Y_2 on Y_1. Little and Rubin also provide the more general form of the bias with more than two variables. This correction is applied to the variance-covariance matrix of the variables. To include this correction, the researcher would need to use the variance-covariance matrix to estimate the desired model.

Table 17.4 compares the results from regression imputation with MCAR and MAR data. The imputation uses both the value of the effect sizes and the dose to generate the regression imputed values. The estimates in both analyses differ from the full data set. The MCAR estimates for the coefficient of high dose are twice as large as those for the full data. Most notably, the standard errors are similar across all analyses, and thus do not reflect the uncertainty in the data. Thus, although regression imputation adds some variation to the estimates, the standard errors still do not reflect the uncertainty caused by missing data. In this simple example, even the MCAR estimates are biased.

Table 17.4 Regression Imputation Results

	MCAR		MAR	
Coefficient	Estimate	SE	Estimate	SE
Intercept	0.223	0.153	0.206	0.136
High dose	0.137	0.191	0.030	0.173
Moderate dose	0.074	0.230	–0.075	0.221
Age of study	–0.006	0.005	0.001	0.006

SOURCE: Author's compilation.

Survey researchers have used a number of other single-value imputation methods to prevent the loss of information from missing data including such strategies as hot-deck imputation, similar response pattern imputation and last observation carried forward (Enders 2010). Any single-value imputation strategy that does not adjust standard errors for the uncertainty caused by missing data will produce biased estimates no matter what mechanism leads to the missing observations.

17.4.4 Summary of Commonly Used Methods

When missing predictors in a meta-analysis are MCAR, complete case analysis can yield unbiased results, and the standard errors will reflect the sample size of the complete cases used. Meta-analysts should be aware of the pattern of missing predictors so that they do not inadvertently use available case analysis when fitting effect-size models. Single-value imputation methods will always underestimate the standard errors of the estimates and are not recommended for any meta-analysis. Researchers who use complete case analysis should be aware and should report that they are assuming that the missing data are MCAR in the analysis.

17.5 MODEL-BASED METHODS

Unfortunately, most commonly used methods for missing data produce biased estimates even when the data can be considered MCAR. Although the assumption of MCAR may be viable for some predictors in a meta-analysis model, the assumption of MAR data—that the probability

for a missing predictor depends on another fully observed variable—may be more defensible in a meta-analysis. None of the commonly used methods produce unbiased estimates in the presence of MAR data. The general problem with commonly used methods is that they do not take into account the distribution of the hypothetically complete data. For example, filing in the complete case mean for a missing observation may be a reasonable assumption, but it is not based on a distribution for that variable. The missing data methods outlined begin with a model for the observed data, namely, that the data are distributed as multivariate normal, and assume that the missing data mechanism is ignorable, an assumption discussed later.

17.5.1 Assumptions for Model-Based Methods

The general approach used in current missing data methods involves using the data at hand to draw valid conclusions, and not to recover all the missing information to create a complete data set. This approach is especially applicable to meta-analysis because missing data frequently occur because a variable was not measured and is not recoverable. Researchers faced with missing data in meta-analysis should aim at testing the sensitivity of results to the presence of missing data rather than attempting to recreate a complete data set.

Model-based methods for missing data make strong assumptions about the distribution of the data, and about the mechanism that causes the missing observations. The set of methods most applicable to meta-analysis require the assumption that the joint distribution of the effect size and predictor variables is multivariate normal. A second assumption is that the reasons for the missing data do not depend on the values of the missing observations, that the missing data mechanism is either MCAR or MAR. As discussed, one major difficulty in applying missing data methods is that assumptions about the nature of the missing data mechanism cannot be tested empirically. These assumptions can only be subjected to the is-it-possible test, that is, is it possible that the reasons for missing observations on a particular variable do not depend directly on the values of that variable? Missing observations on income usually fail the test, because a well-known result in survey sampling is that respondents with higher incomes tend not to report their earnings. The following section examines the assumptions needed for model-based methods for missing data in the context of meta-analysis.

17.5.1.1 Multivariate Normality The missing data methods rely on the assumption that the joint distribution of the data is multivariate normal. Thus, meta-analysts must assume that the joint distribution of the effect sizes and the variables coded from the studies in the review follow a normal distribution. One problematic issue in meta-analysis concerns the common incidence of categorical predictors in effect-size models. Codes for characteristics of studies often take on values that indicate whether a primary author used a particular method, such as random assignment, or a certain assessment for the outcome, such as standardized protocol or test, researcher developed rating scale, and so on. Joseph Schafer indicates that in the case of categorical predictors, the normal model will still prove useful if the categorical variables are completely observed, and the variables with missing observations can be assumed multivariate normal conditional on the variables with complete data (1997). This assumption holds in the oral anticoagulant data because the missing values of study age in the MAR data are randomly deleted within the studies that provided a high dose of the drug. We can still fulfill the multivariate normality condition if we can assume that the variable with missing observations is normally distributed conditional on a fully observed categorical variable. Some ordered categorical predictors can also be transformed to allow the normal assumption to apply. If key moderators of interest are non-ordered categorical variables, and these variables are missing observations, then missing data methods based on the multinomial model may apply. Although currently no research addresses how to handle missing categorical predictors in meta-analysis, researchers do discuss methods for nonnormal missing data that may hold promise for meta-analysis (see, for example, Schafer 1997; White, Royston, and Wood 2011).

17.5.1.2 Ignorable Response Mechanism Rubin discusses in detail the conditions under which a missing data mechanism is ignorable (1987). One of these conditions is that the probability of observing a value does not depend on the value that is missing, a condition that holds for both MCAR and MAR data. The two major model-based methods, maximum likelihood estimation using the EM algorithm and multiple imputation, will provide unbiased estimates with MCAR and MAR data. The MAR assumption holds only when the completely observed variables related to the probability of response for missing values are included in the model. Thus, research synthesists using multiple imputation, for example, should include as many variables as possible when creating multiple imputations for a meta-analysis.

17.5.2 Maximum Likelihood Methods Using EM Algorithm

Although maximum likelihood methods for missing data are widely used for missing data analysis, they have limited use in meta-analysis. Maximum likelihood methods using the EM algorithm (Dempster, Laird, and Rubin 1977) provide unbiased estimates for the means and variance-covariance matrix given an ignorable response mechanism and distributional assumptions about the data. As discussed elsewhere, maximum likelihood methods for missing data in regression provide estimates of the sample means and covariance matrix, which can then be used to obtain the coefficients of the regression model (Little and Rubin 2002). A meta-regression model uses weighted least squares to account for the differences in precision of effect sizes rather than ordinary least squares assumed when using maximum likelihood methods for missing data using the EM algorithm. It is not clear how to estimate effect-size regression models using weighted least squares from the means and variance-covariance matrix of meta-analytic data (Draper and Smith 1981). Thus, these methods currently have limited application in the meta-analysis literature. Little and Rubin provide a number of extensions of the EM algorithm that may prove useful for meta-analysis (2002).

17.5.3 Multiple Imputation for Multivariate Normal Data

Multiple imputation has become the method of choice in many contexts of missing data. The main advantage of multiple imputation is that the analyst uses the same statistical procedures in the analysis phase that were planned for completely observed data (Rubin 1987). In other words, in the analysis phase of multiple imputation, the researcher does not need to adjust standard errors as in Buck's method from 1960, and does not need to estimate a regression from the covariance matrix as in maximum likelihood with the EM algorithm. Multiple imputation, as its name implies, is a technique that generates multiple possible values for each missing observation in the data. Each of these values is used in turn to create a complete data set. The analyst uses standard statistical procedures to analyze each of these multiply-imputed data sets, and then combines the results of these analyses for statistical inference.

Multiple imputation consists of three phases. The first involves the generation of the possible values for each missing observation. The second phase then analyzes each completed data set using standard statistical procedures. The third phase combines the estimates from the analyses of the second phase to obtain results to use for statistical inference. Each of these phases is discussed conceptually in the following sections (for more detail, see Enders 2010; Schafer 1997). A final note concerns the use of multiple imputation in small samples. John Graham and his colleagues provide evidence that multiple imputation performs best in samples of at least fifty cases (Graham 2009; Graham and Schafer 1999). Although many meta-analyses in the social sciences include at least fifty studies, in other contexts (such as medicine), the number of available studies is much smaller. Research is needed to understand the performance of multiple imputation in small meta-analytic data sets.

17.5.3.1 Generating Multiple Imputations Multiple imputation relies on a model for the distribution of missing data given the observed data under the condition of MAR or MCAR data. Multiple imputation uses Bayesian methods to obtain random draws from the posterior predictive distribution of the missing observations given the observed observations. These random draws are completed in an iterative process. Given the means and covariance matrix of our hypothetically complete multivariate normal data, we can obtain the form of the distribution of the missing observations given the observed data, and draw a random observation from that distribution. That observation would be one plausible value for a missing value for a given case. Once we have drawn plausible values for all our missing observations, we obtain a new estimate of our means and covariance matrix, and repeat the process. We assume that our response mechanism is ignorable so that the posterior distribution does not include a specification of the response mechanism.

To generate these random draws, however, we need to use simulation techniques such as Markov Chain Monte Carlo. These methods allow the use of simulation to obtain random draws from a complex distribution. This phase is the most complex statistically, but many commercial software packages, including freeware, are available to generate these imputations, especially in cases when we can assume the complete data is multivariate normal. In R, the program Amelia II can generate these imputations. Major statistical packages such as SAS, STATA and SPSS also include programs for multiple imputation.

One basic issue in multiple imputations is the choice of the number of imputed data sets to generate and analyze. Little and Rubin (2002) and Schafer (1997) both recommend between three to five imputed data sets. However,

more recent recommendations suggest many more imputations to obtain the lowest possible standard errors and to improve the power of the analysis (Enders 2010). John Graham, Allison Olchowski, and Tamika Gilreath show that using more than ten imputations improves the power of the analysis and generally recommend one hundred imputations (2007). Craig Enders indicates that researchers use a minimum of twenty (2010). Ian White and his colleagues also discuss the number of imputations needed as a function of the percentage of missing data in the sample (2011).

Another issue is the question of what variables to include in the imputation model. Schafer suggests using as many complete variables in the data set as possible in order to capture the mechanism that may cause the missing data under the assumption of MAR (1997). Linda Collins, Schafer, and Chi-Ming Kam provide a thorough discussion and simulation study exploring strategies for choosing variables for the multiple imputation model (2001). In general, they recommend using a more inclusive strategy for the imputations, even if the data analysis does not include all of the variables used in the imputation model. Most meta-analysts code a large number of characteristics of studies, and thus should have many variables to use to create imputations. In this chapter, I assume that effect sizes are completely observed in the data set, and thus, effect sizes would also be included in the model used to generate the imputations. Another issue is whether the standard errors of the standardized mean difference should be included in the imputation model. The examples provided later in the chapter omit the standard errors from the imputation model because the standard errors are not strictly predictors of the effect-size magnitude in a meta-regression. More research is needed about whether standard errors of the standardized mean difference should be included in the imputation model.

17.5.3.2 Analyzing Completed Data Sets In this second step, the researcher obtains a series of completed data sets, with each missing observation filled in using the methods in the prior section. Once the imputations are generated, the analyst uses whatever methods were originally planned for the data. These analyses are repeated for each completed data set. In this phase, the analyst takes each completed data set and obtains estimates for the originally planned model. The researcher also needs to take into account the effort involved in conducting the planned analysis for each imputed data set. If the analysis is complex, twenty imputations may increase the overall cost of the meta-analysis.

17.5.3.3 Combining Estimates Rubin (1987) and Enders (2010) provide the formulas for combining the multiply-imputed estimates to obtain overall estimates and their standard errors. Let us denote the mean of our target estimate for the *j*th parameter across all *j* imputations as

$$\overline{q_i} = \frac{1}{m}\sum_{j=1}^{m} q_{ij},$$

where q_{ij} is the estimate of the *i*th parameter from the *j*th completed data set, and m is the number of imputations. To obtain the standard errors of the q_{ij}, we need two estimates of variance. The within-imputation variance of the estimate of q_i is the arithmetic average of the m sampling variances of q_{ij}, or

$$vw_i = \frac{1}{m}\sum_{j=1}^{m} se(q_i)^2.$$

The between-imputation variance of q_{ij} is the variability of the estimate of q_i across the m data sets, or

$$vb_i = \frac{1}{m-1}\left(q_{ij} - \overline{q_i}\right)^2.$$

The standard error of the point estimates of the q_i is given by

$$SE(q_i) = \sqrt{vw_i + vb_i + \frac{vb_i}{m}}.$$

As discussed earlier, multiple imputation is based on large sample theory, and using these methods with small samples may lead to biased estimates. John Barnard and Rubin provide a correction to account for small sample sizes when conducting hypothesis tests of multiply-imputed estimates for small samples (1999).

Table 17.5 presents the results using multiple imputation with ten imputations for the oral anticoagulant data with study age deleted using an MAR mechanism. The effect size and dose were used in the model to generate the missing values for study age. All of the estimates of the coefficients of the meta-regression model are in the same direction as the full data. Notably, the standard errors are much larger, providing a more conservative estimate and reflecting the fact that missing data causes uncertainty in the data. If we were checking the sensitiv-

Table 17.5 Multiple Imputation Results

	MAR	
Coefficient	Estimate	SE
Intercept	0.216	0.380
High dose	0.078	0.428
Moderate dose	–0.004	0.485
Age of study	–0.002	0.076

SOURCE: Author's compilation.

ity of our meta-analysis results to the presence of missing data, we would compare our complete case results in the MAR data in table 17.2 to those in table 17.5. In this small example, two of the coefficients have opposite signs in the two analyses, suggesting that our results may be sensitive to missing information. Note that this simple example uses less than the minimum of imputations recommended, and that the data set includes fewer than fifty cases. The results from this example suggest that results are sensitive to missing data, and caution is needed in the interpretation of our results.

Multiple imputation methods are now widely used in statistical analysis, and much guidance exists about the use of multiple imputation in a variety of contexts (Enders 2010). Although meta-analysts have not uses multiple imputation extensively, many statistical packages can generate multiple imputations for meta-analysis that can then be used within many of the meta-analysis programs available. For the example here, I used the Amelia program in R (Honaker, King, and Blackwell 2015) to generate multiple imputations, and *metafor* in R (Viechtbauer 2010) to estimate the random-effects meta-regression model. The multiply-imputed estimates of the model can be combined in R or Excel.

17.6 RECOMMENDATIONS

Missing data are ubiquitous in meta-analysis. Research synthesists will encounter missing data because studies differ not only in the methods used but also in the completeness of their reporting of information. When missing data occur among predictors in meta-analytic models, the researcher should first understand the patterns of missing data and how those patterns might impact the models that the researchers want to estimate. Fitting a series of models to meta-analytic data without exploring the missing data patterns may lead a researcher to use widely different sets of cases to estimate each model.

Once a research synthesis understands the pattern of missing data, a next step is to explore the sensitivity of the results to the missing data. If a researcher can make the case that the missing predictors are MCAR, then the complete case analysis will provide unbiased estimates. The cost, however, of MCAR data may be power for the analysis if the quantity of missing data is large. A more realistic assumption is MAR, particularly if the meta-analysis includes a large number of study descriptors that are completely observed. In the case of a rich meta-analysis data set, multiple imputation provides a method to test the sensitivity of results to missing data in predictors. Given that methods for multiple imputation are implemented in many different computing packages, meta-analysts should increase their use of multiple imputation to test the robustness of their results. As discussed, multiple imputation is a model-based method that requires the researcher to make clear assumptions about their data. Model-based methods for missing data are thus more defensible than the ad hoc methods typically used in meta-analysis, particularly for sensitivity analysis.

Advances in missing data methods for primary studies may lead to additional strategies for missing data in meta-analysis. Social scientists need to explore the use of methods for missing outcome data in meta-analysis that have been developed in medicine. In addition, methods for testing the sensitivity of meta-analysis results when data are not MCAR or MAR are also needed.

Missing data in meta-analysis occurs because many primary studies do not report on variables that could be used to examine the variation in results across studies. Efforts to increase the transparency and reporting quality of primary studies will alleviate some of the problems missing data cause in meta-analysis. However, it is likely that missing data will continue to impact meta-analysis. In order to ensure that the results of meta-analysis are unbiased, researchers should test the sensitivity of their results to missing data, increasing the transparency of both the potential implications and limitations of their results.

17.7 REFERENCES

Allison, Paul D. 2002. *Missing Data.* Thousand Oaks, Calif.: Sage Publications.

Anand, Sonia S., and Salim Yusuf. 1999. "Oral Anticoagulant Therapy in Patients With Coronary Artery Disease." *Journal of the American Medical Association* 282(21): 2058–67.

Andridge, Rebecca R. and Roderick J. A. Little. 2011. "Proxy Pattern-Mixture Analysis for Survey Nonresponse." *Journal of Official Statistics* 27(2): 153–80.

Barnard, John, and Donald B. Rubin. 1999. "Small-Sample Degrees of Freedom with Multiple Imputation." *Biometrika* 86(4): 948–55.

Begg, Colin B., and Jesse A. Berlin. 1988. "Publication Bias: A Problem in Interpreting Medical Data (with Discussion)." *Journal of the Royal Statistical Society Series A* 151(2): 419–63.

Buck, S. F. 1960. "A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer." *Journal of the Royal Statistical Society Series B* 22(2): 302–303.

Burgess, Stephen, Ian R. White, Matthieu Resche-Rigon, and Angela M. Wood. 2013. "Combining Multiple Imputation and Meta-Analysis with Individual Participant Data." *Statistics in Medicine* 32: 4499–514.

Chan, An-Wen, Asbjørn Hróbjartsson, Mette T. Haahr, Peter, C. Gøtzsche, and Douglas G. Altman. 2004. "Empirical Evidence for Selective Reporting of Outcomes in Randomized Trials." *Journal of the American Medical Association* 291(20): 2457–65.

Chowdhry, Amit K., Robert H. Dworkin, and Michael P. McDermott. 2015. "Meta-Analysis with Missing Study-Level Sample Variance Data." *Statistics in Medicine* 35(17): 3021–32.

Collins, Linda M., Joseph L. Schafer, and Chi-Ming Kam. 2001. "A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures." *Psychological Methods* 6(4): 330–51.

Cooper, Harris M. 2017. *Research Synthesis and Meta-Analysis: A Step-by-Step Approach,* 5th ed. Los Angeles, Calif.: Sage Publications.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society Series B* 39(1): 1–38.

Draper, Normal, and Harry Smith. 1981. *Applied Regression Analysis*, 2nd ed. New York: John Wiley & Sons.

Egger, Matthias, and George Davey. 1998. "Meta-Analysis: Bias in Location and Selection of Studies." *British Medical Journal* 316(7124): 61–66.

Enders, Craig K. 2010. "Applied Missing Data Analysis." New York: The Guildford Press.

Fahrbach, Kyle R. 2001. "An Investigation of Mixed-Model Meta-Analysis in the Presence of Missing Data." Ph.D. diss., Michigan State University.

Glasser, Marvin. 1964. "Linear Regression Analysis with Missing Observations among the Independent Variables." *Journal of the American Statistical Association* 59(307): 834–44.

Graham, John W. 2009. "Missing Data Analysis: Making It Work in the Real World." *Annual Review of Psychology* 60(1): 549–76.

Graham, John W., Allison E. Olchowski, and Tamika D. Gilreath. 2007. "How Many Imputations Are Really Needed? Some Practical Clarifications of Multiple Imputation Theory." *Prevention Science* 8(3): 206–13.

Graham, John W., and Joseph L. Schafer. 1999. "On the Performance of Multiple Imputation for Multivariate Data with Small Sample Size." In *Statistical Strategies for Small Sample Research*, edited by Rick H. Hoyle. Thousand Oaks, Calif.: Sage Publications.

Haitovsky, Yoel. 1968. "Missing Data in Regression Analysis." *Journal of the Royal Statistical Society Series B* 30(1): 67–82.

Hedges, Larry V., and Jack L. Vevea. 1996. "Estimating Effect Size Under Publication Bias: Small Sample Properties and Robustness of a Random Effects Selection Model." *Journal of Educational and Behavioral Statistics* 21(4): 299–332.

Hemminki, Elina 1980. "Study of Information Submitted by Drug Companies to Licensing Authorities." *British Medical Journal* 280(6217): 833–36.

Hahn, Seokyung, Paula R. Williamson, Jane L. Hutton, Paul Gamer, and E. Victor Flynn. 2000. "Assessing the Potential for Bias in Meta-Analysis Due to Selective Reporting of Subgroup Analyses Within Studies." *Statistics in Medicine* 19(24): 3325–36.

Higgins, Julian P. T., Ian R. White, and Angela M. Wood. 2008. "Imputation Methods for Missing Outcome Data in the Meta-Analysis of Clinical Trials." *Clinical Trials* 5(3): 225–39.

Honaker, James, Gary King, and Matthew Blackwell. 2015. Amelia II: A Program for Missing Data, Version 1.7.4.

Hutton, J. L., and Paula R. Williamson. 2000. "Bias in Meta-Analysis Due to Outcome Variable Selection Within Studies." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 49(3): 359–70.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Med* 2(8): e124. DOI: 10.1371/journal.pmed.0020124.

Jackson, Dan, Ian R. White, Dan Mason, and Stephen Sutton. 2014. "A General Method for Handling Missing Binary Outcome Data in Randomized Controlled Trials." *Addiction* 109(12): 1986–93.

Jolani, Shahab, Thomas P. A. Debray, Hendrik Koffijberg, Stef van Buuren, and Karel G. M. Moons. 2015. "Imputation

of Systematically Missing Predictors in an Individual Participant Data Meta-Analysis: A Generalized Approach Using MICE." *Statistics in Medicine* 34: 1841–63.

Jüni, Peter, Franziska Holenstein, Jonathon Sterne, Christopher Bartlett, and Matthias Egger. 2002. "Direction and Impact of Language Bias in Meta-Analyses of Controlled Trials: Empirical Study." *International Journal of Epidemiology* 31(1): 115–23.

Kim, Jae-On, and James Curry. 1977. "The Treatment of Missing Data in Multivariate Analysis." *Sociological Methods and Research* 6(2): 215–40.

Kirkham, Jamie J., Kerry M. Dwan, Douglas G. Altman, Carrol Gamble, Susanna Dodd, Rebecca Smyth, and Paula R. Williamson. 2010. "The Impact of Outcome Reporting Bias in Randomised Controlled Trials on a Cohort of Systematic Reviews." *British Medical Journal* 340: c365.

Little, Roderick J. A. 1993. "Pattern-Mixture Models for Multivariate Incomplete Data." *Journal of the American Statistical Association* 88(421): 125–34.

Little, Roderick J. A., and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*, 2nd ed. New York: John Wiley & Sons.

Mavridis, Dimitris, Anna Chaimani, Orestis Efthimiou, Stefan Leucht, and Georgia Salanti. 2014. "Addressing Missing Outcome Data in Meta-Analysis." *Evidence Based Mental Health* 17(3): 85–89.

Mavridis, Dimitris, Ian R. White, Julian P. T. Higgins, Andrea Capriani, and Georgia Salanti. 2015. "Allowing for Uncertainty Due to Missing Continuous Outcome Data in Pairwise and Network Meta-Analysis." *Statistics in Medicine* 34(5): 721–41.

Orwin, Robert G., and David S. Cordray. 1985. "Effects of Deficient Reporting on Meta-Analysis: A Conceptual Framework and Reanalysis." *Psychological Bulletin* 97(1): 134–47.

Pigott, Terri D., Jeffrey C. Valentine, Joshua R. Polanin, Ryan T, Williams, and Dericka D. Canada. 2013. "Outcome-Reporting Bias in Education Research." *Educational Researcher* 42(8): 424–32.

Quartagno, Matteo, and James R. Carpenter. 2016. "Multiple Imputation for IPD Meta-Analysis: Allowing for Heterogeneity and Studies with Missing Covariates." *Statistics in Medicine* 35(17): 2938–54.

Rosenthal, Robert. 1979. "The 'File Drawer Problem' and Tolerance for Null Results." *Psychological Bulletin* 86(3): 638–41.

Rothstein, Hannah R., Alexander J. Sutton, and Michael Borenstein, eds. 2005. *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments.* Chichester, UK: John Wiley & Sons.

Rubin, Donald B. 1976. "Inference and Missing Data." *Biometrika* 63(3): 581–92.

———. 1987. *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data.* New York: Chapman & Hall/CRC.

Sirin, Selcuk R. 2005. "Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research." *Review of Educational Research* 75(3): 417–53.

Smith, Mary L. 1980. "Publication Bias and Meta-Analysis." *Evaluation in Education* 4(1): 22–24.

Vedula, S. Swaroop, Lisa Bero, Roberta W. Scherer, and Kay Dickersin. 2009. "Outcome Reporting In Industry-Sponsored Trials of Gabapentin for Off-Label Use." *New England Journal of Medicine* 361 (November): 1963–71.

Vevea, Jack L., and Carol M. Woods. 2005. "Publication Bias in Research Synthesis: Sensitivity Analysis Using A Priori Weight Functions." *Psychological Methods* 10(4): 428–43.

Viechtbauer, Wolfgang. 2010. "Conducting Meta-Analysis in R with the *metafor* Package." *Journal of Statistical Software* 36(3): 1–48.

White, Ian R., Patrick Royston, and Angela M. Wood. 2011. "Multiple Imputation Using Chained Equations: Issues and Guidance for Practice." *Statistics in Medicine* 30(4): 377–99.

Willamson, Paula R., Carrol Gamble, Douglas G. Altman, and J. L. Hutton. 2005. "Outcome Selection Bias in Meta-Analysis." *Statistical Methods in Medical Research* 14(5): 515–24.

Wilson, David B. 2016. "Practical Meta-Analysis Effect Size Calculator." Accessed December 13, 2018. http://www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-Home.php.

18

PUBLICATION BIAS

JACK L. VEVEA
University of California, Merced

KATHLEEN COBURN
University of California, Merced

ALEXANDER SUTTON
University of Leicester

CONTENTS

18.1 INTRODUCTION

There is good evidence to suggest that unpublished scientific results may systematically differ from published results, because selectivity may exist in deciding what to publish (see Dickersin, Min, and Meinert 1991, 1992; Song et al. 2000; Dickersin 2005). That phenomenon is frequently referred to as publication bias. For example, researchers may choose not to write up and submit studies with uninteresting or nonsignificant findings, or such studies may not be accepted for publication. Although publication bias refers to whether work is published, unpublished work still available for inclusion in meta-analyses does not technically contribute to bias in those

specific meta-analyses, even though the published studies themselves are a biased sample.

Examples of publication bias are everywhere. Philippa Easterbrook and her colleagues document the role of the perceived importance of findings in determining which to submit for publication (1991). Allan Coursol and Edwin Wagner present evidence of the role of statistical significance in the publication process (1986). Jerome Stern and John Simes offer evidence that significant results are often published more quickly (1997). An-Wen Chan and his colleagues point out that, even if a study is published, there may be selectivity in which aspects are presented; significant outcomes may be given precedent over nonsignificant ones (Chan, Hrobjartsson, et al. 2004; Chan, Krleza-Jeric, et al. 2004). That is, any selection mechanism may operate through suppression of particular results within a study, or all results from a particular sample may be affected. Sven Kepes and his colleagues discuss the distinction (2012). Additionally, research with positive or statistically significant results may be published in more prestigious venues and cited more times, making it more visible and easier to find (Koricheva 2003; Egger and Smith 1998). Indeed, the publication process should be thought of as a continuum and not a dichotomy (Smith 1999). For example, material that has been published with incomplete reporting in a journal may have been circulated with full reporting as a working paper. In keeping with the previous literature, these biases will be referred to simply as publication bias throughout the chapter, although dissemination bias is perhaps a more accurate name for the collection (Song et al. 2000).

In areas where any such selectivity exists, the literature is biased. That is true whether one is reading a single journal article or conducting a synthesis of many. Publication bias is therefore a major threat to the validity not only of meta-analysis and other synthesis methodologies, but also of the research literature itself. Indeed, one could argue that meta-analysis provides a partial solution to the problem, because researchers can at least attempt to identify and estimate the effect of such bias by considering the information contained in the distribution of effect sizes from the available studies. That is the basis of the majority of statistical methods described here. It is important to note that most of these methods have been developed for use with the meta-analytic models advanced in the tradition of Larry Hedges and Ingram Olkin (1985). Many methods for testing and correcting publication bias are not suitable for the psychometric meta-analysis approaches proposed by James Hunter and Frank Schmidt (1990).

Researchers agree that prevention is the best solution to the problem of selectively reported research. Indeed, with advances in electronic publishing making the presentation of large amounts of information more economically viable than traditional paper-based publishing methods, there is some hope that the problem will diminish, if not disappear. Many have suggested that open-access publication can assist with the problem (see, for example, Joober et al. 2012), but much of this advocacy appears in blog entries or in the mission statements of electronic journals. There is still little or no empirical evidence of such an effect. Ridha Joober and his colleagues also point to the possibility that high fees associated with open-access publication could actually lead to publication bias (2012). Moreover, open access does not offer a solution to the suppression of information due to vested economic interests (Halpern and Berlin 2005).

Jesse Berlin and Davina Ghersi, among others, have advocated the use of prospective registries of studies for selecting studies to be included in systematic reviews (2005). The practice provides an unbiased sampling frame guaranteeing the elimination of publication bias (relating to the suppression of whole studies, at least). However, trial registration does not guarantee availability of data, and an obligation to disclose results in an accessible form is also required. Registries exist for randomized controlled trials in numerous medical areas, and there is an expectation that this practice will ultimately reduce publication bias (Zarin et al. 2011). Such a solution will not be feasible for some forms of research, however, including research relating to analysis of observational data, where the notion of a study that can be registered before analysis may be nonexistent. The idea of registries for research in the social sciences has been put forth but it is far from the norm, and controversy surrounds the effectiveness of preregistration in that context (Anderson 2013; Gelman 2013; Humphreys, de la Sierra, and van der Windt 2013; Monogan 2013). The notion of prospectively designing multiple studies with the intention of carrying out a meta-analysis in the future has also been put forward as a solution to the problem (Berlin and Ghersi 2005), but again may be difficult to orchestrate in many situations.

Carrying out as comprehensive a search as possible when obtaining literature for a synthesis will help minimize the influence of publication bias. In particular, this may involve searching for studies not formally published (chapter 6 in this volume; Hopewell, Clarke, and Mallett 2005), as well as using methods other than simple electronic

searches (such as journal browsing and reference chasing). Since the beginning of the internet, the feasibility and accessibility of publication by means other than commercial publishing houses have greatly increased.

Despite researchers' best efforts, at least in the current climate, alleviation of the problem of publication bias may not be possible in many areas of science. In such instances, graphical and statistical tools have been developed to address publication bias within a meta-analysis framework. The remainder of this chapter provides an overview of these methods. If a research synthesis does not contain a quantitative synthesis (for example, if the data being synthesized are not quantitative), publication bias may still be a problem, but methods to deal with it are limited to prevention through registration and rigorous literature searches (Petticrew et al. 2006). Terese Bondas and Elisabeth Hall suggest that careful identification of unpublished studies, such as dissertations, may help, but that has not proven to be consistently effective for quantitative synthesis, so it may be of limited value for qualitative synthesis (2016). Simon Lewin and his colleagues observe that evidence of publication bias in qualitative literature is lacking (2015). They also state that methodological advances are in development, but are not currently available.

18.2 MECHANISMS THAT CAUSE PUBLICATION BIAS

There is considerable discussion in the literature about the precise nature of the mechanisms that lead to suppression of whole studies and other forms of publication bias. These mechanisms may operate on specific results within a particular study (outcome bias) or on the entire study (dissemination bias). Both of these levels can contribute to the overall presence of publication bias (Kepes et al. 2012). If these mechanisms could be accurately specified and quantified, then the appropriate adjustments to a meta-analytic data set would be straightforward. However, measuring such effects is difficult, and the mechanisms vary with data set and subject area.

Evidence is ample that statistical significance, effect magnitude and direction, study size, and other factors can all influence the likelihood of a study being published. Colin Begg and Jesse Berlin address the role of *p*-values and direction of effect (1988). Harris Cooper, Kristina DeNeve, and Kelly Charlton confirm the existence of filters in the research process other than bias against the null hypothesis (1997). Robert Rosenthal and John Gaito present evidence for cliff effects associated with conventional levels of significance (1963, 1964), as do Nanette Nelson, Robert Rosenthal, and Ralph Rosnow (1986). Deborah Barnes and Lisa Bero show that funding source can lead to selection bias (1998). Justin Bekelman, Yan Li, and Cary Gross discuss the role of industry funding (2003). Kathleen Coburn and Jack Vevea mention industry funding and preferences for results that are consistent with current beliefs, trends, and cultural expectations as sources of bias (2015; see also Kepes, Banks, and Oh 2014; Kepes, Bennett, and McDaniel 2014). José Duarte and his colleagues also provide evidence that social preferences can influence publication (2015), citing the work of Stephen Abramowitz, Beverly Gomes, and Christine Abramowitz, that liberal reviewers were less likely to publish research with results favoring conservatives (1975), and of Stephen Ceci, Douglas Peters, and Jonathan Plotkin, that "reverse discrimination" proposals were approved less often (1985).

The sections that follow outline and demonstrate methods to identify and adjust for publication bias. These methods assume different underlying mechanisms for publication bias, and all of those assumptions are wrong. Accordingly, the focus in these sections includes not only an up-to-date overview of available methods, but also attention to the assumptions of each approach. It is not plausible, for example, that publication bias occurs solely because of statistical significance, or that it arises purely from a relationship between effect size and standard error, or that it follows a deterministic pattern, such as elimination of the largest negative effects. For that reason, these methods should be regarded as tools for sensitivity analysis, and triangulation using multiple techniques is essential. Kepes and McDaniel propose reporting a range of estimates across various methods to assess the effect of publication bias (2015). Reporting standards for meta-analysis endorsed by the American Psychological Association (2008) and the Cochrane Collaboration (Higgins and Green 2011) also recommend this approach. Despite these suggestions, evidence indicates that only about 3 percent of meta-analyses use more than two procedures to address publication bias (Ferguson and Brannick 2012; van Enst et al. 2014).

18.3 METHODS FOR IDENTIFYING PUBLICATION BIAS

18.3.1 The Funnel Plot

Since its introduction by Richard Light and David Pillemer in 1984, the funnel plot has been a preferred exploratory

tool for investigating publication bias and, like the forest plot, for presenting a visual summary of a meta-analytic data set (Sterne, Becker, and Egger 2005). In its original form, the funnel plot is a scatterplot with effect estimates on the horizontal axis and sample size on the vertical axis. Sample size is closely related to study precision, which is usually defined as the reciprocal of either the sampling variances or the standard errors of the effect sizes. More recent forms of the funnel plot typically use such a measure of precision (or, alternatively, the reciprocal of precision) in place of sample size. The expectation is that the plot should appear symmetric with respect to the distribution of effect sizes and should resemble a funnel. The effect sizes should be evenly distributed around the underlying true effect size, and show more variability in the smaller studies than the larger ones because of the greater influence of sampling error. This results in a funnel-shaped plot that narrows as study precision increases. If publication bias is present, we might expect some suppression of smaller, unfavorable, and nonsignificant studies that could be identified by a gap in one corner of the funnel or a decrease in density nearer the center of the funnel, inducing asymmetry in the plot.

Figure 18.1 depicts a relatively symmetric funnel plot produced in the traditional manner, using simulated data. However, that mode of presentation goes against the convention of plotting an unknown quantity on the y-axis and a fixed quantity (such as *N*) on the x-axis. Figure 18.2 shows the same plot with the more standard graphics conventions. Funnels in both orientations exist in the literature. The true effect in both plots is 0.5.

It is interesting that study suppression caused by study size, effect size, or statistical significance (one-sided), either individually or in combination, could produce an asymmetric funnel plot. It is also possible for two-sided statistical significance suppression mechanisms (that is, significant studies in either direction are more likely to be published) to could create a tunnel or hole in the middle of the funnel, particularly when the underlying effect size is close to zero. However, in most circumstances, it is implausible that a selection mechanism based on two-tailed *p*-values would function the same way in both tails.

The most appropriate axes for the funnel plot is debated, particularly with respect to the measure of study precision (Vevea and Hedges 1995; Sterne and Egger 2001). Variance (and its inverse) and standard error (and its inverse) are options for use in place of sample size. This choice can affect the appearance of the plot considerably. For instance, if the variance or standard error is used, the

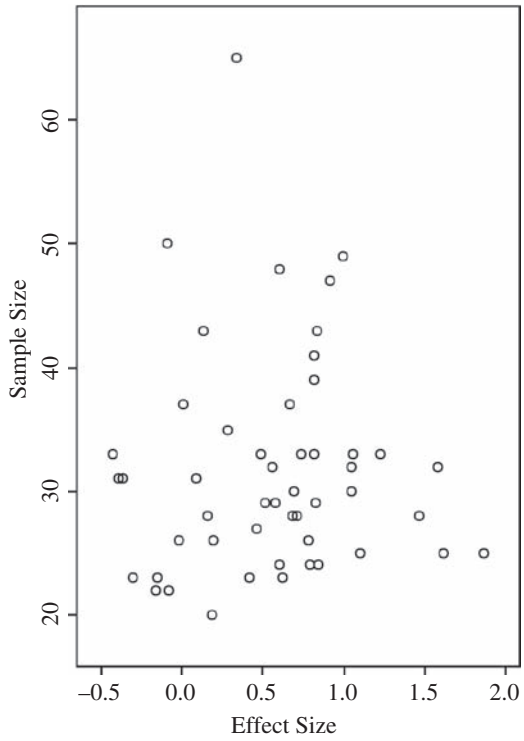

Figure 18.1 Traditional Funnel Plot, Unbiased Data
SOURCE: Author's tabulation.

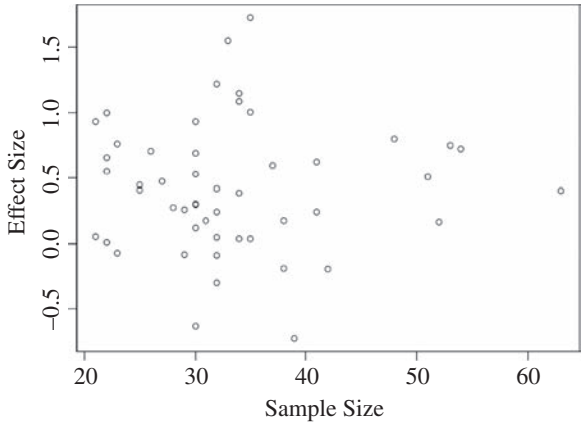

Figure 18.2 Horizontal Funnel Plot, Unbiased Data
SOURCE: Author's tabulation.

distribution of effect sizes covers an expanded range for smaller studies. This gives more plot space to the smaller studies, among which publication bias is more likely to be evident. Jonathan Sterne and Matthias Egger have published comparative plots (2001). Figure 18.3 plots the same effects as figure 18.2, this time against standard error rather than sample size. In figure 18.3, the larger studies appear at the left of the plot, rather than the right, as in figure 18.2, and the range of the plot associated with smaller sample sizes (and larger standard errors) is expanded. Figure 18.4 shows a highly asymmetrical funnel plot, using standard error on the x-axis.

When interpreting funnel plots, the meta-analyst should bear in mind that asymmetry may be due to phenomena other than publication bias. Any external influence associated with both study size and effect size could confound the observed relationship. For example, small studies could be conducted under more carefully controlled experimental conditions than large studies, resulting in differences in effect sizes. In other situations, a higher intensity of the intervention might be possible for the smaller studies, causing their true effect sizes to be larger. Conversely, smaller studies might be carried out under less rigorous conditions; for example, consider a meta-analysis that mixes results from large clinical trials with smaller observational studies that tend to have larger effects. Figure 18.4 actually depicts such a situation; the asymmetry is difficult to miss. Figure 18.5 portrays the same data, but adds information about study type. Neither type of study appears asymmetric, even though the combined distribution is.

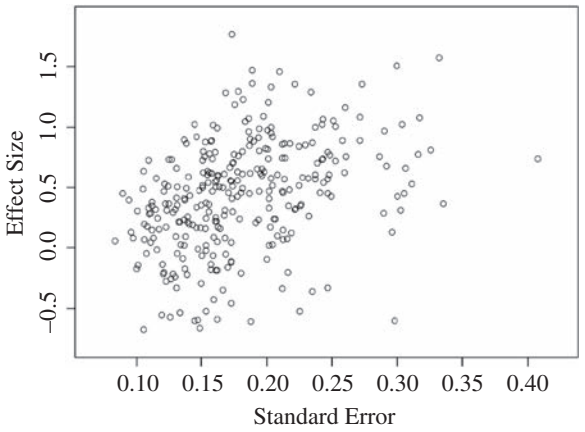

Figure 18.4 Effect Size Against Standard Error, Strong Asymmetry

SOURCE: Author's tabulation.

The utility of the funnel plot has been questioned because of the subjective nature of its interpretation. Norma Terrin, Christopher Schmid, and Joseph Lau find that researchers faced with an assortment of funnel plots cannot correctly identify which plots show bias (2005). Joseph Lau and his colleagues present similar evidence of inconsistent inter-

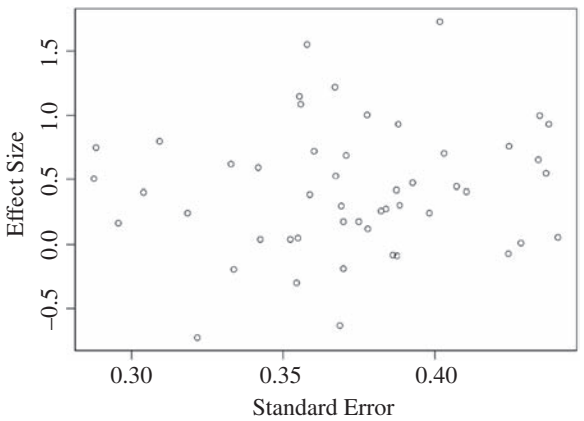

Figure 18.3 Effect Size Against Standard Error, Unbiased Data

SOURCE: Author's tabulation.

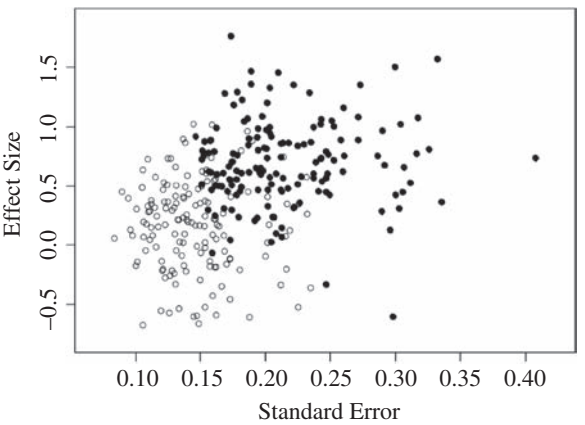

Figure 18.5 Asymmetry Due to Moderator

SOURCE: Author's tabulation.

pretation (2006). Jin-Ling Tang and Joseph Liu (2000), as well as James Hunter and his colleagues (2014), describe problems with interpretation in circumstances where the magnitude of effect sizes is associated with a measure of precision. This is particularly problematic for some outcome measures, such as odds ratios, for which the estimated effect size and its standard error are positively correlated. As a result, in such cases the funnel plot will not be symmetric even in the absence of publication bias. However, Jaime Peters and his colleagues find that the induced asymmetry is small for small and moderate effect sizes (2006).

If variance, standard error, or their inverses are used, it is possible to construct expected 95 percent confidence intervals around the pooled estimate that form guidelines for the expected shape of the funnel under a fixed-effect assumption. Such contour-enhanced funnel plots can also represent other confidence levels. Sterne et al. argue that they can aid in interpretation of the plot (2011). Peters and his colleagues propose that contour-enhanced plots can help the analyst distinguish between asymmetry due to publication bias and asymmetry caused by other factors (2008). Contour guidelines are approximate because they are constructed around the pooled meta-analytic estimate, which may itself be biased. Sometimes the main indication of asymmetry may be differences in the density of plotted points. This frequently occurs well within the bounds of the contour lines. In that case, the contour lines can deflect attention from such changes in density.

18.3.2 Cumulative Meta-Analysis

Cumulative meta-analyses are often used to determine when a meta-analytic effect size appears to stabilize in relation to some variable of interest, such as time of publication. Such a plot consists of a forest plot of meta-analytic results, starting with the oldest study, followed by a meta-analysis of the two oldest studies, and so on, until the final line in the forest plot represents the meta-analysis that includes all of the effect-size estimates. These plots can help identify circumstances in which large effects or results with very small *p*-values are published more quickly than others, a form of publication bias known as time-lag bias.

Recently, the approach has gained a role in the assessment of other forms of publication bias as well. Kepes, Bennett, and McDaniel demonstrate the utility of cumulative meta-analysis (2014). The analyst adds effect sizes one at a time in increasing (or decreasing) order of precision and creates a forest plot of this cumulative meta-analysis. If horizontal drift is present in the plot as studies are added, there is evidence of a relationship between study size and effect size, and therefore the possibility of publication bias. Note, however, that this method, like the funnel plot, cannot distinguish between associations due to publication bias and those due to other causes. A similar approach that bases the meta-analytic estimate on a subset of the most precise studies has also been proposed. Researchers differ on the number or percentage of studies to include for this method. Kepes, Brad Bushman, and Craig Anderson recently used the method using the five most precise studies (2017). In contrast, Tom Stanley, Stephen Jarrell, and Hristos Doucouliagos used the most precise 10 percent of the effect sizes (2010).

18.3.3 Nonparametric Correlation Test

Colin Begg and Madhuchhanda Mazumdar's rank correlation approach makes available a formal test for the presence of funnel plot asymmetry (1994). Along with the other methods presented in this section, it assesses whether a relationship between study size and effect size is present, but does not provide an adjusted estimate of effect size.

The rank correlation test works by estimating a fixed-effect meta-analytic mean, calculating deviations of individual effects from that mean, and standardizing them using the standard error of the deviations from the mean (accounting for both the sampling uncertainty of the individual effects and the standard error of the fixed-effect mean). Subsequently, one calculates a rank correlation between the deviations and their variances using a normalized version of Kendall's tau. Under the null hypothesis of no association between effect and variance, the statistic is tested by reference to the standard normal distribution. As with any hypothesis test, nonsignificance should not be interpreted as a confirmation of the null hypothesis (in this case, that publication bias is absent). Any effect-size scale can be used as long as it is distributed asymptotically normal. The test is available in many statistical packages for meta-analysis.

Begg and Mazumdar's original publication of the method acknowledged that it was underpowered for small meta-analyses (1994). Others echo that finding (Sterne et al. 2000). From comparisons between this and the linear regression test described later in this chapter (Sterne, Gavaghan, and Egger 2000), it appears that the linear regression test is more powerful, though results of

the two tests can sometimes be discrepant. Modifications of the rank correlation test have been proposed to address the power issue, with varying success (for a discussion of these extensions, see Kepes et al. 2012). Because of this concern about power, Jonathan Sterne and his colleagues propose interpreting the results of this test only when a data set includes more than ten effects (2011).

18.4 METHODS FOR ASSESSING THE IMPACT OF PUBLICATION BIAS

Here we examine a variety of approaches for assessing the impact of publication bias on a meta-analysis. Some of these provide an adjusted estimate, a formal test, or both.

For those methods that produce adjusted estimates, meta-analysts likely wish to label the degree of adjustment, or the amount of bias present in their data. Several guidelines for doing so are proposed. Hannah Rothstein, Alexander Sutton, and Michael Borenstein refer to bias as "minimal" when the estimates of effect size are very similar, "modest" when the difference is substantial but the key finding does not change, and "severe" when the key finding is called into question (2005). Kepes, George Banks, and In-Sue Oh refine these definitions, classifying bias as "absent/negligible" when the difference between the unadjusted and adjusted estimates is less than 20 percent, "moderate" when the difference is between 20 percent and 40 percent, and "severe" when the difference is greater than 40 percent (2014). Toward the end of this chapter, we use both guidelines. However, assessing the degree of adjustment is fundamentally subjective, and these guidelines should not be viewed as ironclad.

18.4.1 Fail-Safe *N*

Robert Rosenthal introduced the method that has come to be called the *fail-safe N* (1979). It remains one of the most popular techniques for assessing publication bias today, particularly in the social sciences. The fail-safe *N* addresses the question of how many effect sizes averaging a null value would need to be missing from a meta-analysis to overturn the conclusion that there is a significant effect. Here, significance is defined in terms of inference based on combined *p*-values using the *Z*–score approach (see Stouffer et al. 1949).

Although the fail-safe *N* may be intuitively appealing, it is now generally regarded as valueless. Begg and Berlin argue that the method should be considered nothing more than a crude guide due to a number of shortcomings (1988). Betsy Becker notes that no statistical model underlies the fail-safe *N*, and there is no clear-cut and justifiable criterion for a "large" fail-safe *N* value; she specifically states that the method should be abandoned (2005). One concern is that combining *Z*-scores does not directly account for the sample sizes of the studies. Another is that the choice of zero for the average effect of the unpublished studies is arbitrary; a glance at a typical asymmetric funnel plot will suggest that it is effects below zero that are missing. Hence, in practice, many fewer studies than the number suggested by the fail-safe *N* might be required to overturn the meta-analytic result. Often, then, it has led to unjustified complacency about publication bias. Still another shortcoming is that the method does not adjust or deal with treatment effects—just *p*-values (and thus it provides no indication of the "true" effect size). Satish Iyengar and Joel Greenhouse point out that heterogeneity among the studies is ignored (1988). Robert Orwin notes that the shape of the funnel plot does not influence the method. For these reasons, it is difficult to recommend using the procedure.

It could be argued that, in addition to being hampered by those issues, the fail-safe *N* poses fundamentally the wrong question. Typically, except for meta-analyses that include very few effects, the power to detect even small combined effects is tremendous. Thus it is much more interesting to focus on the question of how many studies would have to be missing for the combined effect to be reduced to a trivial magnitude. Orwin presents an alternative fail-safe *N* that addresses that more interesting question (1983). His method allows the user to specify an average value for missing effects that may or may not be zero, and estimates the number of such effects that would need to be added to the analysis to move the estimated effect below the specified value. This variation is still used with some frequency. However, a glance at the literature that uses the method shows that many who use it accept the default value of zero for the average of the missing effects.

In short, with the possible exception of Orwin's variant, the fail-safe *N* is not a valid method for assessing publication bias.

18.4.2 Methods Based on Observed *p*-Values

The three methods in this category are based not on effect size or study size but on the *p*-values associated with the effects. They have recently gained popularity among meta-analysts, although Blakeley McShane, Ulf Böckenholt,

and Karsten Hansen point out that two of these methods may be viewed as alternative implementations of existing (and more effective) selection models, and all three methods have restrictive assumptions and a series of documented flaws (2016; van Aert, Wicherts, and van Assen 2016; Bruns and Ioannidis 2016; Bishop and Thompson 2016; Ulrich and Miller 2015).

18.4.2.1 *p*-Curve and *p*-Uniform *P*-curve is a method for assessing publication bias, first published by Uri Simonsohn, Lief Nelson, and Joseph Simmons, that has gained popularity (2014). It is based on the notion that, if a given set of studies has evidential value (the average effect size represents a real effect and is not an artifact of bias), the distribution of those one-tailed *p*-values will be right skewed. This means that very small *p*-values (such as $p < .025$) will be more numerous than larger *p*-values if studies have evidential value. If the distribution of significant *p*-values is left skewed and large *p*-values are more numerous than expected, Simonsohn and his colleagues conclude that it is evidence of *p*-hacking—researchers may be striving to obtain *p*-values that fall just below .05.

P-curve uses two tests to assess whether the distribution is right skewed. The first is a binomial test comparing the proportion of observed *p*-values above and below .025; the second is a continuous test that calculates the probability of observing each individual *p*-value under the null hypothesis. The probabilities produced by this second test are then dubbed the studies' "pp" values. These tests for right skew assess what is called the full *p*-curve. To test for "ambitious *p*-hacking," or *p*-hacking to reach below .025 rather than .05, *p*-curve conducts the same tests for right skew on only the observed *p*-values that are below .025, or the "half *p*-curve." If these tests for right skew are not significant, indicating that the studies lack evidential value and no true effect may be present, *p*-curve conducts another pair of binomial and continuous tests to assess whether the studies were underpowered (defined as having power below 33 percent). Simonsohn, Nelson, and Simmons mention that an adjusted effect size can be calculated, and provide supplementary *R* code (R Core Team 2016) on their website (www.p-curve.com) for doing so, but the code is somewhat complicated, and adjusted effect sizes are not discussed here (2014). This method of obtaining an adjusted effect size based on the *p*-curve is an aspect of the *p*-curve approach that functions similarly to *p*-uniform.

P-uniform is also a new method, first published by Marcel van Assen, Robbie van Aert, and Jelte Wicherts (2015). It assumes that the population effect size is fixed—that is, the observed effect sizes are homogeneous with neither systematic nor random heterogeneity. It also assumes that all studies with significant results are equally likely to be published or available for inclusion in a literature review, and that no significant studies are withheld. These are both restrictive assumptions. Most meta-analytic data are heterogeneous to some degree (Gelman 2015; McShane and Böckenholt 2014; McShane and Gal 2015), often despite researchers' best attempts to maintain homogeneity (Klein et al. 2014). In addition, it is extremely unlikely that every significant test has been published and that every nonsignificant test remains unpublished.

Like *p*-curve, *p*-uniform is based on the idea that *p*-values, conditional on a true effect size, are uniformly distributed. If, for example, the hypothesized population effect is 0.50, and the conditional *p*-values for each study are not uniform when calculated under the null hypothesis that the true effect is 0.50, both methods assume that the studies do not reflect the true underlying effect, or that publication bias is present.

P-uniform performs two tests. The first assesses the null hypothesis that the population effect size is zero by transforming the observed significant *p*-values using Ronald Fisher's (1932) method and assessing whether their conditional distribution is uniform. If it is, the test fails to reject the null hypothesis and concludes that there is no evidence of an effect. The second test is a one-tailed test of whether the population effect size equals the effect-size estimate produced by a traditional fixed-effect meta-analysis (van Assen, van Aert, and Wicherts 2015). Again, significant *p*-values are transformed, this time to represent the probability of observing a given effect size conditional on both the fixed-effect average estimate and statistical significance. If this distribution deviates from a uniform distribution, *p*-uniform rejects the null hypothesis and concludes that publication bias may be a threat. Finally, *p*-uniform provides an adjusted effect-size estimate and confidence interval by searching for the population effect size that does meet its qualification—the value where the distribution of conditional *p*-values is uniform. This is similar to the method *p*-curve employs, but the two methods use different algorithms for defining fit to the uniform distribution.

Both *p*-curve and *p*-uniform have several flaws. Robbie van Aert, Jelte Wicherts, and Marcel van Assen point out that these tests perform poorly when meta-analytic data contain *p*-values close to significance levels like .05 (2015). Obviously, many meta-analyses likely con-

tain *p*-values in this range. These methods will always underestimate the true effect when *p*-hacking is present (although, to be fair, many of the other methods in this chapter will as well), and neither method can perform well with heterogeneous data (van Aert, Wicherts, and van Assen 2016). This leads to a recommendation that meta-analysts whose data are heterogeneous should divide their data into homogeneous subgroups prior to estimating the models, but this is often impractical. A traditional random-effects model actually outperforms *p*-curve and *p*-uniform with heterogeneous data, even in the presence of publication bias, the very situation in which both models are designed to work (van Aert, Wicherts, and van Assen 2016). The models assume that all significant studies are published (or otherwise widely available), and are calculated involving only significant *p*-values (McShane, Böckenholt, and Hansen 2016). Finally, Stephan Bruns and John Ioannidis discovered that *p*-curve has difficulty distinguishing between *p*-hacking and the presence of a true effect—the primary purpose for which the method exists (2016). Dorothy Bishop and Paul Thompson confirm their findings (2016), as do Rolf Ulrich and Jeff Miller (2015).

McShane, Böckenholt, and Hansen (2016) note that *p*-curve and *p*-uniform are both a modification of an early selection model presented by Hedges (1984), which forgo maximum likelihood estimation in favor of less efficient alternatives. The benefit of these models is their recent publicity and accessibility, which may increase awareness of publication bias. Beyond that, however, simulations demonstrate that earlier selection models remain more effective under realistic assumptions (McShane, Böckenholt, and Hansen 2016). McShane et al. (2016) also provide mathematical evidence of their ineffectiveness; they argue, based on Jensen's Inequality (Jensen 1906), that *p*-curve and *p*-uniform (as well as the early Hedges 1984 model) will be biased in the presence of heterogeneity.

These models appear to reinvent a wheel first discovered over thirty years ago. The *p*-curve and *p*-uniform methods are certainly superior to some, like the fail-safe *N* or the excess significance test (see following section), and *p*-curve in particular has gained popularity, likely due to its accessibility. Any assessment of publication bias is better than none, and using all methods available is better than using only one. However, meta-analysts should remember that *p*-curve and *p*-uniform are modified versions of simplistic early weight-function models, and should consider using more sophisticated weight-function models as well.

18.4.2.2 Excess Significance Test The excess significance test (or TES), first proposed by John Ioannidis and Thomas Trikalinos (2007), has been the subject of considerable debate. The method is a null hypothesis significance test that takes a given set of studies and asks whether too many are statistically significant or "positive." For example, if a meta-analysis collected three studies and all three were significant with $p < .05$, the excess significance test instructs the meta-analyst to calculate the post hoc power of each study (assuming that the estimated effects are the true effects). The expected number of positive studies is calculated based on the studies' power, and that expected number is compared with the observed number of positive studies using the chi-square statistic. Assuming that each study had 60 percent power, the probability that all three studies would reject the null hypothesis with $p < .05$ works out to the product of the power values, 0.60^3, or 0.22. Guidelines for the TES indicate that a *p*-value less than .10 should be considered significant (Francis 2014). Therefore, given that $0.22 > 0.10$, the meta-analyst will fail to reject the null hypothesis and can conclude that the observed number of positive studies does not exceed the expected.

The simplicity of the excess significance test may initially be appealing. In 2012, Gregory Francis published a series of papers employing the test in various subfields to argue for the presence of publication bias (2012a, b, c, d, e, f, g) and concluded that the results from those subfields should be ignored (Simonsohn 2013). By late 2012, the test of excess significance was attracting more attention, primarily criticism (Balcetis and Dunning 2012; Galak and Meyvis 2012; Piff et al. 2012; Simonsohn 2012), and the test became the focus of a special issue in the *Journal of Mathematical Psychology*.

Criticism of Francis's work and of the excess significance test itself is rooted in a number of important issues. First, as Simonsohn (2013) pointed out, even the presence of publication bias should not result in the rejection of a field of research. The excess significance test does not assess the evidential value (or practical significance) of results. The method is a null hypothesis significance test, implying that it is a form of confirmatory research, despite the fact that it is exploratory at best. Perhaps most notably, Simonsohn writes that the excess significance test actually answers this question: "Has a large enough set of published studies been compiled to reject the obviously false null that all studies, regardless of outcome, would be reported?" (175).

Other criticism of the excess significance test includes the fact that no guidelines are in place for choosing which

tests to examine from a study, especially considering that studies often base multiple tests on the same data, which results in dependencies that affect the outcome of the excess significance test (Johnson 2013). The test does not provide any idea of the magnitude of publication bias or its implications; it also makes assumptions that violate the sequential nature of the publication process (Morey 2013). The excess significance test itself suffers from a substantial lack of power, and simulations demonstrate that it cannot detect even extreme bias without prior knowledge of the true population effect size (Vandekerckhove, Guan, and Styrcula 2013). The test is also likely to perform poorly when effect sizes are heterogeneous (Ioannidis and Trikalinos 2007). Finally, Kepes and Michael McDaniel's simulations revealed that the test for excess significance is not robust to outliers (2015).

In comparison with the many other methods described in this chapter, which have fewer flaws and more redeeming qualities, the excess significance test falls short. Much like the fail-safe *N*, the excess significance test is not useful as a test for publication bias. Researchers should not popularize the test because of its simplicity. Researchers would also be wise to refrain from condemning entire fields of study on the basis of a single severely flawed test.

18.4.3 Trim and Fill

The nonparametric trim and fill method was developed as a simpler alternative to parametric selection models (Duval and Tweedie 2000a, 2000b). It is one of the most popular methods for adjusting for publication bias (Borenstein 2005; Moreno et al. 2009). Also, because the method formalizes the use of a funnel plot, the ideas underlying its statistical calculations can be communicated visually, increasing its accessibility.

The method is based on rectifying funnel plot asymmetry. It assumes that the studies on either the far left-hand or right-hand side of the funnel are suppressed and, therefore, it is a one-sided procedure. First, the method uses an iterative process to determine how many studies would have to be removed, or "trimmed," from one side of the funnel for the remaining effect sizes to be symmetric. It trims the asymmetric effect sizes, then uses one of three estimators to generate, or "fill in," new effects that are mirror images of the remaining ones. The adjusted pooled effect size is then calculated based on this augmented symmetrical data set, which can also be used to calculate an adjusted variance component (Jennions and Moller 2002). Although Eric Weinhandl and Sue Duval (2012) are currently working on allowing trim and fill to include a linear model for the mean effect, it is not yet developed for more than one linear predictor.

Either a fixed- or random-effects meta-analytic model can be used for the iterative trimming and filling parts of this method, and once effect sizes are filled in, either model can be used to obtain adjusted estimates. In this way, the method can accommodate random (or between-studies) heterogeneity, and it can produce an adjusted variance component estimate. The choice between fixed- and random-effects is important. Sue Duval and Richard Tweedie originally advised using a random-effects model for both steps, a process that they referred to as "random-random," because doing so would yield more conservative confidence intervals (2000a, 2000b). However, if publication bias is present and smaller studies are clustered together, the random-effects model (which allows more weight to smaller studies) may be biased; as a result, Alexander Sutton advocates a "fixed-fixed" process (2005). The "fixed-random" process is a compromise, using a fixed-effect model for trimming and a random-effects model to estimate the adjusted effect, although the adjusted estimate may be overly conservative (Peters et al. 2007). Duval recommends the more conservative approaches, although she emphasizes that all three should be estimated and compared (2005).

During the iterative procedure, three possible estimators of the number of missing studies may be employed. Two of these estimators, known as R_0 and L_0, are recommended; the third, Q_0, is merely a linear transformation of one of the others (Duval and Tweedie 2000a, 2000b). If choosing between the two (R_0 and L_0), L_0 sometimes performs better (Jennions and Moller 2002). Others have found that R_0 performs better (Peters et al. 2007). Duval, however, recommends estimating both and comparing the results, especially because the performance of the estimators can depend on the number of observed versus missing effects (Duval 2005; Duval and Tweedie 2000b). Given space constraints, the details of these estimators are not provided here, but thorough examples of their calculation are available elsewhere (Duval and Tweedie 1998; Duval 2005).

Duval and Tweedie initially evaluated this method through simulation, under homogeneous conditions and with a data suppression mechanism matching the model—that is, where the most extreme effect sizes were suppressed (2000a, 2000b). Under those conditions, it performed well. However, other simulations suggest it may perform poorly in the presence of between-study heterogeneity in the absence of any evidence of publication bias (Terrin et al. 2003). Peters and his colleagues conducted

simulations further evaluating the performance of trim and fill, finding that trim and fill underestimates the true effect size in the absence of publication bias (2007). They note that trim and fill is not ideal, in part because it can impute unrealistic effect sizes, although it *can* outperform the unadjusted random-effects model in the presence of publication bias, but should be considered a sensitivity analysis, as originally intended (Peters et al. 2007; Duval and Tweedie 2000a, 2000b). Guido Schwarzer, James Carpenter, and Gerta Rücker (2010) compare trim and fill to the Copas selection model (Copas and Shi 2000) and confirm that trim and fill is more conservative due to inflated standard errors.

All publication bias methods should be regarded as sensitivity analyses; therefore, saying the same of trim and fill is not a slight. Although trim and fill has its share of problems, it is both popular and accessible, and meta-analysts will likely benefit from including it in their arsenal of assessment methods.

18.4.4 Linear Regression Adjustment

In 1997, Matthias Egger and his colleagues described a parametric test for funnel plot asymmetry based on linear regression. The test regresses the standard normal deviate, or the effect sizes divided by their standard errors, on precision (defined as the inverse of the standard error). This regression fits a line to Rex Galbraith's (1994) radial plot, in which the regression line is not constrained to go through the origin. Effect sizes from small studies will have a standard normal deviate that is close to zero regardless of their magnitude, and large studies will produce large standard normal deviates. Therefore, in the absence of publication bias, the regression line will run through the origin. If bias is present, small studies may differ systematically from larger studies, and the line will no longer run through the origin (Egger et al. 1997).

The regression intercept measures the magnitude and direction of asymmetry, and a significant *t*-test on the intercept indicates that asymmetry (and, by extension, publication bias) may be present. A negative intercept indicates that smaller studies have larger effects; a positive intercept indicates that they have smaller effects than expected. The regression model, as proposed above, is equivalent to a weighted meta-regression model, and the regression line can be displayed on a funnel plot for clarity of interpretation.

The model just presented is Egger's linear regression in its original form. Several researchers have proposed extensions or modifications of this model (Macaskill, Walter, and Irwig 2001; Sterne and Egger 2005; Harbord, Egger, and Sterne 2006; Peters et al. 2006; Rücker, Schwarzer, and Carpenter 2008; Deeks, Macaskill, and Irwig 2005). Petra Macaskill, Stephen Walter, and Lesley Irwig propose a variation in which the effect sizes, rather than their standard normal deviates, are regressed on their study size and weighted by their inverse pooled variance (2001). This model and the next three variations reverse the role of the intercept and slope; the slope is expected to be zero in the absence of publication bias. Peters and his colleagues prefer regressing effect sizes on the inverse of their sample size (2006). Sterne and Egger advocate the regression of effect sizes on their standard errors, weighted by their inverse variance (2005). Gerta Rücker, Guido Schwarzer, and James Carpenter describe a variation of this for binary outcome data that has been arcsine-transformed (2008). Jonathan Deeks, Petra Macaskill, and Les Irwig propose a regression of the effect size involving the effective sample size (ESS), defined as $4n_1n_2/(n_1+n_2)$ (2005). The effect size is regressed on the reciprocal of the square root of ESS and weighted by ESS. Finally, Roger Harbord, Egger, Jonathan Sterne recommend regressing the efficient scores (defined as the first derivative of the log-likelihood) against the score variance (Fisher information), for which the intercept is a measure of bias (2006). For binary outcomes, because of the correlation between odds ratios and their standard errors, the original Egger's regression has an inflated type I error rate, and variations are preferable (Moreno et al. 2009).

Clearly, there are several variations of the original Egger's linear regression (Egger et al. 1997). Although these models differ in terms of outcome measure and predictor, and although the role of the intercept and slope occasionally change, the models are not discussed individually for the sake of brevity. A mention of Egger's linear regression or Egger's test here refers to the entire class of models, unless otherwise specified.

If a meta-analytic data set contains systematic heterogeneity due to covariates, these must be considered when using any funnel plot–based assessment. With discrete covariates, separate assessments can be made for each group, although this approach may result in a considerable reduction of power. The meta-analyst can also extend the regression model to include study-level covariates, therefore estimating a mixed-effects weighted regression. In this way, Egger's regression is capable of accommodating some forms of heterogeneity, but it does not incor-

porate random (or between-studies) heterogeneity and cannot estimate a variance component. Egger's regression does, however, produce an adjusted estimate of the average effect size (the slope), although its estimate is biased because its predictor variable is subject to sampling error and therefore violates the assumptions of linear regression (Macaskill, Walter, and Irwig 2001). Although we are not aware of any research on the subject, it is theoretically possible to fit a measurement error model to overcome this bias; exploring such an idea could be promising.

Peters and his colleagues note that Egger's regression is widely used in the medical literature (2006). In the social science literature, the fail-safe *N* is still the most common procedure, despite its deep-seated flaws, but Egger's regression is gaining popularity (Ferguson and Brannick 2012). Egger's regression does suffer from low power and poor performance when the number of studies is small, especially when there are fewer than twenty, or when the treatment effect is large (Moreno et al. 2009; Sterne, Egger, and Smith 2001; Macaskill, Walter, and Irwig 2001). Egger's regression is most powerful with a large number of effect sizes that range widely in terms of study size (Macaskill, Walter, and Irwig 2001). Its problems with power, however, are not unique among publication bias assessment methods, and it is still a useful tool.

18.4.5 PET-PEESE

In 2014, Tom Stanley and Hristos Doucouliagos proposed PET-PEESE, and since then the method has appeared occasionally in the meta-analytic literature (Carter and McCullough 2014; Carter et al. 2015). PET-PEESE is actually an extended modification of Egger's regression (Egger et al. 1997). Stanley and Doucouliagos instruct the meta-analyst first to estimate a regression of effect size on standard error, weighted by the inverse variance (2014). This is the exact model that Sterne and Egger propose (2005). Stanley and Doucouliagos point out that, though the slope of this model is a measure of bias, the intercept is also informative: it represents an estimate of the effect size when the standard error is zero (2014). Therefore, they argue that the intercept is an estimate of a perfectly precise study, or an effect size uninfluenced by publication bias (Stanley 2005). They call this first regression the Precision-Effect Test (or PET). Thus far, PET is a restatement of the fact that, for certain variations of Egger's regression, the intercept is an effect-size estimate adjusted for publication bias. A *t*-test on the intercept, using the null hypothesis that the intercept is zero, indicates whether a true effect is present.

Assuming that a test on the intercept is significant, or that a nonzero effect exists, results in a second problem. The issue with using this adjusted estimate is that, as mentioned previously, the estimate is biased. To avoid this problem, Stanley and Doucouliagos propose a second conditional test (2014). If the intercept from PET is significant, they advise meta-analysts to conduct another regression, this time with effect size predicted by sampling variance rather than standard error. This regression is called the Precision-Effect Estimate with Standard Error (PEESE). PEESE produces an intercept that is *still* biased, but simulations demonstrate that it is less biased than the intercept from PET (Stanley and Doucouliagos 2014). Therefore, they advise that, if the PET test is significant, meta-analysts should estimate PEESE and accept its intercept as an adjusted effect-size estimate. If PET is nonsignificant, there is not enough evidence that the true effect size differs from zero.

A problem with this approach is that bias in the intercept estimate does not vanish when using variance as a predictor rather than its square root. PET-PEESE is not a new technique, although it is described as such; it is a combination of existing variations of Egger's regression. Macaskill, Walter, and Irwig have explained the source of bias in the intercept (2001). The intercept is a biased estimate not because of the choice of predictor, but because of a violation of one of the assumptions for linear regression. Both predictors, variance and standard error, are not fixed; they are random, and are estimated from the observed data. Therefore, using either PET or PEESE, measurement error is inherently present in the independent variable, and the estimate of the intercept will be biased downward—the exact result that Stanley and Doucouliagos (2014) describe. Stanley and Doucouliagos also argue that PET-PEESE outperforms random-effects meta-regression in the presence of publication bias, although it still performs worse overall in the presence of high levels of heterogeneity.

Because it is a combination of two Egger's regression variations, PET-PEESE possesses the same flaws. It has low power and cannot incorporate random or between-studies heterogeneity. When heterogeneity is present, it performs poorly (Stanley and Doucouliagos 2014) and the coverage rate of its confidence intervals is persistently low (Moreno et al. 2009); in the presence of severe bias or when the data are homogeneous, its confidence intervals are too wide (Moreno et al. 2009). Finally, its

effect-size estimate is biased. When we consider the many other methods presented in this chapter that possess more redeeming features, including other variations of Egger's regression, PET-PEESE appears to be a flawed method.

18.4.6 Selection Modeling

Selection models adjust meta-analytic data sets by specifying a model that describes the mechanism by which effect sizes may be suppressed. This model is combined with an effect-size model that describes the distribution of effect sizes in the absence of publication bias.

If the selection model were known, selection methods would be straightforward, but the precise nature of the suppression will almost always be unknown. Instead, selection approaches attempt to estimate the selection model, along with adjusted estimates of the meta-analytic parameters. Although complex to implement, they are recommended over other methods, which can produce misleading results when effect sizes are heterogeneous (Terrin et al. 2003). Selection methods may perform poorly when the number of observed effects is small; an alternative involves specifying selection models of varying severity and estimating the meta-analytic parameters contingent on each hypothetical selection pattern. Jack Vevea and Carol Woods present such an approach (2005; see also table 18.1). These methods, which do not estimate parameter values from the data, are sensitivity analyses by nature, although, of course, all bias assessments are.

Two classes of selection models have been developed: those that model suppression as a function of an effect size's *p*-value, and those that model suppression as a function of a study's effect size and standard error simultaneously. Both are implemented using weighted distributions that represent the likelihood of observing a given effect estimate if it occurs. These methods have gained popularity in the publication bias literature. Descriptions of the more complex selection models are presented here with limited statistical detail. Hedges and Vevea published a comprehensive review of selection models available by the early 2000s that provides a more statistically rigorous account of some of the approaches described here (2005).

Although they do have flaws, namely, their complexity and sample size requirements, both classes of selection model tend to perform well in simulations and allow meta-analysts to evaluate data under a range of selection patterns. Therefore, they are valuable tools in an arsenal of bias assessments.

18.4.6.1 Suppression as a Function of *p*-Value Only Selection models that depend solely on effect sizes'

Table 18.1 Sample Selection Patterns for the Vevea and Woods Method

	Probability of Observing Effect			
p Interval	Moderate One-Tailed Selection	Severe One-Tailed Selection	Moderate Two-Tailed Selection	Severe Two-Tailed Selection
.000–.005	1.00	1.00	1.00	1.00
.005–.010	.99	.99	.99	.99
.010–.050	.95	.90	.95	.90
.050–.100	.90	.75	.90	.75
.100–.250	.80	.60	.80	.60
.250–.350	.75	.50	.75	.50
.350–.500	.65	.40	.60	.25
.500–.650	.60	.35	.60	.25
.650–.750	.55	.30	.75	.50
.750–.900	.50	.25	.80	.60
.900–.950	.50	.10	.90	.75
.950–.990	.50	.10	.95	.90
.990–.995	.50	.10	.99	.99
.995–1.000	.50	.10	1.00	1.00

SOURCE: Author's tabulation.

p-values propose or estimate the likelihood of surviving selection as a function of those *p*-values. Hedges (1984) as well as David Lane and William Dunlap (1978) propose simple selection models that assume all statistically significant effect sizes are observed (for example, $p < .05$ two-tailed, or $p > .975$ or $p < .025$ one-tailed) and all others are suppressed. With this approach, any effect size with a *p*-value < .05 has a probability of one (certainty) of being observed and a probability of zero otherwise. Iyengar and Greenhouse propose somewhat more sophisticated models, assuming that the likelihood of publication is a decreasing function of the *p*-value for studies that are not statistically significant (1988). In the years following, various authors have proposed more sophisticated models.

18.4.6.1.1 Dear and Begg. Keith Dear and Colin Begg introduce a semi-parametric method for assessing publication bias that uses a nonparametric weight function on the two-tailed *p*-value scale (1992). The method, they note, can easily be adapted for one-tailed *p*-values. The weight function is a step function with discontinuities at the alternate individual observed values of *p*. In other words, the Dear and Begg model takes the observed *p*-values of all effect sizes in the data set and orders them. It then includes every other *p*-value as discontinuities in the weight function. For example, if the first four *p*-values of a data set were .001, .01, .03, and .04, the first discontinuity in the weight function would be set at $p = .01$, and weights would be estimated for *p*-values below .01 and *p*-values between .01 and .04. This means there are $k/2$ weight parameters for a meta-analytic data set of size k. The model estimates a weight for each interval that represents the relative probability of surviving the selection process. To identify the model, the weights are constrained to fall between zero and one. However, they are not directly interpretable as probabilities because we lack information about the base rate of publication. No effect has 100 percent probability of publication.

Although the model can provide both an adjusted estimate of the average effect size and a statistical test, Dear and Begg focus on using plots of the weights against *p*-values as a tool for visual assessment (1992). Spikes in the plot indicate that the weight for *p*-values in that particular range is large, meaning that studies with *p*-values in that range are more likely to be published and therefore observed. Valleys or dips in the plot indicate the opposite; studies with *p*-values in those ranges are less likely to be observed.

With this approach, weights for larger (less significant) *p*-values sometimes exceed the weights for the most significant values, making visual assessment of bias difficult. If slight fluctuations in the weights are numerous, identifying the overall pattern may be complicated. Kaspar Rufibach presents an extension of the model that addresses this problem (2011). His approach is identical to Dear and Begg's (1992), except that Rufibach has imposed a constraint, forcing the weights to be a monotone nonincreasing function of *p*-values. Rufibach notes that the constraint improves the performance of estimates, yields more insight into the selection process, and leads to a more realistic weight function. The constraint also makes it easier for meta-analysts to interpret the function from plots.

The Rufibach model provides a useful plot of the weight function, and can be informative (2011). However, as the number of effect sizes in the meta-analysis increases, both the Dear and Begg (1992) and Rufibach models become difficult, if not impossible, to estimate. This problem occurs because, rather than allowing meta-analysts to restrict the number of *p*-value discontinuities, the models determine the number of discontinuities as $k/2$. For a meta-analysis with $k = 20$, this is manageable; for a meta-analysis with $k = 200$, estimating more than one hundred parameters (including a mean and variance component) may be impossible. Furthermore, the difference in assumptions between the two models can point to radically different conclusions (see the example later in this chapter).

In keeping with the importance of triangulation, we encourage the use of these models as part of a toolbox of assessments, but warn meta-analysts that the models may be inestimable under some circumstances.

18.4.6.1.2 Hedges. Hedges proposed a similar model that assumes a step function over *p*-values (1992). His model differs from Dear and Begg's (1992) approach because the analyst must specify steps at perceived milestones in statistical significance. These milestones are based on the perception that a *p*-value of .049 is considerably different from one of .051, that .011 is different from .009, and so on. (Often $p = .50$ is a particularly relevant cut point because it reflects the point at which many effect-size metrics change from positive to negative.) Weights representing the relative likelihood of survival for the intervals are estimated in the context of a random-effects model, and all parameters (weights, the mean effect, and the variance component) are estimated simultaneously by the method of maximum likelihood. The model uses only two-tailed *p*-values, which cannot represent the direction of the effect. (Software for estimating

the two-tailed model is no longer available.) Vevea, Nancy Clements, and Hedges modified the model to use one-tailed *p*-values (1993). Models based on one-tailed *p*-values can still represent a two-tailed selection pattern. A pair of one-tailed *p*-values can define a two-tailed value by employing, for example, .025 and .975 in place of .05. This provides freedom from the constraint that selection must operate identically for positive and negative effects, which is an unlikely phenomenon. One- and two-tailed selection patterns fundamentally reflect the assumption of asymmetry (one-tailed) versus symmetry (two-tailed) of the weight-function model.

The method employs weighted distribution theory: the usual random-effects likelihood is multiplied by the weight for the *p*-value interval of each study, then renormalized. The software first estimates a conventional fixed- or random-effects model. Then the meta-analytic model is reestimated using the weighted likelihood. In addition to the mean and variance component, the model estimates all but one of the weights associated with the *p*-value intervals. To identify the model, the weight for the most significant range of *p*-values is fixed at 1.0. Other weights are interpreted relative to that first weight, and can actually exceed 1.0. Hence, they are not directly interpretable as probabilities. This weighted model provides mean and variance component estimates adjusted for publication bias, as well as estimated weights reflecting the relative likelihood of observing effect sizes in each interval. In addition, a likelihood-ratio test for publication bias compares the conventional model to the adjusted model.

18.4.6.1.3 Vevea and Hedges. Vevea and Hedges (1995) later added the possibility of including study-level covariates to the Vevea, Clements, and Hedges (1993) model. This can remove confounding effects in the distribution of effect sizes if asymmetry in the funnel plot is partly due to the presence of covariates. This weight function model can accommodate a full linear model for the mean effect, including dichotomous and continuous predictors, and can provide estimates of those predictors adjusted for publication bias. A particular advantage is that in some cases, certain classes of effects may remain virtually unaffected by the presence of the selection model, while others may be strongly affected.

The model does have some flaws—in particular, it does not perform as well with smaller meta-analyses, and it cannot estimate weights for ranges of *p*-values in which no observed effect sizes fall. It requires no precise number of effect sizes. Instead, meta-analysts must ensure that there are at least some observed effects in each range of *p*-values they specify, and must keep in mind that weights for intervals with few observed effects will be poorly estimated. Additionally, the model is a selection model, and selection models are often dismissed for their complexity. The model does require users to think about the selection process and to specify some relevant *p*-value breakpoints, but this is not necessarily a flaw. It is unlikely that a phenomenon as complex and multifaceted as publication bias could be adequately handled without some careful consideration.

Despite its flaws, the model has a number of positive features. First and most important, it is capable of handling both random and systematic heterogeneity. Many other assessments cannot accommodate linear models; meta-analysts can still use them on homogeneous subsets, but this may not be practical and, for continuous moderators, may be impossible. Additionally, in terms of performance, a simulation shows that variations of the Hedges model outperform both *p*-curve and *p*-uniform (1992). Such variations have narrower confidence intervals and are robust both to heterogeneity and to differing selection strengths (McShane, Böckenholt, and Hansen 2016). An earlier study by Hedges and Vevea also finds that such models are robust to violations of assumptions about the distribution of random effects—that is, to nonnormal distributions (1996).

Thus far, the Vevea and Hedges method has not seen much use, likely because no user-friendly software has been available (1995). However, Coburn and Vevea have released an *R* package to *CRAN* (the Comprehensive R Archive Network) titled *weightr* (2016a). The program is capable of estimating both the Vevea and Hedges model and the modified Vevea and Woods version described in the following section (2005). The same software is also available through a web-based point-and-click Shiny application (Coburn and Vevea 2016b).

Simulation studies of similar selection models indicate that the Vevea and Hedges model bears promise and will likely perform well under realistic circumstances, whether in the presence of systematic heterogeneity, random heterogeneity, or both (Vevea and Hedges 1995; McShane, Böckenholt, and Hansen 2016). This class of models also appears robust to different patterns of selection, which is a crucial trait given that researchers can never know the true underlying selection pattern. Meta-analysts would be remiss to overlook this model in favor of its simpler counterparts. The release of software will allow the Vevea and Hedges model to see increased use.

18.4.6.1.4 Vevea and Woods. In 2005, Vevea and Woods published a paper presenting a modification of

the Vevea and Hedges (1995) model. Some meta-analysts were disappointed because their data sets were too small to allow estimation of the Vevea and Hedges model. (With small data sets, it is often not possible to estimate weights for more than one or two *p*-value intervals.) There was interest in a method that could allow the user to specify not only *p*-value cut points, but also weights for the *p*-value intervals—a sensitivity analysis tool that would enable the user to explore how the conditional means of a data set might vary under different bias patterns.

The Vevea and Woods model provides this adaptation (2005). With it, there is no need to ensure that the data set is large enough, or even that there are observed effect sizes in every *p*-value interval. The meta-analyst merely specifies the *p*-value cut points of interest and a set of hypothetical weights for the corresponding *p*-value intervals, and the model produces estimates of the adjusted conditional means and variance component under the specified conditions. Because the model is not actually estimating parameter values, the standard errors and confidence intervals are no longer meaningful, nor is the likelihood-ratio test comparing the unadjusted and adjusted models. This does not reduce the impact of the model, however. It is still a valid sensitivity analysis tool that can provide the curious meta-analyst information about how specified selection bias patterns could affect their data, or about how robust their data are to selection bias. When moderators are included in the analysis, the results may show that a subset of effects identified by the linear model are virtually unaffected by any trial bias pattern.

Because the pattern of bias is imposed by the researcher rather than estimated from the data, sometimes the mean and variance component estimates can be adjusted relative to an extreme or unrealistic selection pattern. Kepes and McDaniel record an example of this (2015). To understand why, imagine a case in which the meta-analyst specifies weights of zero for all *p*-value cut points (indicating that no effect sizes can occur). In such a scenario, the estimates will obviously be nonsensical, if the model even converges; estimates may blow up or reduce to zero if extreme selection patterns are imposed. Researchers must remember that they are merely observing the reaction of the estimates to varying scenarios; they should assess the change in estimates *across* scenarios to determine whether their data set is robust to different selection patterns.

The Vevea and Woods (2005) model is a convenient workaround for meta-analysts who wish to implement the Vevea and Hedges (1995) model, but who do not have enough effect sizes to estimate weights. In this way, it is a useful addition to the literature, and helps make selection modeling a feasible option for more researchers.

18.4.6.2 Suppression as a Function of Effect Size and Its Standard Error Another class of models addresses publication bias by assuming a relationship among effect sizes, their standard errors, and the likelihood of their surviving the selection process.

18.4.6.2.1 Copas and Shi. John Copas and Hu Li initially proposed a selection model that functions as a sensitivity analysis in 1997; in subsequent years, Copas and Shi (2000, 2001) published several variations of the model. The method is frequently cited in discussions of selection models, as demonstrated by the fact that the original paper has received more than 340 citations, but it has not seen much practical use. Recently, Schwarzer, Carpenter, and Rücker (2016) created a software package called *metasens* using *R* (R Core Team 2016). The package implements the model and provides guidelines for its interpretation. As a result, the approach is gaining in popularity as more meta-analysts use it (Preston, Ashby, and Smyth 2004; Bennett et al. 2004). Some researchers even advocate a Bayesian implementation, which may avoid the issue of specifying values for the *a* and *b* parameters (Mavridis et al. 2012).

The Copas and Shi selection method (2001) combines two models: a population model that is equivalent to the usual random-effects meta-analytic model, and a model in which the probability of a study being published is a linear function of its reported standard error. There are two parameters in this linear model, the intercept (*a*, or the overall proportion of studies published when the standard error of those studies is zero) and the slope (*b*, or the relationship between standard error and publication). This linear model can be rewritten as a propensity model, where a study is selected for publication if and only if its propensity is greater than zero (Copas and Shi 2001; Copas and Li 1997). A correlation parameter links the observed effect sizes and their estimated propensities. A correlation of zero indicates the complete absence of publication bias, or a case in which effect sizes are published regardless of their standard error, while a positive correlation indicates the presence of bias (Copas and Shi 2001). Therefore, the conditional random-effects model represents the observed effect sizes, given that their propensity score is greater than zero (Copas and Shi 2000).

The selection model involves a total of five parameters—the mean effect size, the variance component, the correlation between effect sizes and propensities, and the slope and intercept (*a* and *b*) for the propensity model. No software to incorporate moderator variables is currently

available, but Copas and Shi indicate that the random-effects population model could easily be replaced with a mixed-effects model (2001). The problem with estimation, however, lies with the *a* and *b* parameters; they are not identified, because not enough information is available (the meta-analyst never knows how many studies remain unpublished). Copas and Shi demonstrate this by proving that the likelihood function for *a* and *b* is almost a plateau, so that using maximum likelihood estimation is near impossible. To solve this problem, they propose entertaining a series of specified values for *a* and *b*, then assessing the impact of those values on the average effect size and variance component. In that way, although all publication bias models should be treated as a sensitivity analysis, their model *must* be; it is a sensitivity analysis by nature, like the Vevea and Woods (2005) approach.

The Copas and Shi (2001) selection model features an algorithm that chooses a range of values for *a* and *b* and then uses maximum likelihood estimation to calculate the average effect size for each pair of values. (On occasions when the model chooses a range that produces uninterpretable results, the user can manually specify a range.) These results demonstrate how the average effect size changes as the likelihood of small studies being published changes. The relationship is easier to observe graphically, and four types of plots aid in its interpretation. The first is a standard contour-enhanced funnel plot. The second is a contour plot of the adjusted effect size against the values of *a* (on the *x*-axis) and *b* (on the *y*-axis), with the values representing no publication bias in the top right (Carpenter et al. 2009). If the contour lines are spread far apart, the adjusted effect size does not change much as the values of *a* and *b* change, and appears robust. The third plot explores this further; it plots the probability of publishing the study with the smallest sample size (on the *x*-axis) against the corresponding adjusted effect size (on the *y*-axis). If this relationship has a slope of zero, the effect size appears to be robust; otherwise, it may be affected by bias (Carpenter et al. 2009). Finally, the fourth plot involves the *p*-values for a likelihood-ratio test that assesses whether selection bias remains. The *p*-values (on the *y*-axis) are plotted against the probability of publishing the smallest-*N* study (again on the x-axis). The point where the plotted curve crosses the horizontal dashed line indicates that the corresponding probability on the x-axis is the most likely probability according to the model (Carpenter et al. 2009).

The Copas and Shi selection model has several positive features (2001). It can accommodate not only random (between-study or unobserved) heterogeneity captured by the variance component but also systematic (or observed) heterogeneity through incorporation of moderators. It can produce adjusted estimates of all the parameters of interest for each specified level of publication bias. In addition, the emphasis on sensitivity analysis encourages meta-analysts to view the model results as flexible, rather than accepting them as truth.

There is a dearth of information from simulations assessing the model's performance. The vast majority of manuscripts exploring the Copas and Shi (2001) model do so empirically, comparing it with other publication bias assessments using a limited set of observed data sets (Carpenter et al. 2009; Mavridis et al. 2012; Schwarzer, Carpenter, and Rücker 2010). Rücker, Carpenter, and Schwarzer (2011) recently presented the results of a small-scale simulation evaluating the Copas and Shi model, but noted that doing so was time-consuming and difficult, and that their simulation neglected to assess extreme heterogeneity and small sample sizes. A thorough simulation of the Copas and Shi model, perhaps along with competing selection models, would be informative.

The Copas and Shi model is a valuable addition to the body of selection models for publication bias, and its increasing popularity is promising (2001). Like the Vevea and Woods (2005) model, that of Copas and Shi does not estimate a selection pattern from the data; it imposes a range of possible patterns and observes the results. Therefore, it may also work well with smaller meta-analyses. Software to implement the model is available, which may encourage its use in the future.

18.4.6.2.2 Rücker. Rücker, Carpenter, and Schwarzer first published the Rücker limit meta-analysis method in 2011. The underlying model is an extended random-effects model that takes account of a possible relationship between effect size and sample size by allowing effect size to depend on standard error. Part of the model is based on earlier simulation work by Rücker, Schwarzer, and Carpenter (2008), which involved artificially inflating the sample size of effect sizes by a given factor of *M*. The model is also based on the original Egger's linear regression (Egger et al. 1997).

The concept of limit meta-analysis begins with the usual random-effects model, with an added parameter *alpha* that represents a small-study effect by allowing effect size to depend on standard error (Rücker, Carpenter, and Schwarzer 2011). The method then considers a situation where the sample size of all observed effect sizes is inflated by a factor of *M*. As *M* approaches positive infin-

ity, so does sample size; the effect sizes become infinitely precise, and variation due to sample size disappears—between-studies heterogeneity is all that remains. Estimating an original Egger's regression (Egger et al. 1997) on the observed effect sizes yields an intercept and slope. The intercept corresponds to the alpha parameter, and the slope is an estimate of the average effect size with a standard error of zero, or the infinitely precise effect. The limit meta-analysis method creates a new data set by transforming the original effect sizes so that they are centered on the slope from the Egger's regression. The slope of an Egger's regression calculated on this centered data set is the estimate of the average effect size, adjusted for small-study effects. A test on the intercept, or alpha, assesses the presence of a small-study effect. Finally, a test for heterogeneity on the centered data addresses the question of whether residual heterogeneity is present after adjustment.

Because it is based on Egger's regression, limit meta-analysis may presumably also incorporate a linear model, although no manuscripts demonstrate this feature. Limit meta-analysis does not produce an adjusted variance component, but does test for the presence of heterogeneity.

Rücker, Carpenter, and Schwarzer (2011) explored the performance of limit meta-analysis, generating and suppressing the data according to the Copas and Shi model, and find that its adjusted estimate was less biased than those from trim and fill and the Copas and Shi model (2001). Of course, the Copas and Shi approach is purely a sensitivity analysis, so its bias depends on the particular parameter settings used in the simulation. Limit meta-analysis was the most conservative of the three. As the size of the small-study effect increases, so does the performance of the limit meta-analysis method in comparison to the usual random-effects model (Rücker, Carpenter, and Schwarzer 2011).

Berlin and Robert Golub briefly explored the performance of the limit meta-analysis method, but further research into its performance would be beneficial (2014). The question of bias also remains. This method relies on the adjusted estimate from Egger's regression, so its overly conservative nature may be due to the same violated assumption that impacts PET-PEESE.

18.4.6.3 Bayesian Approaches After early interest in developing Bayesian approaches to address publication bias, there was a lengthy gap in new developments. Recently, however, there has been a resurgence of activity in Bayesian methods.

M. J. Bayarri and Morris DeGroot (1987) introduced a Bayesian method similar to Hedges's (1984) early approach in that it restricts attention to statistically significant outcomes. Geof Givens, David Smith, and Richard Tweedie developed a method similar to Hedges's (1992) early version of the step-function model (1997). Nancy Silliman (1997) presented, in a random-effects context, Bayesian models that estimate weight functions similar to those that both Hedges and Olkin (1985) as well as Iyengar and Greenhouse (1988) describe. Silliman also developed more complex weight-function models, including one that estimates weights as a step function of *p*-values with unknown cut points between intervals. Daniel Larose and Dipak Dey (1998) offered a similar method, emulating Iyengar and Greenhouse with a random-effects model.

More recently, Dimitris Mavridis and his colleagues developed a Bayesian implementation of the Copas approach (2012). Maime Guan and Joachim Vandekerckhove (2016) describe a method that considers four possible models—no selection, extreme selection with only statistically significant effects, nonsignificant results published with unknown but constant probability, and the Givens and colleagues model (1997). Their approach is to use Bayesian model averaging over the four competing models. Although it could be argued that these four models are not necessarily the best choices, the idea of Bayesian model averaging in this context is intriguing.

Other papers proposing new Bayesian approaches to addressing publication bias are currently under review. Thus, the Bayesian toolbox is likely to be expanded in the near future.

All of these Bayesian approaches have a common shortcoming: to our knowledge, accessible estimation software is not available. Hence, the meta-analyst who is not well versed in Bayesian estimation would find it difficult to employ these methods.

18.5 METHODS TO ADDRESS SPECIFIC DISSEMINATION BIASES

Incomplete data reporting may occur at various levels below the suppression of whole studies. Analyses of specific outcomes or subgroups, for example, may have been conducted but not written up and therefore are not available for meta-analysis. Similarly, all the details of an analysis required for meta-analysis may not be reported. For example, the standard error of an effect size or a study-level covariate, which is required for meta-regression, may not have been published.

This latter category of missing data may be entirely innocent, simply due to a lack of journal space or awareness about the importance of reporting such information.

Such missingness may not be related to outcome, and data can be assumed missing (completely) at random. If that is the case, standard methods for dealing with missing data can be applied to the meta-analytic data set, though this is rarely done in practice (Little and Rubin 1987; Pigott 2001; Sutton and Pigott 2004). More ad hoc methods can also be applied as necessary (Song et al. 1993; Abrams, Gillies, and Lambert 2005). However, meta-analyses in which data are missing completely at random are most likely quite rare.

If the data are missing for less innocent reasons, then it is probably safest to assume data are not missing at random. That is the typical assumption made when addressing publication bias, though it is not often framed as a missing data problem. Outcomes and subgroups may be suppressed under mechanisms similar to those acting on whole studies, so missingness may manifest itself in a similar way, and therefore the methods covered may be appropriate to address it. There may be advantages, however, to developing and applying methods that address specific forms of missing information. This area of research is in its infancy, although Coburn and Vevea have taken steps toward developing models in which the bias pattern may vary with study characteristics (2015).

18.5.1 Outcome Reporting Biases

Outcome reporting bias occurs when a study measures multiple outcomes, and those outcomes that are statistically significant are more likely to be published than those that are not. The issue of outcome reporting bias has received considerable attention in recent years, and empirical research indicates that it is a serious problem, especially for randomized controlled trials in medicine (Hahn, Williamson, and Hutton 2002; Chan, Hrobjartsson, et al. 2004; Chan, Krleza-Jeric, et al. 2004; Chan and Altman 2005). Although few studies examine outcome bias in the social sciences, some evidence indicates that it affects education research (Pigott et al. 2013). A recent survey of psychologists found that at least 63 percent did not report all outcome measures that they assessed (John, Loewenstein, and Prelec 2012). Together, this evidence reinforces the presence and severity of outcome bias.

Although most methods for assessing publication bias are sensitive to outcome reporting bias, they cannot distinguish between that and publication bias from other sources. These methods cannot accommodate patterns of missing data across multiple outcomes measured across all studies, or information across all reported outcomes. A method has been developed that does consider such patterns; it assumes that the most statistically significant outcomes from some fixed number of identically distributed independent outcomes are reported (Hutton and Williamson 2000; Williamson and Gamble 2005). Although its assumptions are unrealistically strict, the model does provide an upper bound on the likely impact of outcome reporting bias, and could help determine whether contacting study investigators for the potentially missing data would be cost effective. However, the bulk of citations of these articles are methodological rather than empirical work, so it appears that employment of the method is not common.

Daniel Jackson, John Copas, and Alexander Sutton developed a selection model for a specific application—success rates of surgery for emergency aneurysm repair—to address outcome reporting bias (2005). The model relies on a specific outcome that it assumes was reported without bias to learn about an outcome that obviously was not. Assuming no other sources of bias at any level, the selection model was identifiable and yielded adjusted estimates. Of course, the assumption that no other bias exists is very restrictive, and Jackson and his colleagues report that a model capable of incorporating both outcome- and study-level bias is in development.

18.5.2 Subgroup Reporting Biases

Subgroup reporting bias is similar to outcome reporting bias in that it involves the omission of one or more uninteresting or nonsignificant subgroup analyses. Either some subgroup results are published and others are not, or all subgroup results may be excluded (Hahn et al. 2000). Subgroup bias has received little attention in the research literature, although there are indications that it exists in medical research (McIntosh and Olliaro 2000). Others note that the prevalence of outcome bias implies the existence of subgroup bias as well (Hahn et al. 2000).

Seokyung Hahn and his colleagues (2000) suggest a sensitivity analysis approach, similar to the Jane Hutton and Paula Williamson (2000) approach for outcome reporting bias, which involves data imputation for missing subgroup analyses under the assumption that the unpublished analyses were nonsignificant. This is a useful start; however, in general, the issue of subgroup bias is underresearched. In the meantime, research guidelines such as PRISMA (preferred reporting items for systematic reviews and meta-analyses) recommend that meta-analysts report all analyses, regardless of significance, and indicate

whether they were planned (Liberati et al. 2009). Doing so may help reduce the impact of subgroup bias.

18.5.3 Time-Lag Bias

Time-lag bias occurs when research with large effect sizes or significant results tends to be stopped earlier than originally planned, published more quickly, or both (Hopewell et al. 2007)—in other words, when the speed of publication depends on the direction or strength of the results (Jadad and Rennie 1998). When such bias operates, the first studies to be published will often show systematically greater effect sizes than subsequently published investigations (Trikalinos and Ioannidis 2005). The cumulative effect size then diminishes over time.

The Proteus effect, named by Thomas Trikalinos and John Ioannidis (2005), is a similar time-related phenomenon in which the exciting findings of the first published study are followed by a series of equally exciting, contradictory studies, while intermediate, less exciting studies are published later on. In the case of the Proteus effect, because large effect sizes in opposite directions are published most quickly, the cumulative effect size may actually increase over time. The effect is named after Proteus, a god who rapidly transformed himself into different figures (Trikalinos and Ioannidis 2005).

Evidence that time-lag biases exist in the field of genetic epidemiology is strong (Ioannidis et al. 2001). It also appears in child psychiatry (Reyes et al. 2011), in clinical trial research (Clarke and Stewart 1998), and in management and industrial-organizational psychology (Banks, Kepes, and McDaniel 2012; Kepes et al. 2012), among others.

Methods for the assessment of time-lag biases are thoroughly described elsewhere (Trikalinos and Ioannidis 2005) and are not included in this chapter.

18.6 EXAMPLES

18.6.1 Data Sets

In this section, we illustrate some of the available methods using two empirical meta-analytic data sets that differ in size: one that is large (containing more than four hundred effect sizes), and one that is small (containing fewer than twenty). Both data sets are available as supplementary material. The large data set is from a social science meta-analysis, consisting of standardized mean differences. The small data set comes from a medical meta-analysis, and consists of log risk ratios.

We use *R* version 3.2.4 for most analyses (R Core Team 2016). As we describe our results, we include relevant sections of *R* code in the text so that interested readers can replicate the examples.

18.6.1.1 Psychotherapy Efficacy The first data set is from a well-known meta-analysis performed by Mary Smith, Gene Glass, and Thomas Miller on the efficacy of psychotherapy (1980). We use a subset of the original data that consists of studies in which the psychotherapy effects being compared include both behavioral and systematic desensitization treatments for phobias. The phobias themselves are also divided into two groups, one consisting of patients suffering from "complex" (multiple) phobias and one of those suffering from "simple" (only one) phobias. The original data set included some effect sizes that modern-day meta-analysts might consider implausibly large; for example, one study reported a standardized mean difference of 25.33. To avoid complications that such huge effects can induce, we deleted five cases with effect sizes larger than 4.0. Of the 489 effect sizes that remain, 216 employ behavioral treatments and 273 employ desensitization therapies. Positive effect sizes indicate effectiveness of psychotherapy.

This is the same data set that Vevea and Hedges (1995) used to demonstrate the use of a linear model for estimating effect size in the presence of publication bias (one of the models included in this chapter). They find that a funnel plot of the effect sizes demonstrated typical one-tailed selection, and the selection model resulted in a reduction of the mean effect size by as much as 25 percent, in the case of desensitization treatment of complex phobias. These results indicate that publication bias does affect this data set.

We read the psychotherapy efficacy data set into *R* with

```
glass <- read.csv("data glass.csv",
header=TRUE)
```

and create variables for the effect sizes and sampling variances:

```
glass_y <- glass$g
```

and

```
glass_v <- glass$v
```

We also create variables for the three moderators we will be using: whether the therapy was behavioral modification,

```
glass_b1 <- glass$b1
```

whether the patients' phobia was simple or complex,

```
glass_b2 <- glass$b2
```

and whether there was an interaction,

```
glass_b3 <- glass$b3
```

which is the product of glass_b1 and glass_b2.

18.6.1.1.1 Funnel Plots. No special software is necessary to create a funnel plot using *R*. For users who prefer to do so, most common meta-analysis packages include a funnel plot function. Wolfgang Viechtbauer's *metafor* includes *funnel()*, which yields both traditional and contour-enhanced funnel plots, with the enhanced plot as the default (2010). Schwarzer's *meta* includes its own *funnel()*, which performs similarly (2016). The funnel plots presented here were created using the basic *R* scatterplot tools, with the margins extended so that there is approximately 5 percent white space between axes and data points. Adding white space aids in the interpretation of asymmetry if data points fall extremely close to the axes. We plot standard error on the *x*-axis and effect size on the *y*-axis.

Figure 18.6 shows the contour-enhanced funnel plot, using *metafor's funnel()* function. This type of plot may be misleading under some circumstances, however, as discussed earlier.

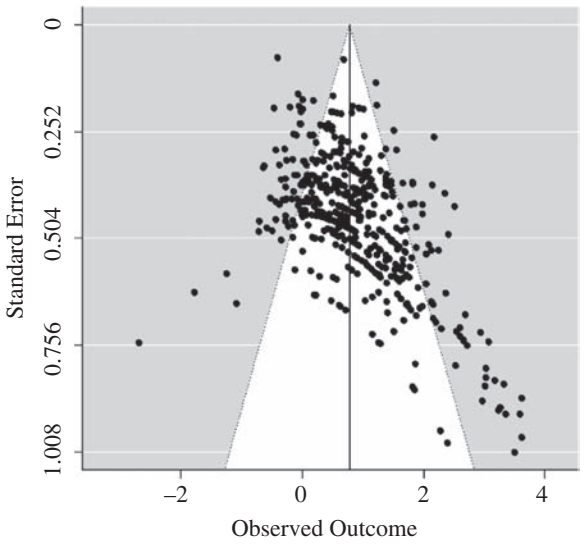

Figure 18.6 Contour-Enhanced Funnel Plot, Psychotherapy Data

SOURCE: Author's tabulation.

We calculated the amount necessary to expand the range of the *x* and *y* axes by 5 percent. Then we created a space for the funnel plot with

```
plot(c(min(sqrt(glass_v))-.0465,
max(sqrt(glass_v))+0.0465),c(min(glass_y)-
0.3145,max(glass_y)+0.3145),type='n',
xlab="Standard Error",ylab="Effect Size")
```

We added the scatterplot points with

```
points(sqrt(glass_v),glass_y)
```

The resulting funnel plot appears in figure 18.7. We computed the mean effect size for this data set using the *metafor* package's *rma()* function:

```
rma(glass_y, glass_v, method='ML')
```

We used a random-effects model and maximum likelihood estimation. This yielded a mean of 0.70 and a variance component of 0.28, which corresponds to I^2 of 66.51 percent. I^2 is fairly large, indicating that the results of some methods—particularly those that cannot accommodate heterogeneity—may be less reliable.

This data set is large, so it is easier to assess asymmetry in the funnel plot. There are many very large effect sizes ($d > 1.00$) with large standard errors (> 0.60) that are not mirrored by small or negative effect sizes; in fact, only *four* effect sizes with standard errors greater than 0.60 fall equally far below. This is an example of asymmetry associated with characteristic one-tailed selection bias. The plot suggests that concern about publication

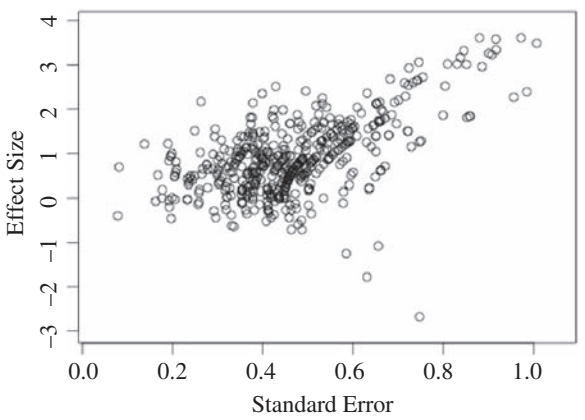

Figure 18.7 Effect Size Against Standard Error, Psychotherapy Data

SOURCE: Author's tabulation.

bias may be appropriate. (The curvature evident in the plot is due to the dependence of standard errors on standardized mean difference; studies with equal sample sizes will have increasingly large standard errors as their effect sizes increase.)

18.6.1.1.2 Cumulative Meta-Analysis. Both *metafor* and *meta* include a function for cumulative meta-analyses; in *metafor* this is *cumul()* and in *meta* it is *metacum()*. Users can create an object to store the results of these analyses and then pass that object to *forest()* using either package. We used *metafor* for our cumulative meta-analyses and forest plots.

The forest plot is not featured here. A forest plot of a cumulative meta-analysis involving 489 effect sizes is more or less impossible to read; the annotations are so tiny as to be invisible, and the average effect estimates resemble a thick black blur. Instead, it is more informative to consider the cumulative meta-analysis itself:

```
glass_cumul <- cumul(glass_rma, order =
order(sqrt(glass_v)))
```

The effect size with the least precision belongs to study 100, with $d = -0.40$. As soon as the second least-precise effect, study 99, is added, the estimate of mean effect jumps to $d = 0.14$. It proceeds from there to $d = 0.49$ with the third least-precise effect and, by the time all effects are included, has drifted all the way to $d = 0.78$. This drift is a sign of a relationship between study size and effect size, which indicates the presence of publication bias.

18.6.1.1.3 Trim and Fill. We used the *trimfill()* function of the *R* package *metafor*. The *R* package *meta* also includes a *trimfill()* function. Users who are interested in separate control of fixed-effect and random-effects models for imputing data and estimating parameters may wish to use *meta*. Duval and Tweedie recommend that trim and fill be used as a sensitivity analysis (2000a, 2000b), so we present the results of trim and fill using both the L_0 and the R_0 estimators, and we consider the possibility of publication bias in favor of both smaller and larger effects.

For each data set, we estimated four trim and fill models. We specified that studies were missing on either the left or the right side of the funnel plot (suppression of smaller or larger effects). In view of the funnel plot, specifying missing effects on the right side makes little practical sense, but we included it for demonstration purposes. We also specified the L_0 or R_0 estimator. The literature is ambiguous about the performance of L_0 versus R_0; see section 18.4.3 for details. We used the *metafor* function *trimfill()*, which can handle either "fixed-fixed" or "random-random" models. (The *meta* package is more versatile in this respect.) The results presented here are "random-random." These factors yielded four separate models.

The models were variations of

```
trimfill(glass_rma, side="left",
estimator="L0")
```

where "L0" was exchanged for "R0" and "left" was exchanged for "right." The results of the trim and fill analyses for the psychotherapy efficacy data set are presented in table 18.2.

Only L_0 added effect sizes, and it did so on the left side of the plot (indicating a suppression of smaller effects). L_0 added 117 effect sizes, reducing the average effect from 0.78 to 0.52 (a difference of 0.26, or 33 percent) and increasing the variance component from 0.29 to 0.58 (a difference of 0.29, or 100 percent).

It seems that this data set may not be robust to publication bias. Nevertheless, an effect persists, even after adjustment.

18.6.1.1.4 Egger's Regression. The package *meta* contains the function *metabias()*, which can be used for both Egger's linear regression and the rank correlation test. The package *metafor* contains *regtest()*, which conducts Egger's linear regression; we used *regtest()*. The function allows users to specify whether they want to estimate a standard Egger's regression, a mixed-effects Egger's regression, or a random-effects Egger's regression. We estimated all three models to compare their conclusions, although because heterogeneity is present in the data set, the mixed- or random-effects models may perform better.

A standard Egger's regression using a weighted regression model

```
regtest(glass_rma, model="lm")
```

was statistically significant, $t(484) = 9.81$, $p < .0001$, indicating that publication bias, or funnel plot asymmetry, is present. The evidence under a random-effects meta-regression model acquired from

```
regtest(glass_rma)
```

was also statistically significant, $z = 12.26$, $p < .0001$.

We incorporated the moderators in this data set as well. There are three dichotomous moderators, as described. Their unadjusted conditional means are presented in the top row of table 18.2.

The mixed-effects variation of Egger's regression was also statistically significant, $z = 11.80$, $p < .0001$. All three variations indicate a relationship between study size and effect size, or that bias may be present.

Table 18.2 Summary of Results for Psychotherapy Data

Method		Overall Mean	BMOD, SP	BMOD, CP
Unadjusted		0.78	0.90	0.63
Trim and fill	left, L_0	0.52 (32.99%, M) *MO*	—	—
	right, L_0	0.78 (0%, A) *MI*	—	—
	left, R_0	0.78 (0%, A) *MI*	—	—
	right, R_0	0.78 (0%, A), *MI*	—	—
PET-PEESE		**–0.04 (105.13%, S) *S***	—	—
Vevea and Hedges	$p = 0.025$	0.68 (12.71%, A) *MI*	—	—
	multiple	0.47 (39.67%, M) *MO*	—	—
	multiple, LM	—	0.65 (27.55%, M) *MO*	0.36 (42.36%, S) *MO*
Vevea and Woods	moderate one-tailed	—	0.78 (13.67%, A) *MI*	0.49 (21.50%, M) *MO*
	severe one-tailed	—	0.56 (37.44%, M) *MO*	**0.21 (65.92%, S) *S***
	moderate two-tailed	—	0.82 (8.78%, A) *MI*	0.57 (9.55%, A) *MI*
	severe two-tailed	—	0.72 (20.11%, M) *MO*	0.49 (21.34%, M) *MO*
Copas and Shi		**0.10 (87.16%, S) *S***	—	—
Rücker		**0.09 (88.45%, S) *S***	—	—
p-uniform		1.06 (36.07%, M) *MO*	—	—

SOURCE: Author's tabulation.

NOTES: Adjusted estimates are reported unless row is labeled "Unadjusted." Percentage adjustment is in parentheses, followed by the Kepes, Banks, and Oh (2012) categorization (A for absent, or < 20 percent adjustment; M for moderate, or adjustment between 20 percent and 40 percent; S for severe, or adjustment > 40 percent). The Rothstein, Sutton, and Borenstein (2005) categorization follows in italics (MI for minimal, or adjustment is similar; MO for moderate, or adjustment is substantial, but key finding remains; S for severe, adjustment that calls the key finding into question). If both categorizations were "Severe," the cell is boldface. BMOD = behavioral modification, SYSDS = systematic desensitization, SP = simple phobia, and CP = complex phobia. ^{1} is the variance component without moderators; ^{2} is the variance component with moderators. Cells marked with a dash either were not or could not be estimated.

18.6.1.1.5 PET-PEESE. We were unable to locate any *R* package capable of implementing PET-PEESE. However, meta-analysts can easily construct the code themselves; PET-PEESE is a pair of linear regressions, which can be estimated using the base *R* function *lm()*.

We estimated the PET regression:

```
pet <- lm(glass_y~sqrt(glass_v),weights =
1/glass_v)
```

followed by the PEESE regression:

```
peese <- lm(glass_y ~ glass_v, weights =
1/glass_v)
```

We stored the estimates from these regressions and kept the PET estimates if PET was nonsignificant; otherwise, we kept PEESE.

PET was nonsignificant, indicating the absence of evidential value, with $p = .56$. This means estimating PEESE is not necessary, and the adjusted estimate of effect size from PET, $d = –0.04$ (see table 18.2) is a result of publication bias. The unadjusted random-effects mean for this data set is 0.78; PET-PEESE reduced the mean by 0.82, or 105 percent. This indicates that the effect size is an artifact of publication bias.

18.6.1.1.6 Nonparametric Correlation Test. The *meta* package contains the function *metabias()*, which can perform the rank correlation test. The *metafor* package contains *ranktest()*. We use *ranktest()* to conduct the analyses. The rank correlation test for funnel plot asymmetry is not model-based, and changing the meta-analytic model does not change the results. The rank correlation results we present here have all used Kendall's tau, but the *R* function can accommodate other correlation coefficients.

The rank correlation test was provided by

```
ranktest(glass_rma)
```

SYSDS, SP	SYSDS, CP	τ_1^2	τ_2^2
0.86	0.70	0.29	0.28
—	—	0.58 (98.98%, S) *MO*	—
—	—	0.29 (0%, A) *MI*	—
—	—	0.29 (0%, A) *MI*	—
—	—	0.29 (0%, A) *MI*	—
—	—	—	—
—	—	0.27 (7.85%, A) *MI*	—
—	—	0.37 (26.28%, M) *MO*	—
0.61 (28.82%, M) *MO*	0.42 (40.63%, S) *MO*	—	0.36 (22.22%, M) *MO*
0.73 (16.76%, A) *MI*	0.56 (20.60%, M) *MO*	—	0.32 (15%, A) *MI*
0.51 (40.26%, S) *MO*	0.28 (59.80%, S) *MO*	—	0.44 (56.07%, S) *MO*
0.78 (8.75%, A) *MI*	0.63 (10.23%, A) *MI*	—	0.27 (5.36%, A) *MI*
0.69 (19.95%, A) *MO*	0.54 (22.87%, M) *MO*	—	0.24 (15.36%, A) *MI*
—	—	—	—
—	—	—	—
—	—	—	—

and was significant, with a Kendall's tau of 0.32 and $p < .0001$. This indicates that we can reject the null hypothesis of no correlation and conclude a danger of publication bias, or of funnel plot asymmetry.

18.6.1.1.7 p-*Curve and* p-*Uniform*. *P*-curve is not available as an *R* package, but Simonsohn, Nelson, and Simmons created a web application (www.p-curve.com) (2014). The application requires users to enter data in the form of the original test statistics rather than effect sizes, likely to encourage meta-analysts to think carefully about which tests are included. We include a caveat here—we are demonstrating *p*-curve using empirical sets of effect sizes and sampling variances, not raw test statistics from the corresponding studies. This exercise is solely for demonstration purposes.

The graph produced by *p*-curve is presented in figure 18.8. The distribution of *p*-values is visibly right skewed. The binomial test for right skew, comparing the proportions of *p*-values below and above .025, was nonsignificant with $p = .13$, but continuous tests, both for the full *p*-curve and for the half *p*-curve, were significant: $z = -3.64$, $p = .0001$ and $z = -4.17$, $p < .0001$, respectively. These results indicate that these studies do contain evidential value; the effect is not completely due to publication bias.

Although not entirely necessary because right skew is present, the binomial test for underpowered studies is significant, $p = .01$. However, the continuous test for underpowered studies is not, with $z = -0.78$ and $p = .22$. Despite the fact that evidential value is present, the continuous test indicates that data may be underpowered, or (according to the application) that the evidential value is inadequate. The studies conducted may have had reduced power; perhaps the researchers did not conduct a priori power analyses.

No *R* package for *p*-uniform is available on the Comprehensive *R* Archive Network, but Robbie van Aert

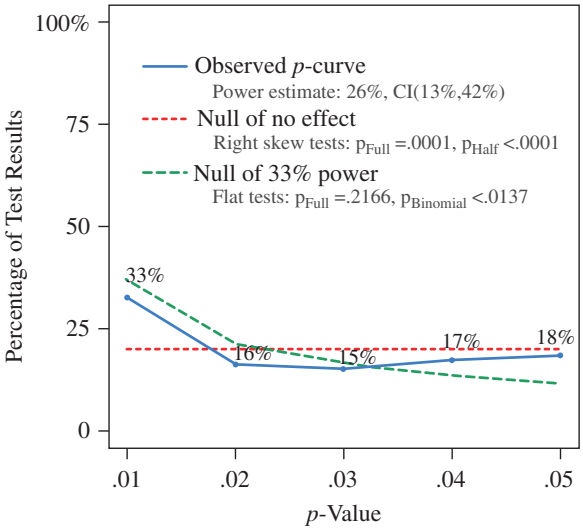

Figure 18.8 *p*-Curve, Psychotherapy Data

SOURCE: Authors' tabulation obtained directly from the *p*-curve application (v. 4.06), available at http://www.p-curve.com/app4/ (accessed January 7, 2019).
NOTE: The observed *p*-curve includes 92 statistically significant ($p<.05$) results, of which 52 are $p<.0.5$. There were 395 additional results entered but excluded from *p*-curve because they were $p>0.5$.

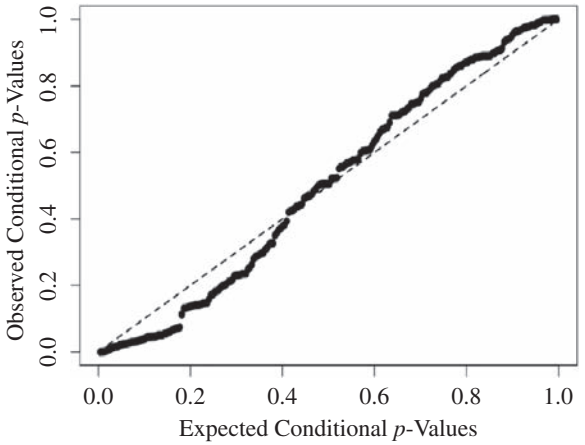

Figure 18.9 *p*-Uniform, Psychotherapy Data

SOURCE: Author's tabulation.

(2015) has uploaded a preliminary version of his package, called *puniform*, on GitHub. GitHub is a less regulated analog of CRAN. To install packages directly to *R* using GitHub, users must first install the *R* package *devtools*, then type *install_github("author/package")*. For *puniform*, this would look like *install_github("RobbievanAert/puniform")*. The function requires that users specify the alpha level of included studies. We entered an alpha level of 0.05. The function also produces a plot of observed conditional *p*-values against expected ones, so users can visually assess deviation from uniformity.

We estimated *p*-uniform:

```
puniform(yi = glass_y, vi = glass_v,
alpha = 0.05,side="right", method="P",
plot=TRUE)
```

The plot of observed versus expected *p*-values is in figure 18.9. There is some deviation from uniformity in the areas of the graph between expected *p*-values of about .60 and 1.00 and between about .10 and .40. The one-tailed test for publication bias was nonsignificant, with $z = -7.96$ and $p > .999$. The adjusted fixed-effect estimate was $d = 1.06$ (see table 18.2), $p < .001$. This is an increase of 0.44, or 71 percent.

P-uniform does not indicate that publication bias is a serious threat for this data set. It is unusual, though, that the mean was adjusted upward, indicating that larger studies were suppressed.

18.6.1.1.8 Excess Significance Test. No *R* package is available to implement the excess significance test. However, it is not complicated to conduct the test in base *R* by calculating the post hoc power for each significant study in your data set and multiplying them. We do not provide the *R* code in text as ours was lengthy and calculating power, of course, depends on the format of your effect sizes.

The product of the power for each significant effect size was 0.00 ($p < 0.10$), so we can reject the null hypothesis and conclude that publication bias may be present.

18.6.1.1.9 Dear and Begg. We faced an estimation problem with the psychotherapy efficacy data set, and were unable to obtain results from either the Dear and Begg (1992) weight-function model or the modified version by Rufibach (2011), because both versions attempted to estimate *n*/2 (or 244) parameters.

18.6.1.1.10 Vevea and Hedges. In 2016, Coburn and Vevea released an *R* package titled *weightr* that can estimate both the Vevea and Hedges (1995) model and the Vevea and Woods (2005) model. We use the package *weightr* to perform these demonstrations.

Meta-analysts who wish to use either the Vevea and Hedges (1995) or the Vevea and Woods (2005) model

will undoubtedly wonder about the *p*-value cut points—how many to specify, whether to specify a one-tailed or two-tailed pattern, which to specify, and so on. The choice of cut points is entirely up to the researcher. We do not recommend any specific pattern over another, nor do we intend for the cut points used here to become a canonical guideline. We cannot emphasize this enough. The researcher must consider the data set and select a series of cut points such that at least some observed effect sizes fall within each interval, and the cut points correspond to *p*-values that may be of psychological relevance. Regarding the latter, it is poor practice to present only the results of a model with a cut point at, say, $p = .129$ if the results drastically differ when the cut point is changed (for instance, to $p = .20$). Do not select cut points to force the model into significance or nonsignificance. Specify several different sets of cut points and observe the change in the adjusted estimates; if there are moderators, calculate conditional means and assess their changes as well. Perhaps most important, present the results of all models specified, to compare the adjusted estimates across models. *Always* report the cut points specified (preferably along with their rationale).

No guideline is in place for the number of cut points meta-analysts should specify for a given *k*—though, of course, it is impossible to estimate more cut points than there are effect sizes, and therefore the number of cut points must be less than *k*. The meta-analyst should survey the distribution of observed *p*-values and ensure at least some observed effect sizes in each specified interval, a fact that can be verified using the table=TRUE argument in *weightr*. If an interval is empty of observed effect sizes or contains only a few effects, this will immediately be evident, because *weightr* will yield a warning and the parameter estimates may be nonsensical or missing due to model nonconvergence.

For these examples, to maintain consistency across data sets, we have specified the same set of *p*-value cut points. Remember: The unadjusted mean for the psychotherapy efficacy data set is $d = 0.78$, and the unadjusted variance component is 0.29. All these results are presented in table 18.2.

First, we specify one *p*-value cut point at $p = .025$.

```
weightfunct(glass_y, glass_v)
```

The $p = .025$ corresponds to the positive tail in a two-tailed test with an alpha of 0.05. This yields a weight for the nonsignificant interval ($.025 < p < 1.00$) of 0.68, indicating that nonsignificant studies are 68 percent as likely to survive selection as significant ones. The mean effect is adjusted downward to 0.68 (a change of 0.10, or 13 percent), and the variance component is also adjusted downward to 0.27 (a change of 0.02, or 7 percent). The likelihood-ratio test comparing the adjusted and unadjusted models is significant, $p < .05$, which indicates that the adjusted model fits the data better, and hence that publication bias is present.

Next, we estimate a more detailed one-tailed selection pattern ($p = .01, .025, .05, .10, .20, .30, .50$, and 1.00). The first four cut points are at *p*-values that correspond to common alpha levels: $p = .10$ is often referred to as *marginal significance*; $p = .025$ corresponds to the positive tail of a two-tailed test at an alpha level of .05; $p = 0.50$ represents the point at which most effect-size measures become negative. There are no cut points after $p = 0.50$ because we wish to specify one-tailed selection; $p = .20$ and $p = .30$ are included because enough observed effects fall in that range for the model to estimate weights.

We enter:

```
weightfunct(glass_y, glass_v, steps=c(0.01,
0.025,0.05, 0.30, 0.50, 1.00))
```

Now the adjusted mean effect size has been reduced even further, to 0.47 (a change of 0.31, or 40 percent), and the variance component has increased to 0.37 (a change of 0.08, or 28 percent). The likelihood-ratio test is still significant, indicating that the adjusted model is a better fit.

Finally, we include a linear model for the mean. We specify the same pattern of one-tailed *p*-value cut points as before. The command is:

```
weightfunct(glass_y, glass_v, mods=~glass_b1
+ glass_b2 + glass_b3, steps=c(0.01, 0.025,
0.05, 0.30, 0.50, 1.00))
```

The likelihood-ratio test comparing this adjusted model to its unadjusted counterpart is still significant, indicating that it is a better fit for the data. We are now presented with not only an adjusted variance component (of 0.36, a change of 0.07 or 24 percent) but also adjusted conditional means for all six groups.

Some conditional means were adjusted more than others, likely based on how much the effect sizes in that particular group are susceptible to publication bias. More specifically, the conditional means for both systematic desensitization and behavioral modification with complex phobias appear to be more heavily affected by publication bias than the other two conditional means (a change of 41 percent and 42 percent, respectively). The other two conditional means were changed by 28 percent and

29 percent. This difference occurs because the less affected subset of effects is systematically larger than other effects, so that the bulk of them fall within the range of the highly significant cut points, where weights tend to be high. The model can adjust some means more than others, as necessary. The results may indicate that more bias is present among these groups of effect sizes.

18.6.1.1.11 Vevea and Woods. The Vevea and Woods (2005) model also requires meta-analysts to specify a series of *p*-value cut points, this time along with a fixed weight for each interval. Because this model does not estimate any weights for the *p*-value intervals, it does not matter how many cut points the meta-analyst specifies. It is possible to specify more cut points than observed effect sizes (that is, the number of cut points can be greater than *k*). It is also possible to specify any weights for those intervals, bearing in mind that for convenience of interpretation the weights should be between 0 and 1. If the specified weights increase as the *p*-value cut points decrease—that is, if *p*-values < .05 are the most likely to survive—this represents traditional one-tailed selection. The model is flexible; researchers can specify two-tailed selection, or any selection pattern, using any *p*-value cut points.

Interpreting the results of the Vevea and Woods (2005) model comes with a caveat. The model does not estimate the pattern of selection; the researcher chooses a pattern of selection and imposes that pattern on the observed effect sizes, then the mean (or set of conditional means) and variance component are adjusted according to the pattern. The idea is to conduct multiple analyses with various weight patterns representing different degrees of selection severity. Some patterns may lead to ludicrous estimates of the mean (or conditional means). But often the mean, or a particular conditional mean, may be relatively unaffected by any reasonable pattern of weights. Under those circumstances, the researcher can be confident that the magnitude of the mean is not principally an artifact of *p*-value based selection. But none of these estimates should be regarded as a true bias-corrected estimate.

The *R* package *weightr* can implement the Vevea and Woods (2005) adaptation. We used *weightr* to conduct the following analyses. We specified the sets of cut points and weights ("moderate" and "severe" one-tailed and two-tailed selection) mentioned in Vevea and Woods, but we emphasize that these are not hard and fast guidelines for severity of selection. They are merely an indication of what severity of selection bias might look like. Bias patterns most likely vary widely both across and within fields, and we are merely using these weights for demonstration purposes. We do not mean for them to become canonical specifications. These sets of weights, and the bias patterns they theoretically represent, are presented in table 18.1.

The results of the analyses appear in table 18.2. The code we used consists of variations on

```
weightfunct(glass_y, glass_v, mods=~glass_b1
+ glass_b2 + glass_b3, steps=c(0.005,
0.010, 0.050, 0.100, 0.250, 0.350, 0.500,
0.650, 0.750, 0.900, 0.950, 0.990, 0.995),
weights=c(1, 0.99, 0.95, 0.80, 0.75, 0.65,
0.60, 0.55, 0.50, 0.50, 0.50, 0.50, 0.50))
```

We replaced the weights vector with the corresponding set of weights for each selection pattern.

Some of the conditional means are affected more than others. All means are most attenuated in the severe one-tailed bias condition, but the means for the group with simple phobias under both treatment conditions are reduced much less than the others. Even though all the conditional means are attenuated to one degree or another, however, these results at least indicate that none of the bias patterns we attempted reversed the direction of the effects, and only two means were reduced below 0.25 for any of the four selection patterns. This implies that, although publication bias has the potential to affect the mean estimates, the changes are only large for specific combinations of treatment and complex phobias. Hence, it is unlikely that publication bias, if present, would overturn conclusions about the positive effects of psychotherapy for these conditions.

18.6.1.1.12 Copas and Shi. The Copas and Shi (2001) model can easily be implemented using the *R* package *metasens* (Schwarzer, Carpenter, and Rücker 2016) and the function *copas()* (Carpenter et al. 2009).

We first estimated a random-effects meta-analysis with maximum likelihood:

```
glass_meta <- metagen(TE = glass_y, seTE =
sqrt(glass_v), method.tau = "ML")
```

Then we estimated the Copas and Shi selection model:

```
cop.glass <- copas(glass_meta)
plot(cop.glass)
summary(cop.glass)
```

The four plots produced by the Copas and Shi (2001) selection model *R* function (Carpenter et al. 2009) are presented in figure 18.10. We begin with the top right plot, the contour plot. The contour lines are straight, indicating

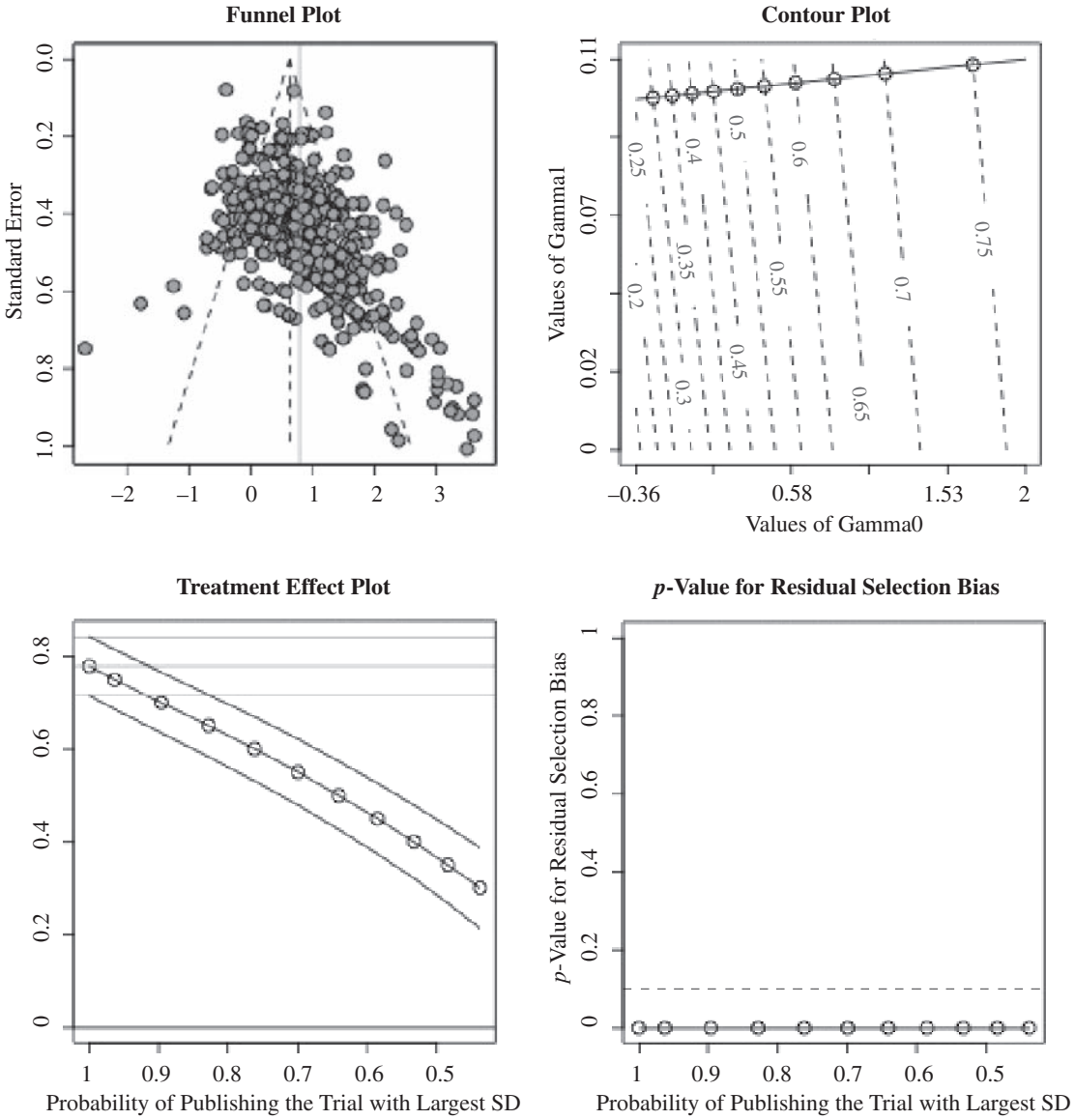

Figure 18.10 Copas and Shi, Psychotherapy Data, $a = -0.36$ to 2, $b = 0$ to 0.11

SOURCE: Author's tabulation.
NOTE: Gamma zero and gamma one represent a and b, respectively.

little difficulty in model estimation, and most of the contour lines are close together, indicating that the data set is not very robust to changes in selection bias patterns. The estimate at the top right of the contour plot, under little to no selection bias, is about $d = 0.75$. The unadjusted random-effects estimate for the psychotherapy data set is $d = 0.78$.

The treatment effect plot indicates that, with no selection bias (a probability of 100 percent), the estimated average effect size is about $d = 0.80$. When the probability of publishing the effect size with the largest standard error reaches 50 percent, the average effect size has plummeted to about 0.30.

The *p*-value plot gives us some problems. The *R* function automatically scanned a range of *a* and *b* values for which the *p*-value associated with residual selection bias never becomes nonsignificant, as it does not cross the horizontal dashed line. This means that the *R* function cannot give us an adjusted estimate. However, the choices for *a* and *b* can be manually altered. The function allows users to specify a range of values for *a* and *b*. We can see the range that the function's algorithm chose by looking at the axes of the contour plot; the algorithm scanned from –0.36 to 2 for *a* (gamma zero in the software nomenclature) and from 0 to 0.11 for *b* (gamma one). We change these and specify a wider range of values—from –2 to 2 for *a* and from 0 to 1 for *b*. This yields the plots in figure 18.11.

It appears that our conclusion was correct; the function was simply not scanning a wide enough range of values by default. Now, if we look at the treatment effect plot, we get an estimated mean effect size for probabilities beyond 20 percent—the effect size has dropped as far as about 0.10. The *p*-value plot shows that, when the probability of publishing an effect size with the largest standard error reaches 10 percent or so, the test for residual selection bias finally becomes nonsignificant. The most likely scenario given these observed data is a probability of 10 percent. This is very strong selection bias.

The adjusted mean effect size that the Copas and Shi (2001) model provides for this most likely scenario is $d = 0.10$ (see table 18.2), a change of 87 percent from the unadjusted $d = 0.78$.

18.6.1.1.13 Rücker Limit Meta-Analysis. To implement the Rücker et al. (2011) limit meta-analysis method, we used the *R* package *metasens* and the function it provides, *limitmeta()*. The *R* function yields a test of heterogeneity, a test of small-study effects (on alpha), a test of residual heterogeneity after accounting for small-study effects, and an adjusted estimate of the average effect size.

We estimated the Rücker limit meta-analysis method:

```
glass_limit <- limitmeta(glass_meta)
summary(glass_limit)
```

The results indicate a significant relationship between effect size and standard error. The test for small-study effects was significant, $Q(1) = 304.35$, $p < .0001$, as was the test for residual heterogeneity, $Q(487) = 1255.21$, $p < .0001$.

The adjusted random-effects estimate for the mean effect is 0.09 (table 18.2), with a 95 percent confidence interval from 0.01 to 0.17. According to the model, the effect size for a study having infinite precision is 0.09. This is a change of 0.69, or 88 percent.

18.6.1.2 Irritable Bowel Syndrome The second data set consists of nineteen trials examining the response rate of patients with irritable bowel syndrome to complementary and alternative medicine (CAM) therapies (Dorn et al. 2007). Only randomized, placebo-controlled trials were included. The CAM response rate was high across trials—more than 40 percent. Risk ratios greater than 1 indicate that patients undergoing CAM therapies had a higher response rate than those undergoing placebo therapies.

Spencer Dorn and his colleagues (2007) assessed publication bias using a funnel plot, Egger's regression, and Begg and Mazumdar's rank correlation test. They concluded that the funnel plot displayed asymmetry and found that Egger's regression was significant ($p = .03$) and the rank correlation was "trend[ing] towards significance" (632), with $p = .06$. They also noted that, of nineteen trials, twelve were statistically significant.

We read the irritable bowel syndrome data set into *R* with

```
ibs <- read.csv("data IBS.csv",
header=TRUE)
```

and create variables for the effect sizes and sampling variances:

```
ibs_y <- ibs$LogRR
```

and

```
ibs_v <- ibs$v
```

(The "header=TRUE" component of the R command is appropriate if the data file contains variable names in the first row, as this one does.)

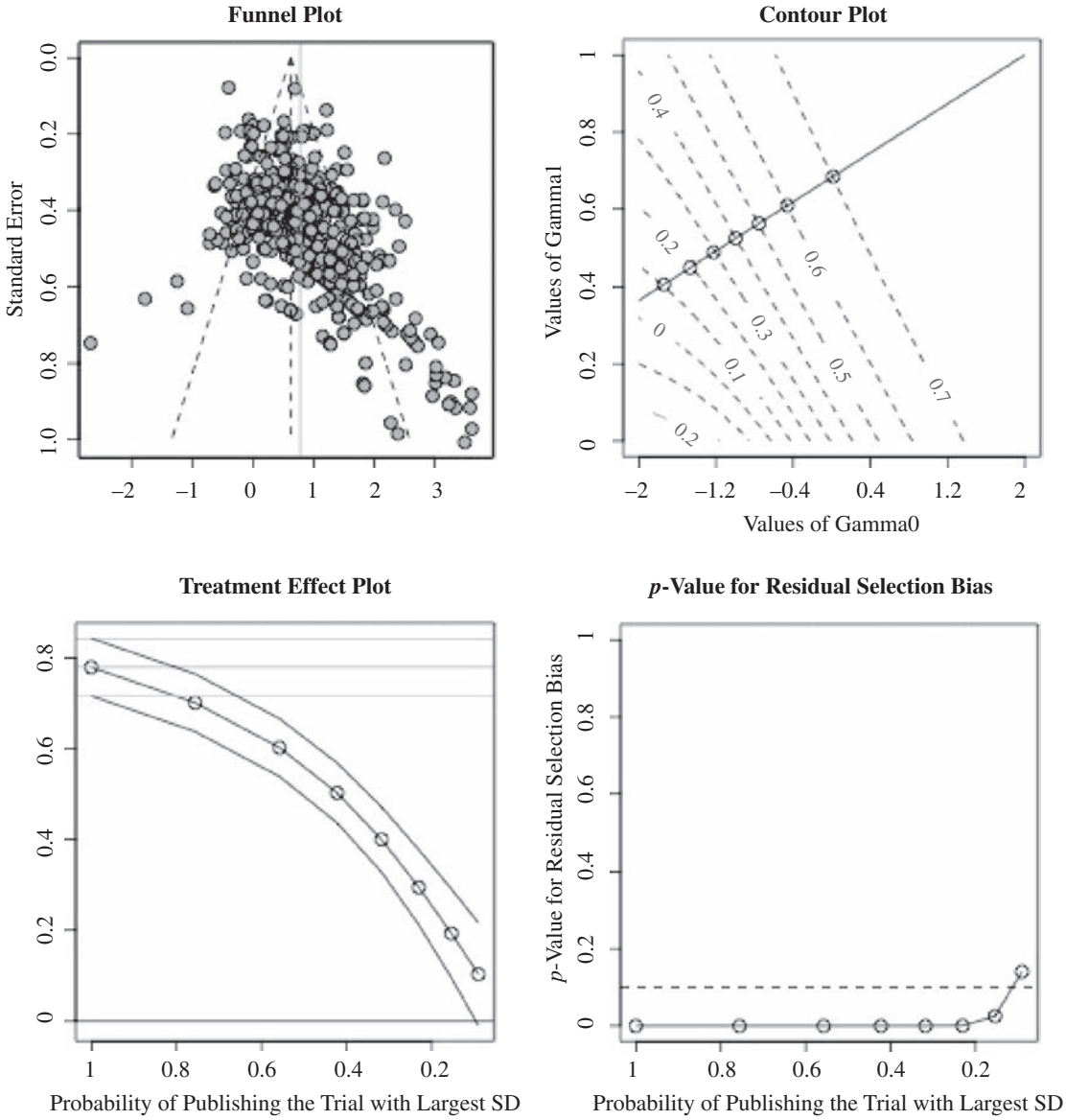

Figure 18.11 Copas and Shi, Psychotherapy Data, $a = -2$ to 2, $b = 0$ to 1

SOURCE: Author's tabulation.
NOTE: Gamma zero and gamma one represent a and b, respectively.

18.6.1.2.1 Funnel Plots. We calculated the amount necessary to expand the range of the *x* and *y* axes by 5 percent. Then we created a space for the funnel plot with

```
plot(c(min(sqrt(ibs_v))-.0193, max(sqrt
(ibs_v))+0.0193),c(min(ibs_y)-.1057,
max(ibs_y) + 0.1057),type='n', xlab=
"Standard Error",ylab="Effect Size")
```

We added the scatterplot points with

```
points(sqrt(ibs_v),ibs_y)
```

The funnel plot for the irritable bowel syndrome data set is featured in figure 18.12. Again, we computed the mean effect size for this data set using the *metafor* package's *rma()* function:

```
rma(ibs_y, ibs_v, method='ML')
```

We used a random-effects model and maximum likelihood estimation. This yielded a mean effect size of 0.42 and a variance component of 0.15 (see table 18.3), which corresponds to a large I^2 (72.90 percent). There are only

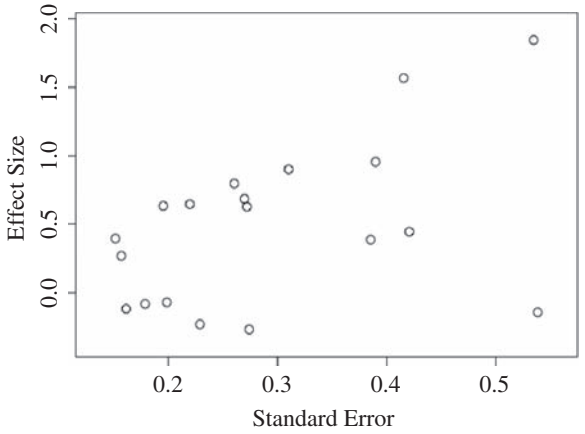

Figure 18.12 Effect Size Against Standard Error, Irritable Bowel Syndrome Data

SOURCE: Author's tabulation.

Table 18.3 Summary of Results for Irritable Bowel Syndrome Data

Method		Overall Mean	τ^2
Unadjusted		0.42	0.15
Trim and fill	left, L0	0.42 (0%, A) *MI*	0.15 (0%, A) *MI*
	right, L0	0.42 (0%, A) *MI*	0.15 (0%, A) *MI*
	left, R0	0.38 (9.09%, A) *MI*	0.18 (20.26%, M) *MO*
	right, R0	0.418 (0%, A) *MI*	0.15 (0%, A) *MI*
PET-PEESE		**–0.23 (155%, S), *S***	—
Vevea and Hedges	$p = 0.025$	0.16 (61.72%, S), *MO*	0.09 (41.18%, S), *MO*
	multiple	—	—
Vevea and Woods	moderate one-tailed	0.32 (24.64%, M) *MO*	0.17 (10.46%, A) *MI*
	severe one-tailed	**0.11 (74.64%, S) *S***	0.23 (49.67%, S) *MO*
	moderate two-tailed	0.37 (11.00%, A) *MI*	0.14 (11.11%, A) *MI*
	severe two-tailed	0.32 (24.16%, M) *MO*	0.11 (26.14%, M) *MO*
Copas and Shi		0.25 (40.19%, S) *MO*	—
Rücker		**0.09 (78.47%, S) *S***	—
p-uniform		0.64 (53.11%, S) *MO*	—

SOURCE: Author's tabulation.

NOTES: Adjusted estimates are reported unless row is labeled "Unadjusted." Percent adjustment is in parentheses, followed by the Kepes, Banks, and Oh (2012) categorization (A for Absent, or < 20% adjustment; M for Moderate, or adjustment between 20% and 40%; S for Severe, or adjustment > 40%). The Rothstein, Sutton, and Borenstein (2005) categorization follows in italics (MI for Minimal, or adjustment is similar; MO for Moderate, or adjustment is substantial, but key finding remains; S for Severe, adjustment that calls the key finding into question). If both categorizations were "Severe," the cell is bolded. Cells with "—" either were not or could not be estimated.

nineteen effect sizes here, so this funnel plot is more difficult to assess. There does appear to be greater density at the top of the funnel than at the bottom, and several large effect sizes (> 0.90) with large standard errors are not mirrored by an equivalent number of smaller effects. Despite the small size of the data set, there are enough signs of asymmetry that bias may still be a concern.

18.6.1.2.2 Cumulative Meta-Analysis. We took the object containing the results of the random-effects meta-analysis, ibs_rma, and created a cumulative meta-analysis:

```
ibs_cumul <- cumul(ibs_rma, order =
order(sqrt(ibs_v)))
```

Then we made a forest plot of the cumulative meta-analysis:

```
forest(ibs_cumul)
```

The forest plot for the irritable bowel syndrome data set appears in figure 18.13. The vertical dashed line represents an effect size—here, log(*RR*)—of 0.00. The least precise study is study 19, with log(*RR*) = 0.40. As more and more precise studies are added, the average effect size drifts to the left a bit, eventually going as far as log(*RR*) = 0.13. However, by the time the most precise studies are added, the average effect size has arrived back where it began, at log(*RR*) = 0.42. This is a very small drift, of about 0.02—not exactly indicative of a relationship between study size and effect size. However, the pattern is unusual. After the first few lines of the plot, the drift is consistently toward larger effects as studies with greater precision are added to the analysis. That would be consistent with publication in the unexpected direction.

18.6.1.2.3 Trim and Fill. Again, we estimated four trim and fill models. The first three were variations of

```
trimfill(ibs_rma, side="left",
estimator="L0")
```

where "L0" was exchanged for "R0." The second two were the same variations of

```
trimfill(ibs_rma, side="right",
estimator="L0")
```

The results of the trim and fill analyses for the irritable bowel syndrome data set are presented in table 18.3. This time, only the R_0 estimator added any additional effects. It imputed one effect on the left side of the funnel plot, reducing the mean from 0.42 to 0.38 (an attenuation of 9.52 percent) and increasing the variance component

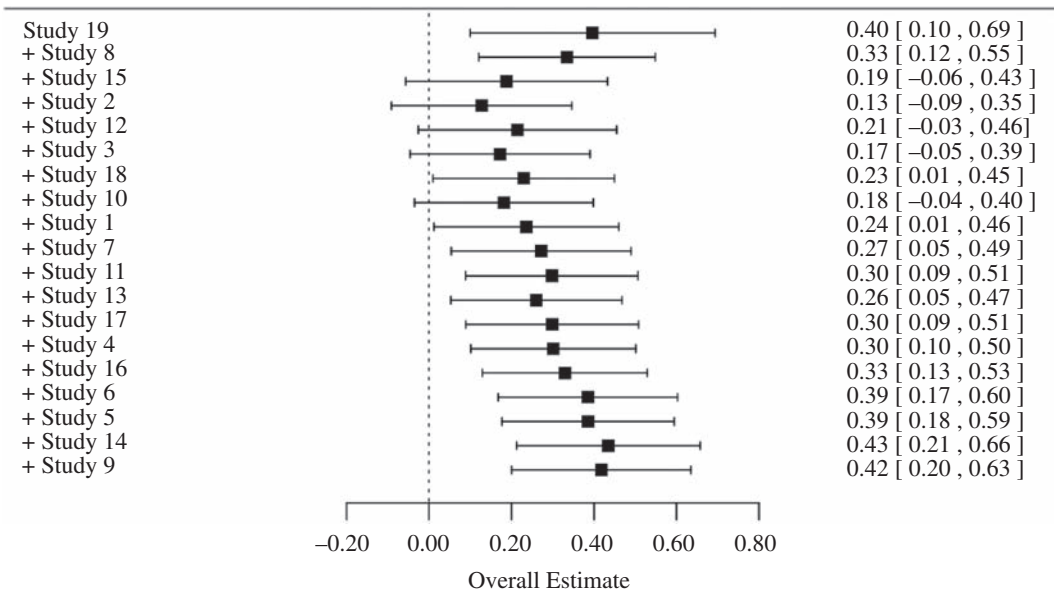

Figure 18.13 Cumulative Meta-Analysis, Irritable Bowel Syndrome Data

SOURCE: Author's tabulation.

from 0.15 to 0.18 (a change of 120 percent). The average effect size has barely been reduced. The trim and fill results for this data set indicate that the irritable bowel data set is very robust to the effects of publication bias.

18.6.1.2.4 Egger's Regression. We estimated a standard Egger's regression of effect size on standard error using weighted regression with multiplicative dispersion

```
regtest(ibs_rma, model="lm")
```

resulted in a test for funnel plot asymmetry on the intercept that was significant, $t(17) = 2.22$, $p = .04$. This indicates that a relationship may exist between study size and effect size.

We also estimated a variation of Egger's regression that predicts effect size with standard error using a random-effects meta-regression model

```
regtest(ibs_rma)
```

A test on the intercept of this model was also significant, $z = 2.64$, $p = .01$. We can reject the null hypothesis and conclude that there may be some evidence of bias.

18.6.1.2.5 PET-PEESE. We estimated PET:

```
pet <- lm(ibs_y ~ sqrt(ibs_v), weights =
1/ibs_v)
```

followed by PEESE:

```
peese <- lm(ibs_y ~ ibs_v, weights = 1/ibs_v)
```

We stored the estimates from these regressions and kept the PET estimates if PET was nonsignificant; otherwise, we kept PEESE.

PET was nonsignificant, indicating the absence of evidential value ($p = .39$). This means that estimating PEESE is not necessary, and the adjusted estimate of effect size from PET, $d = –0.23$ (table 18.3) is due to publication bias. the adjustment is an attenuation of about 156 percent.

18.6.1.2.6 Nonparametric Rank Correlation. The rank correlation returned by

```
ranktest(ibs_rma)
```

was nonsignificant, with Kendall's tau = 0.26 ($p = .13$). This indicates a lack of any significant correlation between effect size and sampling variance, or that no significant evidence of publication bias exists.

18.6.1.2.7 p-*Curve and* p-*Uniform.* The graph *p*-curve produced is presented in figure 18.14. This distribution of *p*-values is also visibly right skewed, and the tests that *p*-curve conducts agree. All three tests for right skew—the binomial ($p = .001$), the continuous full *p*-curve ($z = –4.47$, $p < .0001$), and the continuous half *p*-curve ($z = –2.87$, $p = .00$)—were significant, indicating that the data set contains evidential value.

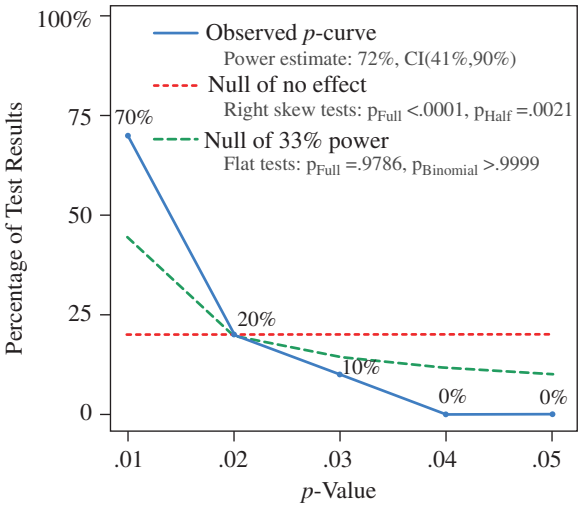

Figure 18.14 *p*-Curve, Irritable Bowel Syndrome Data

SOURCE: Authors' tabulation obtained directly from the *p*-curve application (v. 4.06), available at http://www.p-curve.com/app4/ (accessed January 7, 2019).

NOTE: The observed *p*-curve includes 10 statistically significant ($p<.05$) results, of which 10 are $p<.0.5$. There were 9 additional results entered but excluded from *p*-curve because they were $p>.0.5$.

The binomial test ($p > .999$) and continuous test ($p = .98$) assessing whether the studies are underpowered were both nonsignificant. This indicates that the studies are not underpowered, which makes sense given that right skew is present (a sign of evidential value).

We estimated *p*-uniform:

```
puniform(yi = ibs_y, vi = ibs_v, alpha =
0.05, side="right", method="P", plot=TRUE)
```

The plot of observed versus expected *p*-values for the irritable bowel syndrome data set is in figure 18.15.

This data set does not include many significant *p*-values. (Recall that *p*-uniform analyses only significant *p*-values.) The pattern of deviations from uniformity appears similar to that from the psychotherapy effectiveness data—deviations occur more at the very small and very large parts of the *x*-axis.

The one-tailed test for publication bias was nonsignificant, with $z = –2.61$ and $p = .99$. The adjusted fixed-effects estimate was $d = 0.64$ (0.41, 0.90), $p < .001$ (see

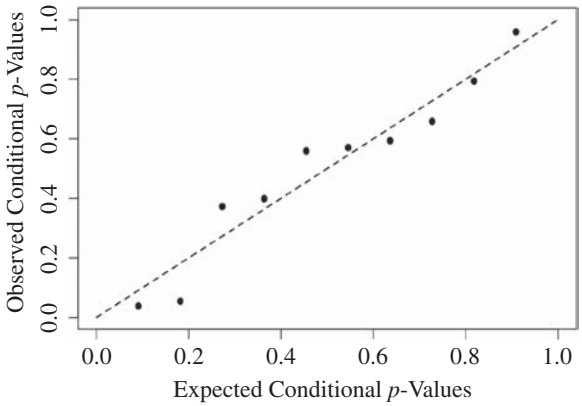

Figure 18.15 *p*-Uniform, Irritable Bowel Syndrome Data

SOURCE: Author's tabulation.

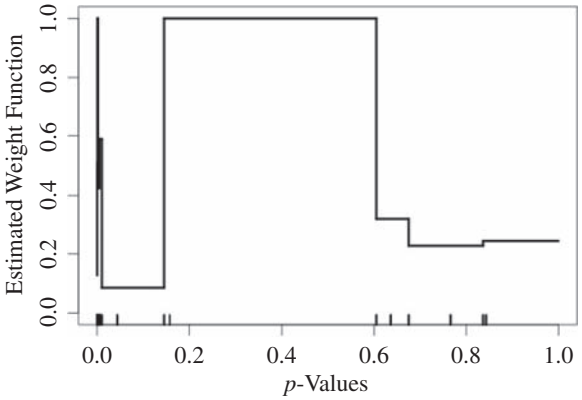

Figure 18.16 Dear and Begg, Irritable Bowel Syndrome Data

SOURCE: Authors' tabulation obtained directly via the *R* package *selectMeta* (v. 1.0.8).

table 18.3). *p*-uniform indicates that publication bias is not a threat for this data set.

18.6.1.2.8 Excess Significance Test. For the irritable bowel syndrome data, the product of the power for each significant effect size was also 0.00. Publication bias may be present.

18.6.1.2.9 Dear and Begg. The Dear and Begg (1992) weight-function model can be implemented in *R* using the *R* package *selectMeta* and its function *DearBegg()*. Meta-analysts must first run *DearBegg()* on their effect sizes and sampling variances, then use the examples in the package manual to create a plot of the resulting weight function. *DearBegg()* itself produces matrices of all the weight estimates for *p*-value intervals, and these must be plotted to be meaningful.

We estimated the Dear and Begg weight-function model:

```
ibs_db <- DearBegg(ibs_y, sqrt(ibs_v),
trace=FALSE)
```

The plot of the Dear and Begg model is presented in figure 18.16.

The weight function has a spike at the far left of the plot, near $p = 0$, indicating that effect sizes with *p*-values in that range are more likely to be observed (a sign of publication bias). However, for observed *p*-values in the range from about .18 to .60, the probability of surviving selection is also high, indicating that nonsignificant studies in that range are also likely to be observed. The spike at the far left, therefore, may not be a matter for concern. It is difficult to determine whether publication bias is a threat based on the plot.

Rufibach's modification can be implemented using *selectMeta*, the same *R* package. The function for Rufibach's modification is *DearBeggMonotone()*, and the results of this function also must be plotted to be meaningful. *R* code to construct plots is featured in the package manual.

We estimated the Rufibach (2011) weight-function model:

```
ibs_db <- DearBeggMonotone(ibs_y,
sqrt(ibs_v), trace=FALSE)
```

The plot of the Rufibach weight-function model is featured in figure 18.17. This weight function indicates that all but the most significant studies have a low probability of surviving publication, and as *p*-value increases the likelihood of surviving decreases even further. For studies with a *p*-value near 1.00, this probability is close to .00. The plot does indicate that publication bias is a concern. The difference between this plot and the one observed for the Dear and Begg method is due to the constraint that the Rufibach model imposes on the weight function—the required monotonicity suppresses the higher weights for the 0.18 to 0.60 *p*-value range.

18.6.1.2.10 Vevea and Hedges. We refresh readers' memory that the unadjusted mean for the irritable bowel syndrome data set is $\log(RR) = 0.42$ and the unadjusted variance component is 0.15.

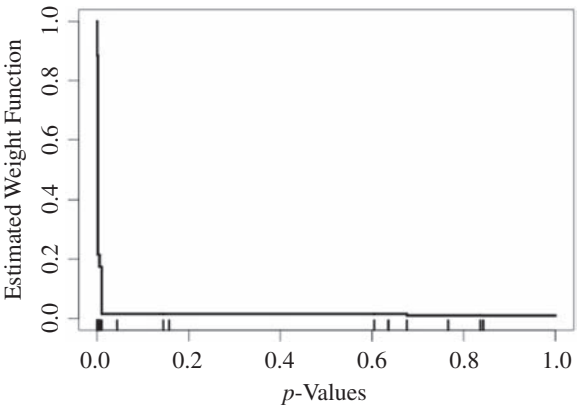

Figure 18.17 Rufibach, Irritable Bowel Syndrome Data

SOURCE: Authors' tabulation obtained directly via the *R* package *selectMeta* (v. 1.0.8).

We begin by specifying one *p*-value cut point, at $p = .05$:

```
weightfunct(ibs_y, ibs_v)
```

The weight for the interval $p < .05$ is, of course, fixed to one. The weight for the interval $0.05 < p < 1.00$ is estimated at 0.19, indicating that nonsignificant studies are 19 percent as likely to survive selection as significant ones. The mean effect size is adjusted downward to 0.16, an attenuation of about 50 percent, and the variance component is also adjusted downward to 0.09, an attenuation of 20 percent (see table 18.3).

Next, we attempt to specify the more detailed one-tailed pattern of *p*-value cut points, but we cannot estimate it—several intervals have no effect sizes. Based on the model that distinguished only between significant and nonsignificant studies, however, the mean effect size was reduced by half. It did not drop below zero, but a reduction this large likely indicates that the data set is not robust to publication bias.

18.6.1.2.11 Vevea and Woods. The results for the irritable bowel syndrome data set are presented in table 18.3. The code we used consists of the same variations that we used for the psychotherapy data. Again, we replaced the weights vector with the corresponding set of weights for each selection pattern. Because weights are fixed, it is irrelevant whether *p*-values actually fall in every interval.

None of the selection bias patterns adjusted the original unadjusted effect size upward. The furthest downward that it is attenuated happens under severe one-tailed selection, where the adjusted effect size reaches 0.11 from its unadjusted 0.42. Despite this reduction, these results are encouraging. Even under the most severe one-tailed bias pattern we created, the average effect size did not become negative or too near zero; there still appears to be a positive effect. This data set does appear to be robust to the effects of publication bias.

18.6.1.2.12 Copas and Shi. We first estimated a random-effects meta-analysis with maximum likelihood:

```
ibs_meta <- metagen(TE = ibs_y, seTE =
sqrt(ibs_v), method.tau = "ML")
```

Then we estimated the Copas and Shi selection model:

```
cop.ibs <- copas(ibs_meta)
plot(cop.ibs)
summary(cop.ibs)
```

The four plots produced by the Copas and Shi (2001) selection model *R* function (Carpenter et al. 2009) are shown in figure 18.18.

The contour plot indicates that the data set is not terribly robust to the effects of selection bias; most of the contour lines are closer together. It also indicates that the model had some difficulty converging, as some of the contour lines curve. The top right estimate, under no selection bias, is about $\log(RR) = 0.40$.

The treatment effect plot shows that, as the probability of publishing the study with the largest standard error decreases, the estimated average effect size decreases as well, moving from about $\log(RR) = 0.40$ under no selection bias to about $\log(RR) = 0.10$ in a situation where studies with the largest standard error are published only about 35 percent of the time.

Finally, the *p*-value plot shows that residual selection bias becomes nonsignificant when the least precise studies are published about 73 percent of the time. This situation includes publication bias, but is far from the extreme of the psychotherapy data set. The Copas and Shi model yields an adjusted estimate in this situation of $\log(RR) = 0.25$, an attenuation of 40 percent (2001; see table 18.3).

18.6.1.2.13 Rücker Limit Meta-Analysis. We estimated the Rücker limit meta-analysis method:

```
ibs_limit <- limitmeta(ibs_meta)
summary(ibs_limit)
```

The results indicate a significant relationship between effect size and standard error. The test for small-study effects was significant, $Q(1) = 14.03$ ($p = .0002$), as was the test for residual heterogeneity, $Q(17) = 48.37$, $p < .0001$.

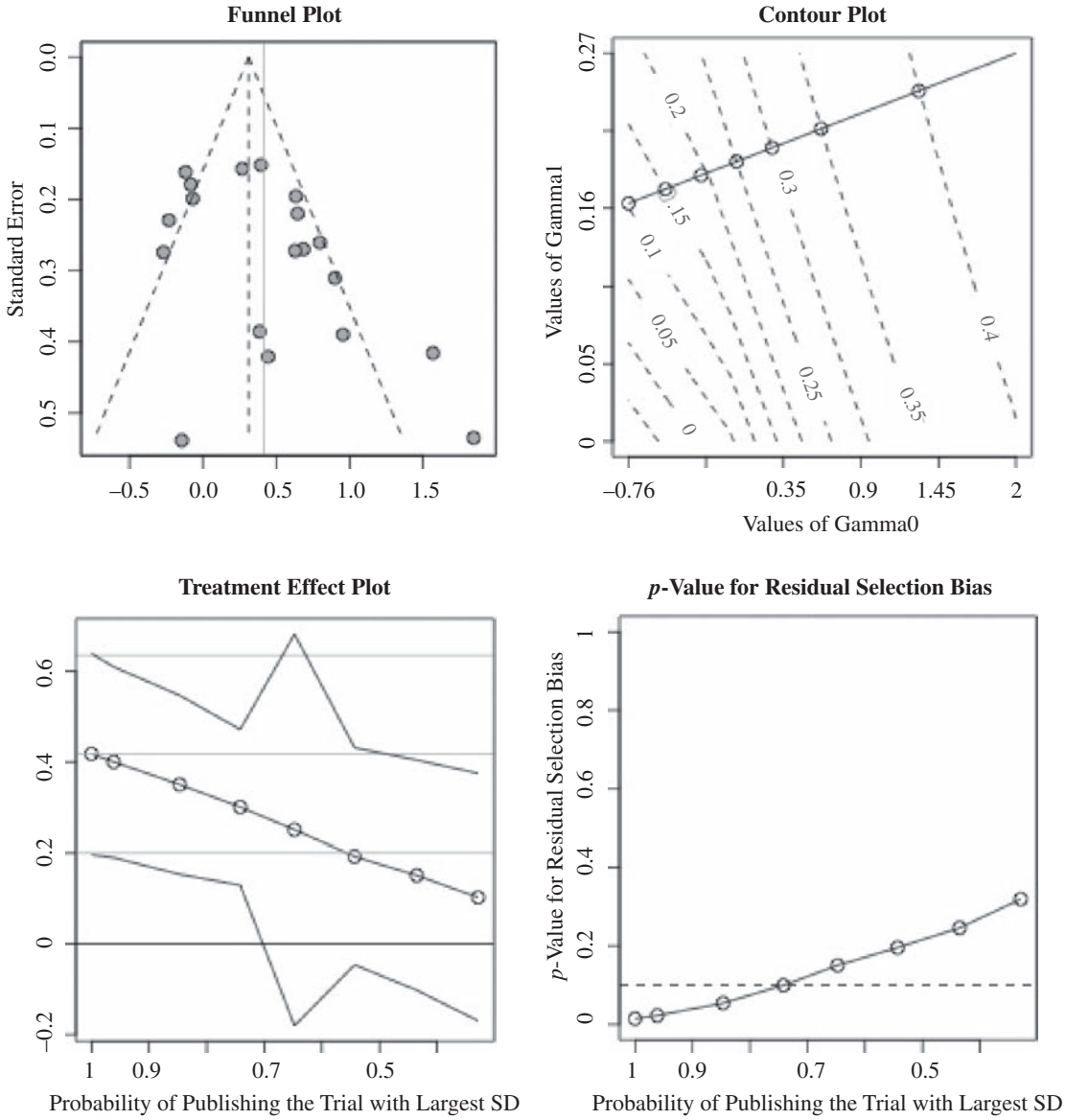

Figure 18.18 Copas and Shi, Irritable Bowel Syndrome Data, $a = -0.76$ to 2, $b = 0$ to 0.27

SOURCE: Author's tabulation.
NOTE: Gamma zero and gamma one represent a and b, respectively.

The adjusted random-effects estimate for the mean effect is 0.09 (see table 18.3), with a 95 percent confidence interval from –0.22 to 0.39. This is a downward adjustment of 78.57 percent.

18.7 DISCUSSION

This chapter provides an overview of the problem of publication bias with a focus on the methods developed to address it in a research-synthesis context. Publication bias is a difficult problem. The mechanisms causing bias are unknown, and the merit of any method to address it depends on the truth of any assumptions the method makes. Therefore, all methods should be viewed as sensitivity analyses, and triangulation across multiple approaches is advisable. Because of the different assumptions, users should not expect triangulation to lead to a consensus across methods and must exercise judgment that includes an assessment of the plausibility of the assumptions. Banks, Kepes, and McDaniel (2012) and Kepes and his colleagues (2012) discuss this issue.

Meta-analysts' toolboxes are now likely to be full of publication bias assessments. The sheer number of possible methods can engender some confusion. As the earlier examples demonstrate, results for a single data set may vary widely across methods, due to the methods' differing assumptions and strategies. Faced with such a range, what should the meta-analyst conclude?

Although all publication bias assessments are sensitivity analyses, some assessments are less sensitive than others. For instance, some methods (such as weight-function methods) are robust to violations of their assumptions; others are not. When conducting a meta-analysis, if the data set does not meet the assumptions of a particular bias assessment, and if that assessment is not robust, the researcher should be less willing to trust its results. On the other hand, if the data set does meet the assumptions, or if the assessment is robust to violations, the researcher should place more credence in its results. This chapter does not provide a specific list of methods that should or should not be included as part of the triangulation process. Instead, the researcher should observe the data set, estimate a range of methods, and assess the body of results, bearing in mind that some results may be more meaningful than others.

Table 18.2 compares the adjusted effect estimates for the psychotherapy data set. One of the more conservative methods (PET-PEESE) gives an adjusted effect size as small as –0.04. At the other end of the spectrum, *p*-uniform adjusts the estimate to 1.06. Other methods (trim and fill and Vevea and Hedges's weight-function model) give estimates that seem more consistent with the funnel plot, ranging from 0.47 to 0.78. The Vevea and Woods model suggests that the data set is robust to the effects of different selection patterns (2005). In no case are the key findings reversed or called into question. The conditional means for complex phobias are attenuated more than simple phobias, due to the smaller magnitude of effects for those groups, which makes them more likely to be affected by weights in the nonsignificant *p*-value ranges. Overall, across all the methods presented in table 18.2, it is plausible that some degree of publication bias is present in this data set. However, with the exception of those methods that can accommodate neither systematic nor random heterogeneity, the key finding is never reversed or called into question.

Table 18.3 compares the adjusted estimates for the irritable bowel syndrome data. PET-PEESE yields a result so extreme that the adjusted finding (–0.23) suggests the treatment is harmful. Once again, *p*-uniform inflates the adjusted effect, likely due to the fact that it disregards nonsignificant effect sizes (in this case, ignoring 50 percent of an already small data set). The Vevea and Woods results yield minimal to moderate adjustment except in the most severe one-tailed scenario (2005). Rücker's method is conservative, producing an estimate similar to the most extreme case of the Vevea and Woods method. The other selection models (Vevea and Hedges, Copas and Shi) reduce the effect dramatically, into the range that Kepes (CITE) define as "severe." It does appear, then, that the true effect may be substantially smaller than estimated in the meta-analysis.

The danger of reliance on a single approach is clear. A responsible analyst here would most likely discount the most extreme results and conclude that, although bias may be a problem, it is not likely to be the primary reason that a positive effect was found. A good sense of what various approaches can and cannot achieve is useful for this triangulation process. Table 18.4 summarizes the characteristics of many methods.

Freely available software has been developed that implements most of the methods described here, with the exception of the Bayesian approaches and methods for outcome, subgroup, and time-lag biases. The new tendency among developers of methods is to make them accessible as open-source packages. *R* (R Core Team 2016) packages implement various models previously inaccessible to typical users. Examples include Wolfgang

Table 18.4 Characteristics of Various Methods

Method	Subjective Interpretation	Tests for Bias	Primarily Visual	Software Available	Linear Model	Adjusted Effect Size(s)	Homogeneity Necessary	Adjusted Variance Component	Based on Relationship Between Study Size and Effect Size	Based on *p*-Values
Funnel plot	X		X	X			X		X	
Cumulative meta-analysis	X		X	X			X		X	
Egger's regression		X		X	X[1]	X	X		X	
Rank correlation		X		X			X		X	
Trim and fill		X		X	X[1]	X	X	X	X	
PET-PEESE		X			X[1]	X	X		X	
Vevea and Hedges		X		X	X	X		X		X
Vevea and Woods				X	X	X		X		X
Dear and Begg	X		X	X						X
Rufibach	X		X	X						X
Copas and Shi			X	X	X[1]	X	X		X	
Limit meta-analysis		X		X	X[1]	X	X		X	
p-curve		X	X	X		X[2]	X			X
p-uniform		X		X		X	X			X
Excess significance test		X					X			X

SOURCE: Author's tabulation.
[1] indicates that the method has the potential to incorporate a linear model, but that software is not readily available to do so.
[2] indicates that the method can yield an adjusted effect size, but software is not readily available to do so.

Viechtbauer's *metafor* (2010), Guido Schwarzer's *meta* package (2016), *metasens* by Guido Schwarzer and his colleagues (2016), and Coburn and Vevea's *weightr* (2016a). Others have made their approaches available through web interfaces. Examples include Coburn and Vevea's Shiny application (2016b), and Uri Simonsohn, Lief Nelson, and Joseph Simmons's web application for *p*-curve (2014).

Future investigation may prove fruitful in several directions. One example is development of methods that simultaneously account for different possible sources of bias (for example, *p*-value as well as magnitude and direction of individual effect estimates). Further development of models that allow various selection patterns for different study designs would be useful (for example, Sutton, Abrams, and Jones 2002). Extension of that idea to account for study characteristics that are not design related (for example, funding source, social preferences, or time) is a developing area (see, for example, Coburn and Vevea 2016b). Dan Jackson points out that little is yet known about the effects of publication bias on the between-studies variance component (2006). Derrick Bennett and his colleagues investigated capture-recapture methods across electronic databases to estimate the number of missing studies, but evidence of further research on that approach is scant (2004). Kepes and McDaniel (2015) mention the need for development of methods that are suitable for psychometric meta-analysis.

Bayesian methods are likely to provide valuable new insights on the publication bias problem. Promising directions could include incorporating Bayesian model averaging or Bayes factors. Bayesian methods also are likely to lead to models that address publication bias for statistical approaches that are more complicated than a standard meta-analysis, such as network meta-analyses.

Publication bias is a pervasive problem in the research literature, and meta-analysis provides a valuable opportunity to assess its impact. This chapter discusses and illustrates a variety of methods that aid in this process. The issue of addressing more nuanced questions about publication is a rapidly developing field, and one that is likely to prove fruitful in the next few years.

18.8 REFERENCES

Abramowitz, Stephen I., Beverly Gomes, and Christine V. Abramowitz. 1975. "Publish or Politic: Referee Bias in Manuscript Review." *Journal of Applied Social Psychology* 5(3): 187–200.

Abrams, Keith R., Clare L. Gillies, and Paul C. Lambert. 2005. "Meta-Analysis of Heterogeneously Reported Trials Assessing Change from Baseline." *Statistics in Medicine* 24(24): 3823–44.

American Psychological Association. 2008. "Reporting Standards for Research in Psychology: Why Do We Need Them? What Might They Be?" *American Psychologist* 63(9): 839–851.

Anderson, Richard. 2013. "Registration and Replication: A Comment." *Political Analysis* 21(1): 38–39.

Balcetis, Emily, and David Dunning. 2012. "A False-Positive Error in Search of Selective Reporting." *i-Perception* 3(3).

Banks, George C., Sven Kepes, and Michael A. McDaniel. 2012. "Publication Bias: A Call for Improved Meta-Analytic Practice in the Organizational Sciences." *International Journal of Selection and Assessment* 20(2): 182–96.

Barnes, Deborah E., and Lisa A. Bero. 1998. "Why Review Articles on the Health Effects of Passive Smoking Reach Different Conclusions." *Journal of the American Medical Association* 279(19): 1566–570.

Bayarri, M. J., and Morris H. DeGroot. 1987. "Bayes Analysis of Selection Models." *Journal of the Royal Statistical Society, Series D (The Statistician)* 36(2/3): 137–46.

Becker, Betsy J. 2005. "Failsafe *N* or File-Drawer Number." In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester: John Wiley & Sons.

Begg, Colin B., and Jesse A. Berlin. 1988. "Publication Bias: A Problem in Interpreting Medical Data (with Discussion)." *Journal of the Royal Statistical Society Series A* 151(3): 419–63.

Begg, Colin B., and Madhuchhanda Mazumdar. 1994. "Operating Characteristics of a Rank Correlation Test for Publication Bias." *Biometrics* 50(4): 1088–101.

Bekelman, Justin E., Yan Li, and Cary P. Gross. 2003. "Scope and Impact of Financial Conflicts of Interest in Biomedical Research: A Systematic Review." *Journal of the American Medical Association* 289(4): 454–65.

Bennett, Derrick A., Nancy K. Latham, Caroline Stretton, and Craig S. Anderson. 2004. "Capture-Recapture Is a Potentially Useful Method for Assessing Publication Bias." *Journal of Clinical Epidemiology* 57(4): 349–57.

Berlin, Jesse A., and Davina Ghersi. 2005. "Preventing Publication Bias: Registries and Prospective Meta-Analysis." In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Berlin, Jesse A., and Robert M. Golub. 2014. "Meta-Analysis as Evidence: Building a Better Pyramid." *Journal of the American Medical Association* 312(6): 603–06.

Bishop, Dorothy V. M., and Paul A. Thompson. 2016. "Problems in Using P-Curve Analysis and Text-Mining to Detect Rate of P-Hacking and Evidential Value." *PeerJ* 4: e1715.

Bondas, Terese, and Elisabeth O. C. Hall. 2016. "Challenges in Approaching Metasynthesis Research." *Qualitative Health Research* 17(1): 113–21.

Borenstein, Michael. 2005. "Software for Publication Bias." In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Bruns, Stephan B., and John P. A. Ioannidis. 2016. "*p*-Curve and *p*-Hacking in Observational Research." *PloS One* 11(2): e0149144.

Carpenter, James R., Guido Schwarzer, Gerta Rücker, and Rita Künstler. 2009. "Empirical Evaluation Showed That the Copas Selection Model Provided a Useful Summary in 80 Percent of Meta-Analyses." *Journal of Clinical Epidemiology* 62(6): 624–31.

Carter, Evan C., Lilly M. Kofler, Daniel E. Forster, and Michael E. McCullough. 2015. "A Series of Meta-Analytic Tests of the Depletion Effect: Self-Control Does Not Seem to Rely on a Limited Resource." *Journal of Experimental Psychology* 144(4): 796–815.

Carter, Evan C., and Michael E. McCullough. 2014. "Publication Bias and the Limited Strength Model of Self-Control: Has the Evidence for Ego Depletion Been Overestimated?" *Frontiers in Psychology* 5: 819.

Ceci, Stephen J., Douglas Peters, and Jonathan Plotkin. 1985. "Human Subjects Review, Personal Values, and the Regulation of Social Science Research." *American Psychologist* 40(9): 994.

Chan, An-Wen, and Douglas G. Altman. 2005. "Identifying Outcome Reporting Bias in Randomised Trials on PubMed: Review of Publications and Survey of Authors." *British Medical Journal* 330(7494): 753.

Chan, An-Wen, Asbjorn Hrobjartsson, Mette T. Haahr, Peter C. Gøtzsche, and Douglas G. Altman. 2004. "Empirical Evidence for Selective Reporting of Outcomes in Randomized Trials. Comparison of Protocols to Published Articles." *Journal of the American Medical Association.* 291(20): 2457–65.

Chan, An-Wen, Karmela Krleza-Jeric, Isabelle Schmid, and Douglas G. Altman. 2004b. "Outcome Reporting Bias in Randomized Trials Funded by the Canadian Institutes of Health Research." *Canadian Medical Association Journal* 171(7): 735–40.

Clarke, Mike, and Lesley Stewart. 1998. "Time Lag Bias in Publishing Clinical Trials." *Journal of the American Medical Association* 279(24): 1952–53.

Coburn, Kathleen, and Jack L. Vevea. 2015. "Publication Bias as a Function of Study Characteristics." *Psychological Methods* 20(3): 310.

———. 2016a. "weightr: Estimating Weight-Function Models for Publication Bias in *R*" (1.0.0). *R* package.

———. 2016b. "The Vevea and Hedges Weight-Function Model for Publication Bias." Computer software. (1.0.0). Accessed December 14, 2018. https://vevealab.shinyapps.io/WeightFunctionModel.

Cooper, Harris M., Kristina M. DeNeve, and Kelly Charlton. 1997. "Finding the Missing Science: The Fate of Studies Submitted for Review by a Human Subjects Committee." *Psychological Method.* 2(4): 447–52.

Copas, John B., and Hu G. Li. 1997. "Inference for Non-Random Samples." *Journal of the Royal Statistical Society, Series B* 59(1): 55–95.

Copas, John, and Jian Qing Shi. 2000. "Meta-Analysis, Funnel Plots and Sensitivity Analysis." *Biostatistics* 1(3): 247–62.

———. 2001. "A Sensitivity Analysis for Publication Bias in Systematic Reviews." *Statistical Methods in Medical Research* 10(4): 251–65.

Coursol, Allan, and Edwin E. Wagner. 1986. "Effect of Positive Findings on Submission and Acceptance Rates: A Note on Meta-Analysis Bias." *Professional Psychology: Research and Practice* 17(2): 136–37.

Dear, Keith B. G., and Colin B. Begg. 1992. "An Approach for Assessing Publication Bias Prior to Performing a Meta-Analysis." *Statistical Science* 7(2): 237–45.

Deeks, Jonathan J., Petra Macaskill, and Les Irwig. 2005. "The Performance of Tests of Publication Bias and Other Sample Size Effects in Systematic Reviews of Diagnostic Test Accuracy Was Assessed." *Journal of Clinical Epidemiology* 58(9): 882–93.

Dickersin, Kay. 2005. "Publication Bias: Recognizing the Problem, Understanding its Origins and Scope, and Preventing Harm." In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Dickersin, Kay, Yi Min, and Curtis L. Meinert. 1991. "The Fate of Controlled Trials Funded by the NIH in 1979." *Controlled Clinical Trials* 12: 634.

———. 1992. "Factors Influencing Publication of Research Results: Follow-Up of Applications Submitted to Two Institutional Review Boards." *Journal of the American Medical Association* 267(3): 374–78.

Dorn, Spencer D., Ted J. Kaptchuk, Jae Berm Park, Long Thanh Nguyen, Katia M. Canenguez, Bong Hyun Nam, Ko Bo Woods, Lisa A. Conboy, William B. Stason, and Anthony J. Lembo. 2007. "A Meta-Analysis of the Placebo Response in Complementary and Alternative Medicine Trials of Irritable Bowel Syndrome." *Neurogastroenterology and Motility* 19(8): 630–37.

Duarte, José L., Jarret T. Crawford, Charlotta Stern, Jonathan Haidt, Lee Jussim, and Philip E. Tetlock. 2015. "Political Diversity Will Improve Social Psychological Science." *Behavioral and Brain Sciences* 38: e130. DOI: 10.1017/S0140525X14000430.

Duval, Sue. 2005. "The Trim and Fill Method." In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Duval, Sue, and Richard Tweedie. 1998. "Practical Estimates of the Effect of Publication Bias in Meta-Analysis." *Australian Epidemiologist.* 5(4): 14–17.

———. 2000a. "A Non-Parametric 'Trim and Fill' Method of Accounting for Publication Bias in Meta-Analysis." *Journal of the American Statistical Association* 95(449): 89–98.

———. 2000b. "Trim and Fill: A Simple Funnel Plot Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis." *Biometrics* 56(2): 455–63.

Easterbrook, Phillipa J., Ramana Gopalan, J. A. Berlin, and David R. Matthews. 1991. "Publication Bias in Clinical Research." *The Lancet* 337(8746): 867–72.

Egger, Matthias, and G. Davey Smith. 1998. "Bias in Location and Selection of Studies." *British Medical Journal* 316(7124): 61–66.

Egger, Matthias, George Davey Smith, Martin Schneider, and Christoph Minder. 1997. "Bias in Meta-Analysis Detected by a Simple, Graphical Test." *British Medical Journal* 315(7109): 629–34.

Ferguson, Christopher J., and Michael T. Brannick. 2012. "Publication Bias in Psychological Science: Prevalence, Methods for Identifying and Controlling, and Implications for the Use of Meta-Analyses." *Psychological Methods* 17(1): 120–28.

Fisher, Ronald A. 1932. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Francis, Gregory. 2012a. "Too Good to Be True: Publication Bias in Two Prominent Studies from Experimental Psychology." *Psychonomic Bulletin & Review* 19(2): 151–56.

———. 2012b. "The Same Old New Look: Publication Bias in a Study of Wishful Seeing." *i-Perception* 3(3): 176–78.

———. 2012c. "Response to Author: Some Clarity About Publication Bias and Wishful Seeing." *i-Perception* 3. DOI: 10.1068/i0519ic.

———. 2012d. "Evidence That Publication Bias Contaminated Studies Relating Social Class and Unethical Behavior." *Proceedings of the National Academy of Sciences* 109(25): E1587.

———. 2012e. "Checking the Counterarguments Confirms That Publication Bias Contaminated Studies Relating Social Class and Unethical Behavior." Accessed December 14, 2018. http://www3.psych.purdue.edu/~gfrancis/Publications/FrancisRebuttal2012.pdf.

———. 2012f. "Publication Bias and the Failure of Replication in Experimental Psychology." *Psychonomic Bulletin & Review* 19(6): 975–91. DOI: 10.3758/s13423-012-0322-y.

———. 2012g. "The Psychology of Replication and Replication in Psychology." *Perspectives on Psychological Science* 7(6): 580–89. DOI: 10.1177/1745691612459520.

———. 2014. "The Frequency of Excess Success for Articles in *Psychological Science*." *Psychonomic Bulletin and Review* 21(5): 1180–87.

Galak, Jeff, and Tom Meyvis. 2012."You Could Have Just Asked: Reply to Francis (2012)." *Perspectives on Psychological Science* 7(6): 595–96.

Galbraith, Rex F. 1994. "Some Applications of Radial Plots." *Journal of the American Statistical Association.* 89(428): 1232–42.

Gelman, Andrew. 2013. "Preregistration of Studies and Mock Reports." *Political Analysis* 21(1): 40–41.

———. 2015. "The Connection Between Varying Treatment Effects and the Crisis of Unreplicable Research: A Bayesian Perspective." *Journal of Management* 41(2): 632–43.

Givens, Geof H., David D. Smith, and Richard L. Tweedie. 1997. "Publication Bias in Meta-Analysis: A Bayesian Data-Augmentation Approach to Account for Issues Exemplified in the Passive Smoking Debate." *Statistical Science* 12(4): 221–50.

Guan, Maime, and Joachim Vandekerckhove. 2016. "A Bayesian Approach to Mitigation of Publication Bias." *Psychonomic Bulletin & Review* 23(1): 74–86.

Hahn, Seokyung, Paula R. Williamson, and Jane L. Hutton. 2002. "Investigation of Within-Study Selective Reporting in Clinical Research: Follow-Up of Applications Submitted

to a Local Research Ethics Committee." *Journal of Evaluation in Clinical Practice* 8(3): 353–59.

Hahn, Seokyung, Paula R. Williamson, Jane L. Hutton, Paul Garner, and E. Victor Flynn. 2000. "Assessing the Potential for Bias in Meta-Analysis Due to Selective Reporting of Subgroup Analyses Within Studies." *Statistics in Medicine* 19(24): 3325–36.

Halpern, Scott D., and Jesse A. Berlin. 2005. "Beyond Conventional Publication Bias: Other Determinants of Data Suppression." In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Harbord, Roger M., Matthias Egger, and Jonathan A. C. Sterne. 2006. "A Modified Test for Small-Study Effects in Meta-Analyses of Controlled Trials with Binary Endpoints." *Statistics in Medicine* 25(20): 3443–57.

Hedges, Larry V. 1984. "Estimation of Effect Size Under Nonrandom Sampling: The Effects of Censoring Studies Yielding Statistically Insignificant Mean Differences." *Journal of Educational Statistics* 9(1): 61–85.

———. 1992. "Modeling Publication Selection Effects in Meta-Analysis." *Statistical Science* 7(2): 246–55.

Hedges, Larry V., and Ingram Olkin. 1985. *Statistical Methods for Meta-Analysis*. San Diego, Calif.: Academic Press.

Hedges, Larry V., and Jack L. Vevea. 1996. "Estimating Effect Size Under Publication Bias: Small Sample Properties and Robustness of a Random Effects Selection Model." *Journal of Educational and Behavioral Statistics* 21(4): 299–332.

———. 2005. "Selection Method Approaches." In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Higgins, Julian P. T., and Sally Green, eds. 2011. *Cochrane Handbook for Systematic Reviews of Interventions*, Version 5.1.0 [updated March 2011]. London: The Cochrane Collaboration. Accessed December 14, 2018. http://handbook-5-1.cochrane.org.

Hopewell, Sally, Michael Clarke, and Sue Mallett. 2005. "Grey Literature and Systematic Reviews." In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Hopewell, Sally, Mike J. Clarke, Lesley Stewart, and Jayne Tierney. 2007. "Time to Publication for Results of Clinical Trials." *Cochrane Database Systematic Reviews* 2 (April): MR000011.

Humphreys, Macartan, Raul Sanchez de la Sierra, and Peter van der Windt. 2013. "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration." *Political Analysis.* 21(1): 1–20.

Hunter, James P., Athanasios Saratzis, Alex J. Sutton, Rebecca H. Boucher, Robert D. Sayers, and Matthew J. Bown. 2014. "In Meta-Analyses of Proportion Studies, Funnel Plots Were Found to Be an Inaccurate Method of Assessing Publication Bias." *Journal of Clinical Epidemiology* 67(8): 897–903.

Hunter, James P., and Frank L. Schmidt. 1990. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings.* Newbury Park, Calif.: Sage Publications.

Hutton, Jane L., and Paula R. Williamson. 2000. "Bias in Meta-Analysis Due to Outcome Variable Selection Within Studies." *Applied Statistics*. 49(3): 359–70.

Ioannidis, John P., Evangelina E. Ntzani, Thomas A. Trikalinos, and Despina G. Contopoulos-Ioannidis. 2001. "Replication Validity of Genetic Association Studies." *Nature Genetics.* 29(3): 306–09.

Ioannidis, John P. A., and Thomas A. Trikalinos. 2007. "An Exploratory Test for an Excess of Significant Findings." *Clinical Trials* 4(3): 245–53.

Iyengar, Satish, and Joel. B. Greenhouse. 1988. "Selection Models and the File Drawer Problem." *Statistical Science* 3(1): 109–35.

Jackson, Daniel. 2006. "The Implication of Publication Bias for Meta-Analysis' Other Parameter." *Statistics in Medicine* 25(17): 2911–21.

Jackson, Daniel, John Copas, and Alexander Sutton. 2005. "Modelling Reporting Bias: The Operative Reporting Rate for Ruptured Abdominal Aortic Aneurysm Repair." *Journal of the Royal Statistical Society Series A* 168(4): 737–52.

Jadad, Alejandro R., and Drummond Rennie. 1998. "The Randomized Controlled Trial Gets a Middle-Aged Checkup." *Journal of the American Medical Association* 279(4): 319–20.

Jennions, Michael D., and Anders P. Moeller. 2002. "Publication Bias in Ecology and Evolution: An Empirical Assessment Using the 'Trim and Fill' Method." *Biological Reviews of the Cambridge Philosophical Society* 77(2): 211–22.

Jensen, Johan Ludwig William Valdemar. 1906. "Sur les fonctions convexes et les inégalités entre les valeurs moyennes." *Acta Mathematica* 30(1): 175–93.

John, Leslie K., George Loewenstein, and Drazen Prelec. 2012. "Measuring the Prevalence of Questionable Research

Practices with Incentives for Truth Telling." *Psychological Science* 23(5): 524–32.

Johnson, Valen E. 2013. "On Biases in Assessing Replicability, Statistical Consistency and Publication Bias." *Journal of Mathematical Psychology* 57(5): 177–79.

Joober, Ridha, Norbert Schmitz, Lawrence Annable, and Patricia Boksa. 2012. "Publication Bias: What Are the Challenges and Can They Be Overcome?" *Journal of Psychiatry and Neuroscience 37*(3): 149–52.

Kepes, Sven, George C. Banks, Michael McDaniel, and Deborah L. Whetzel. 2012. "Publication Bias in the Organizational Sciences." *Organizational Research Methods* 15(4): 624–62.

Kepes, Sven, George C. Banks, and In-Sue Oh. 2014. "Avoiding Bias in Publication Bias Research: The Value of 'Null' Findings." *Journal of Business and Psychology* 29(2): 183–203.

Kepes, Sven, Andrew A. Bennett, and Michael A. McDaniel. 2014. "Evidence-Based Management and the Trustworthiness of Our Cumulative Scientific Knowledge: Implications for Teaching, Research, and Practice." *Academy of Management Learning & Education* 13(3): 446–466.

Kepes, Sven, Brad J. Bushman, and Craig A. Anderson. 2017. "Violent Video Game Effects Remain a Societal Concern: Reply to Hilgard, Engelhardt, and Rouder (2017)." *Psychological Bulletin,* 143(7): 775–82.

Kepes, Sven, and Michael A. McDaniel. 2015. "The Validity of Conscientiousness Is Overestimated in the Prediction of Job Performance." *PLoS ONE* 10(10): e0141468. DOI: 10.1371/journal.pone.0141468.

Klein, Richard A., Kate A. Ratliff, Michelangelo Vianello, Reginald B. Adams Jr., Štěpán Bahník, Michael J. Bernstein, Konrad Bocian, et al. 2014. "Investigating Variation in Replicability: A 'Many Labs' Replication Project." *Social Psychology* 45(3): 107–12.

Koricheva, Julia. 2003. "Non-Significant Results in Ecology: A Burden or a Blessing in Disguise?" *Oikos* 102(2): 397–401.

Lane, David M., and William P. Dunlap. 1978. "Estimating Effect-Size Bias Resulting from Significance Criterion in Editorial Decisions." *British Journal of Mathematical and Statistical Psychology* 31(2): 107–12.

Larose, Daniel T., and Dipak K. Dey. 1998. "Modeling Publication Bias Using Weighted Distributions in a Bayesian Framework." *Computational Statistics & Data Analysis* 26(3): 279–302.

Lau, Joseph, John P. A. Ioannidis, Norma Terrin, Christopher H. Schmid, and Ingram Olkin. 2006. "The Case of the Misleading Funnel Plot." *British Medical Journal.* 333(7568): 597–600.

Lewin, Simon, Claire Glenton, Heather Munthe-Kaas, Benedicte Carlsen, Christopher J. Colvin, Metin Gülmezoglu, Jane Noyes, Andrew Booth, Ruth Garside, and Arash Rashidian. 2015. "Using Qualitative Evidence in Decision Making for Health and Social Interventions: An Approach to Assess Confidence in Findings from Qualitative Evidence Syntheses (GRADE-CERQual)." *PLoS Medicine* 12(10): e1001895.

Liberati, Alessandro, Douglas G. Altman, Jennifer Tetzlaff, Cynthia Mulrow, Peter C. Gøtzsche, John PA Ioannidis, Mike Clarke, P. J. Devereaux, Jos Kleijnen, and David Moher. 2009. "The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration." *PLoS Medicine* 6(7).

Light, Richard J., and David B. Pillemer. 1984. *Summing Up: The Science of Reviewing Research.* Cambridge, Mass.: Harvard University Press.

Little, Roderick J. A., and Don B. Rubin. 1987. *Statistical Analysis with Missing Data.* New York: John Wiley & Sons.

Macaskill, Petra, Stephen D. Walter, and Lesley Irwig. 2001. "A Comparison of Methods to Detect Publication Bias in Meta-Analysis." *Statistics in Medicine* 20(4): 641–54.

Mavridis, Dimitris, Alex Sutton, Andrea Cipriani, and Georgia Salanti. 2012. "A fully Bayesian Application of the Copas Selection Model for Publication Bias Extended to Network Meta-Analysis." *Statistics in Medicine* 32(1): 51–66.

McIntosh, Heather, and Piero Olliaro. 2000. "Artemisinin Derivatives for Treating Severe Malaria." *Cochrane Database Systematic Reviews* 2(2): CD000527.

McShane, Blakeley, and Ulf Böckenholt. 2014. "You Cannot Step into the Same River Twice: When Power Analyses Are Optimistic." *Perspectives on Psychological Science* 9(6): 612–25.

McShane, Blakeley, Ulf Böckenholt, and Karsten Hansen. 2016. "Adjusting for Publication Bias in Meta-Analysis: An Evaluation of Selection Methods and Some Cautionary Notes." *Perspectives on Psychological Science* 11(5): 730–49.

McShane, Blakeley, and David Gal. 2015. "Blinding Us to the Obvious? The Effect of Statistical Training on the Evaluation of Evidence." *Management Science* 62(6): 1707–18.

Monogan, James E. 2013. "A Case for Registering Studies of Political Outcomes: An Application in the 2010 House Elections." *Political Analysis* 21(1): 21–37.

Moreno, Santiago G., Alex J. Sutton, A. E. Ades, Tom D. Stanley, Keith R. Abrams, Jaime L. Peters, and Nicola J. Cooper. 2009. "Assessment of Regression-Based Methods to Adjust for Publication Bias Through a Comprehensive

Simulation Study." *BMC Medical Research Methodology* 9(1): 2.

Morey, Richard D. 2013. "The Consistency Test Does No—and Cannot—Deliver What Is Advertised: A Comment on Francis (2013)." *Journal of Mathematical Psychology* 57(5): 180–83.

Nelson, Nanette, Robert Rosenthal, and Ralph L. Rosnow. 1986. "Interpretation of Significance Levels and Effect Sizes by Psychological Researchers." *American Psychologist* 41(11): 1299–301.

Orwin, Robert G. 1983. "A Fail-Safe N for Effect Size in Meta-Analysis." *Journal of Educational Statistics* 8(2): 157–59.

Peters, Jaime L., Alexander J. Sutton, David R. Jones, Keith R. Abrams, and Lesley Rushton. 2006. "Comparison of Two Methods to Detect Publication Bias in Meta-Analysis." *Journal of the American Medical Association* 295(6): 676–80.

———. 2007. "Performance of the Trim and Fill Method in the Presence of Publication Bias and Between-Study Heterogeneity." *Statistics in Medicine* 26(25): 4544–62.

———. 2008. "Contour-Enhanced Meta-Analysis Funnel Plots Help Distinguish Publication Bias from Other Causes of Asymmetry." *Journal of Clinical Epidemiology* 61(10): 991–96.

Petticrew, Mark, Matt Egan, Hilary Thomson, Val Hamilton, Renée Kunkler, and Helen Roberts. 2006. "Publication Bias in Qualitative Research: What Becomes of Qualitative Research Presented at Conferences?" *British Medical Journal* 62(6): 552–54.

Piff, Paul K., Daniel M. Stancato, Stéphane Côté, Rodolfo Mendoza-Denton, and Dacher Keltner. 2012. "Reply to Francis: Cumulative Power Calculations Are Faulty When Based on Observed Power and a Small Sample of Studies." *Proceedings of the National Academy of Sciences* 109(25): E1588.

Pigott, Therese D. 2001. "Missing Predictors in Models of Effect Size." *Evaluation and the Health Professions* 24(3): 277–307.

Pigott, Therese D., Jeffrey C. Valentine, Joshua R. Polanin, Ryan T. Williams, and Dericka D. Canada. 2013. "Outcome-Reporting Bias in Education Research." *Educational Researcher* 42(8): 424–32.

Preston, Carrol, Deborah Ashby, and Rosalind Smyth. 2004. "Adjusting for Publication Bias: Modelling the Selection Process." *Journal of Evaluation in Clinical Practice* 10(2): 313–22.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Reyes, Magdalena M., Kaitlyn E. Panza, Andrés Martin, and Michael H. Bloch. 2011. "Time-Lag Bias in Trials of Pediatric Antidepressants: A Systematic Review and Meta-Analysis." *Journal of the American Academy of Child & Adolescent Psychiatry* 50(1): 63–72.

Rosenthal, Robert. 1979. "The File Drawer Problem and Tolerance for Null Results." *Psychological Bulletin.* 86(3): 638–41.

Rosenthal, Robert, and John Gaito. 1963. "The Interpretation of Levels of Significance by Psychological Researchers." *Journal of Psychology* 55(1): 33–38.

———. 1964. "Further Evidence for the Cliff Effect in Interpretation of Levels of Significance." *Psychological Reports* 15(2): 570. DOI: 10.2466/pr0.1964.15.2.570.

Rothstein, Hannah R., Alexander J. Sutton, and Michael Borenstein. 2005. "Publication Bias in Meta-Analysis." In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Rücker, Gerta, James R. Carpenter, and Guido Schwarzer. 2011. "Detecting and Adjusting for Small-Study Effects in Meta-Analysis." *Biometrical Journal* 53(2): 351–68.

Rücker, Gerta, Guido Schwarzer, and James R. Carpenter. 2008. "Arcsine Test for Publication Bias in Meta-Analyses with Binary Outcomes." *Statistics in Medicine* 27(5): 746–63.

Rücker, Gerta, Guido Schwarzer, James R. Carpenter, Harald Binder, and Martin Schumacher. 2011. "Treatment-Effect Estimates Adjusted for Small-Study Effects Via a Limit Meta-Analysis." *Biostatistics* 12(1): 122–42.

Rufibach, Kaspar. 2011. "Selection Models with Monotone Weight Functions in Meta Analysis." *Biometrical Journal* 53(4): 689–704.

Schwarzer, Guido. 2016. "meta: General Package for Meta-Analysis" (4.4-0). *R* package.

Schwarzer, Guido, James Carpenter, and Gerta Rücker. 2010. "Empirical Evaluation Suggests Copas Selection Model Preferable to Trim-and-Fill Method for Selection Bias in Meta-Analysis." *Journal of Clinical Epidemiology* 63(3): 282–88.

———. 2016. "metasens: Advanced Statistical Models to Model and Adjust for Bias in Meta-Analysis" (0.3–0). *R* package.

Silliman, Nancy P. 1997. "Hierarchical Selection Models with Applications in Meta-Analysis." *Journal of the American Statistical Association.* 92(429): 926–36.

Simonsohn, Uri. 2012. "It Does Not Follow: Evaluating the One-Off Publication Bias Critiques by Francis (2012a, b,

c, d, e, f).” *Perspectives on Psychological Science* 7(6) 597–99.

———. 2013. “It Really Just Does Not Follow, Comments on.” *Journal of Mathematical Psychology* 57(5): 174–76.

Simonsohn, Uri., Lief D. Nelson, and Joseph P. Simmons. 2014. “*p*-Curve: A Key to the File-Drawer.” *Journal of Experimental Psychology: General* 143(2): 534.

Smith, Richard. 1999. “What Is Publication? A Continuum.” *British Medical Journal* 318(7177): 142.

Smith, Mary Lee, Gene V. Glass, and Thomas I. Miller. 1980. *The Benefits of Psychotherapy*. Baltimore, Md.: Johns Hopkins University Press.

Song, Fujan, Alison Easterwood, Simon Guilbody, Lelia Duley, and Alexander J. Sutton. 2000. “Publication and Other Selection Biases in Systematic Reviews.” *Health Technology Assessment* 4(10): 1–115.

Song, Fujan, Nick Freemantle, Trevor A. Sheldon, Allan House, Paul Watson, Andrew Long. 1993. “Selective Serotonin Reuptake Inhibitors: Meta-Analysis of Efficacy and Acceptability.” *British Medical Journal* 306(6879): 683–87.

Stanley, Tom D. 2005. “Beyond Publication Bias.” *Journal of Economic Surveys* 19(3): 309–45.

Stanley, Tom D., and Hristos Doucouliagos. 2014. “Meta-Regression Approximations to Reduce Publication Selection Bias.” *Research Synthesis Method* 5(1): 60–78.

Stanley, Tom D., Stephen B. Jarrell, and Hristos Doucouliagos. 2010. “Could It Be Better to Discard 90% of the Data? A Statistical Paradox.” *American Statistician* 64(1): 70–77.

Stern, Jerome M., and R. John Simes. 1997. “Publication Bias: Evidence of Delayed Publication in a Cohort Study of Clinical Research Projects.” *British Medical Journal* 315(7109): 640–45.

Sterne, Jonathan A. C., Betsy J. Becker, and Matthias Egger. 2005. “The Funnel Plot.” In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Sterne, Jonathan A. C., and Matthias Egger. 2001. “Funnel Plots for Detecting Bias in Meta-Analysis: Guidelines on Choice of Axis.” *Journal of Clinical Epidemiology* 54(10): 1046–55.

———. 2005. “Regression Methods to Detect Publication and Other Bias in Meta-Analysis.” In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Sterne, Jonathan A. C., Matthias Egger, and George Davey Smith. 2001. “Investigating and Dealing with Publication and Other Biases in Meta-Analysis.” *BMJ: British Medical Journal* 323(7304): 101.

Sterne, Jonathan A. C., David Gavaghan, and Matthias Egger. 2000. “Publication and Related Bias in Meta-Analysis: Power of Statistical Tests and Prevalence in the Literature.” *Journal of Clinical Epidemiology* 53(11): 1119–29.

Sterne, Jonathan A., Alex J. Sutton, John P. Ioannidis, Norma Terrin, David R. Jones, Joseph Lau, James Carpenter, Gerta Rücker, Roger M. Harbord, Christopher H. Schmid, Jennifer Tetzlaff, Jonathan J. Deeks, Jaime Peters, Petra Macaskill, Guido Schwarzer, Sue Duval, Douglas G. Altman, David Moher, and Julian P. T. Higgins. 2011. “Recommendations for Examining and Interpreting Funnel Plot Asymmetry in Meta-Analyses of Randomised Controlled Trials.” *British Medical Journal* 343(7818): 302.

Stouffer, Samuel A., Edward A. Suchman, Leland C. DeVinney, Shirley A. Star, and Robin M. Williams Jr. 1949. *The American Soldier: Adjustment During Army Life*. Studies in Social Psychology in World War II, vol. 1, edited by Samuel Stouffer and Edward A. Suchman. Princeton, N.J.: Princeton University Press.

Sutton, Alexander J. 2005. “Evidence Concerning the Consequences of Publication and Related Biases.” In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Sutton, Alexander J., Keith R. Abrams, and David R. Jones. 2002. “Generalized Synthesis of Evidence and the Threat of Dissemination Bias: The Example of Electronic Fetal Heart Rate Monitoring (EFM).” *Journal of Clinical Epidemiology* 55(10): 1013–24.

Sutton, Alexander J., and Therese D. Pigott. 2004. “Bias in Meta-Analysis Induced by Incompletely Reported Studies.” In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Tang, Jin-Ling, and Joseph L. Y. Liu. 2000. “Misleading Funnel Plot for Detection of Bias in Meta-Analysis.” *Journal of Clinical Epidemiology* 53(5): 477–84.

Terrin, Norma, Christopher H. Schmid, and Joseph Lau. 2005. “In an Empirical Evaluation of the Funnel Plot, Researchers Could Not Visually Identify Publication Bias.” *Journal of Clinical Epidemiology* 58(9): 894–901.

Terrin, Norma, Christopher H. Schmid, Joseph Lau, and Ingram Olkin. 2003. “Adjusting for Publication Bias in the Presence of Heterogeneity.” *Statistics in Medicine* 22(13): 2113–26.

Trikalinos, Thomas A., and John P. A. Ioannidis. 2005. “Assessing the Evolution of Effect Sizes Over Time.”

In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Ulrich, Rolf, and Jeff Miller. 2015. "p-hacking by Post Hoc Selection with Multiple Opportunities: Detectability by Skewness Test?: Comment on Simonsohn, Nelson, and Simmons (2014)." *Journal of Experimental Psychology: General* 144(6): 1137–45.

van Aert, Robbie C. M. 2015. "puniform: Meta-analysis with p-uniform" (0.0.0). *R* package.

van Aert, Robbie C. M., Jelte M. Wicherts, and Marcel A. van Assen. 2016. "Conducting Meta-Analyses Based on *p*-Values: Reservations and Recommendations for Applying *p*-Uniform and *P*-Curve." *Perspectives on Psychological Science* 11(5): 713–29.

van Assen, Marcel A., Robbie C. M. van Aert, and Jelte M. Wicherts. 2015. "Meta-Analysis Using Effect Size Distributions of Only Statistically Significant Studies." *Psychological Methods* 20(3): 293.

Vandekerckhove, Joachim, Maime Guan, and Steven A. Styrcula. 2013. "The Consistency Test May Be Too Weak to Be Useful: Its Systematic Application Would Not Improve Effect Size Estimation in Meta-Analyses." *Journal of Mathematical Psychology* 57(5): 170–73.

van Enst, W. Annefloor, Eleanor Ochodo, Rob J.P.M. Scholten, Lotty Hooft, and Mariska M. Leeflang. 2014. "Investigation of Publication Bias in Meta-Analyses of Diagnostic Test Accuracy: A Meta-Epidemiological Study." *BMC Medical Research Methodology* 14(1): 70–81.

Vevea, Jack L., Nancy C. Clements, and Larry V. Hedges. 1993. "Assessing the Effects of Selection Bias on Validity Data for the General Aptitude Test Battery." *Journal of Applied Psychology* 78(6): 981–87.

Vevea, Jack L., and Larry V. Hedges. 1995. "A General Linear Model for Estimating Effect Size in the Presence of Publication Bias." *Psychometrika* 60(3): 419–35.

Vevea, Jack L., and Carol M. Woods. 2005. "Publication Bias in Research Synthesis: Sensitivity Analysis Using A Priori Weight Functions." *Psychological Methods* 10(4): 428–43.

Viechtbauer, Wolfgang. 2010. "Conducting Meta-Analyses in R with the *metafor* Package." *Journal of Statistical Software* 36(3): 1–48.

Weinhandl, Eric. D., and Sue Duval. 2012. "Generalization of Trim and Fill for Application in Meta-Regression." *Research Synthesis Methods* 3(1): 51–67.

Williamson, Paula R. and Carol Gamble. 2005. "Identification and Impact of Outcome Selection Bias in Meta-Analysis." *Statistics in Medicine* 24(10): 1547–61.

Zarin, Deborah A., Tony Tse, Rebecca J. Williams, Robert M. Califf, and Nicholas C. Ide. 2011. "The ClinicalTrials.gov Results Database—Update and Key Issues." *New England Journal of Medicine* 364(9): 852–60.

PART VII

DATA INTERPRETATION

19

INTERPRETING EFFECT SIZES

JEFFREY C. VALENTINE
University of Louisville

ARIEL M. ALOE
University of Iowa

SANDRA JO WILSON
Vanderbilt University

CONTENTS

19.1 INTRODUCTION

Imagine you are responsible for overseeing the training of medical students. As part of their training, students are exposed to evidence-based medicine. This is simply the idea that clinical judgment and patient values should be used in conjunction with rigorous research evidence to make treatment decisions, but it has important implications, because among other things it means that students need to learn how to read, evaluate, and interpret research studies. For example, in the process of learning about chronic pain management, assume that students read two studies that evaluate the effectiveness of one approach to managing pain. One study found that an intervention improved pain symptoms by 10 points on a measure that is scaled to have a standard deviation of 15 points. Another study reported that the intervention improved pain symptoms by 5 points on a different pain measure that was scaled to have a standard deviation of 10 points. A meta-analysis of the two studies suggested that the overall weighted average effect size for the intervention is d +0.55. Is this effect large enough to have a meaningful impact on patient pain? More to the point from the perspective of a director of clinical training, what skills do your doctors-in-training need in order to help them make this determination?

Michael Borenstein and Larry Hedges introduce effect sizes and how to compute them in chapter 11 of this book. This chapter address how to interpret effect sizes. If you are like most people, knowing that $d = +0.55$ is not terribly helpful (imagine trying to explain what this effect size means to a patient suffering from chronic pain). We therefore start our discussion with an overview of effect sizes and why they are necessary. We then describe two general approaches to interpreting—making meaning of—effect sizes. We call these the descriptive approach and the benchmarking approach. (Ross Crosby, Ronette Kolotkin, and Rhys Williams, 2003, used the terms "distribution-based" and "anchor-based.") We also show how these two approaches can be used together. In describing the approaches, we introduce several effect-size translation metrics. These translations are statistics that express the effect size in different ways, and our hope is that they will help both researchers and their consumers better understand how much of an impact an intervention had, whether you are describing the results of a single study or the results of a meta-analysis. We provide definitional formulas for these translations, and in the appendices provide R code that will allow you to easily compute them with some additional input, such as the results from a meta-analysis (R Core Team 2016). Many of these translations are also available in a separate R package (Del Re 2014). Our primary recommendations are that researchers should express results in terms of the original measure as much as possible, routinely present results in easy to digest tables, provide readers with a suite of effect-size translations to help them understand the effects observed in their studies, and when possible provide external references against which study effects can be compared. We illustrate how to implement each of these recommendations.

19.2 WHAT ARE EFFECT SIZES AND WHY DO THEY NEED TO BE INTERPRETED?

Formally, an effect size is a statistic that expresses the magnitude of a relationship observed in a study. "The effect size was $d = +0.55$," "The correlation was $r = -.10$," and "Self-reported pain symptoms improved by 10 points" are all expressions of the magnitude of the effects observed in a study, and hence, all of these formulations express effect sizes. Occasionally, studies will report outcomes that need little or no additional interpretation. Outcomes such as lived or died, graduated or did not graduate, minutes spent in traffic, hourly wage, and annual salary are outcomes like this. But what if some studies report wages in U.S. dollars and others in Singapore dollars? You will need to convert one form of currency to the other in order to make sense of the results. Similarly, a group of studies on traffic patterns might report minutes spent in traffic, while others will report fractions of an hour spent in traffic. Again, you'll need to convert one formulation to the other to make sense of the results.

With a measure such as a pain scale, things are not so simple. In the example, the two pain scales resulted in measurements with different standard deviations. This implies two really important points. First, just as ten U.S. dollars have different purchasing power than ten Singapore dollars, a ten on one pain scale means something different than a ten on another pain scale. Similarly, the same one-point change on two different pain scales means different things, just like the difference between ten and eleven U.S. dollars is not the same as the difference between ten and eleven Singapore dollars. But pain scales have an additional complication. Most people understand the currency of the country in which they live (for example, someone living in Singapore has a good idea about what one dollar buys). Almost no one has a deep conceptual understanding of what one point on a particular pain scale means. This suggests that consumers of research will benefit from additional interpretations of study effects, and this is the focus of our chapter.

19.2.1 How *Not* to Interpret Effect Sizes

In our introduction, we asked whether the effect size $d = +0.55$ is large enough to be meaningful. Some readers may have been disconcerted by the fact that we did not accompany the effect size with an indication of whether the corresponding analysis resulted in a rejection of the null hypothesis (in other words, that we did not say "$d = +0.55$, $p < .05$" or something like that). In fact, in the history of the social and medical sciences, probability values from null hypothesis significance tests have been the most widely used indication of the magnitude of a study's effect. That is, researchers and research consumers have tended to believe that if $p < .05$ then the effect must be "big" (and "bigger" if $p < .01$, "really big" if $p < .001$) and, conversely, if $p > .05$ that the effect was zero. Unfortunately, these judgments reflect a mistaken belief that stems from the widespread and persistent misconceptions many researchers have about what probability values mean (Cohen 1994). Probability values arising from inferential tests are a function of two independent dimensions: the observed effect size and the sample size used to estimate it. As a result, any nonzero effect size can be statistically significant if the sample size is large enough. A correlation of $r = +.001$ will be statistically significant if the sample has about four million observations (as a very large survey or epidemiological study might). Similarly, an effect size that looks large (like an increase in the high school graduation rate from 30 percent to 60 percent) will not be statistically significant if the sample is small enough. Probability values arising from a null hypothesis significance test should never be used to describe the magnitude of an effect.

After probability values, probably the next most common way that people have used to describe the magnitude of effects observed in a study is to use Jacob Cohen's guidelines for what constitutes a small, medium, or large effect size (1988). Unfortunately, despite what is implied by the descriptors, Cohen never intended these to indicate the importance of effect sizes, nor did he intend them to be applied without regard to context (Cooper 2008). Instead, his interest was in a priori statistical power analysis for sample size planning. To use power analysis for study planning, one needs to be able to reasonably guess the population effect size. To help users with this difficult judgment, Cohen examined studies published in volume 61 of the *Journal of Abnormal and Social Psychology* to get a sense of the magnitude of the effects that researchers might expect to observe (1962). To do this, it was convenient for him to develop working definitions for small, medium, and large effects. This work generated the now familiar thresholds of $d = 0.20$ is a small effect size, $d = 0.50$ is a medium effect size, and $d = 0.80$ is a large effect size (he presents corresponding values for correlation coefficients and odds ratios). The distinction between Cohen's intent (to inform guesses about what effect sizes might be expected) and how his guidelines are used (to describe the importance of effect sizes, without regard to context) is critical, and you will not be surprised to learn that we believe Cohen's rules are generally unhelpful as descriptions of effect size importance. There are two main reasons for this assertion. First, judgments of importance are just that—judgments. Different judges can value different aspects of a decision differently and hence, reach different judgments about whether or not a given effect size is important. A fixed set of rules does not respect these valid differences, and instead imposes a single standard for all judges. Related to this point, judgments about importance are inextricably bound to context. This is perhaps easiest to see with different outcomes. There is a big difference between a 5 percent increase in whether clients attend follow-up visits and a 5 percent increase in survival rates. Yet if the underlying base rates are the same the researcher blindly applying Cohen's rules of thumb will treat these as similarly important. Like probability values arising from a null hypothesis significance test, we believe that Cohen's rules should not be used to describe the magnitude of effects observed in a study.

Now that we have discussed what researchers should *not* to perhaps it is time to start being helpful and provide advice on what we believe researchers *should* do. Cohen's rules are so widely used in part because few have a good idea about what the standardized mean difference means. That is, as we suggested earlier not many people can express how much of a difference an effect size of $d = +0.55$ means. Instead of relying on probability values from null hypothesis tests or Cohen's guidelines, we believe that researchers should consider using translations of effect sizes that we believe have the potential to help explain how much of an effect the intervention under study had.

19.3 DESCRIBING EFFECT SIZES

As a starting point, researchers reporting the results of a single study should always use the original measure as one way to communicate study results. For binary outcomes such as survived or did not survive, presenting the percentages of study participants who experienced each outcome is likely go a long way toward describing study results in an understandable way. For some outcomes scaled continuously, expressing the results in the original measure will be helpful (for example, minutes spent in traffic). Generally, if the scale is likely to be understood by the intended audience, this is the best way to go, and is even preferable to using commonly-reported standardized effect sizes like Cohen's *d*. For example, Susan Carter, Kyle Greenberg, and Michael Walker conducted a study examining the effects of allowing students in college classrooms to access the internet during class (2016). They find that students who were allowed access scored lower ($d = -0.18$) on the final exam relative to students who were not allowed internet access, a statistically significant difference. In this case, the difference in the scores on the final (71 percent versus 73 percent) is a more natural and understandable way of thinking about the magnitude of the experimental effect in this case. That said, as we have seen, many continuously scaled outcomes are like the pain scale example we used—one point on a pain scale does not have an inherent meaning, and as such simply saying that one group scored two points higher than another often will not be terribly helpful.

We should note here that research synthesists face a challenge that researchers reporting the results of a single study do not. That is, it is often the case that in a collection of studies on the same research question, the dependent variable will be operationalized in different ways across studies, even when the construct of interest is the same. Pain is one example. Several measures are commonly used (see, for example, Galer and Jensen 1997; Jensen, Turner, and Romano 1994). All are scaled differently. As another example, in the United States many students who want to attend college must take an entrance examination. The two major tests are the SAT and the ACT, and most colleges and universities will accept either. The SAT typically has a standard deviation of around 117 points, and the ACT of about 5 points. If you are interested in carrying out a research synthesis on the effect of programs that aim to prepare students in the United States to take college entrance exams, some studies in your meta-analytic database will use the ACT and others will use the SAT as the primary outcome, which means that the results are not in the same metric across studies. It is for exactly this reason that the standardized mean difference effect size is so handy. However, if you accept our contention that most people do not understand what a standardized mean difference of, say, $d = -0.10$ means, and that study results should be reported in the original metric, what should a researcher do when the original metric differed across studies?

The answer is that unless there is reason to believe otherwise, it is reasonable to assume operational exchangeability among measurements that share the same underlying conceptual variable. That is, make the assumption that the effect size observed in studies that used one operationalization of the outcome is the same as the effect size in studies that used a different operationalization of the same outcome. Continuing with the SAT-ACT example, this means assuming that the effect size observed for the ACT is the same effect size that would have been observed had the researchers used the SAT instead. Therefore, in this case following our advice would lead researchers to report the results in terms of points on the ACT and in terms of points on the SAT. Extending the example might help illustrate this point. Assume that a good systematic review and meta-analysis examines the effects of coaching on performance on college entrance exams. The overall effect size is $d = +0.08$, a statistically significant result. Because *d* is the effect of the treatment expressed in standard deviation units, we can multiply *d* by the standard deviation to compute the treatment effect expressed on each test's scale. Here, the effect of the intervention is to increase scores on the SAT by $117 \times 0.08 = 9.4$ points, and scores on the ACT by $5 \times 0.08 = 0.4$ points. Assume that we have data suggesting that the mean score on the SAT is about 515, and that

the mean score on the ACT is about 21. Following the logic articulated, we suggest reporting results like this:

> The effect of test coaching on college entrance test scores was $d = +0.08$, $p = 0.02$. Put in context, this effect size implies that the average student's score would improve from 515 to 524.4 on the SAT, and from 21 to 21.4 on the ACT.

Relative to simply stating that the intervention's effect size was $d = +0.08$, we believe that when expressed this way, readers should have a much better chance of understanding what the effect size means in practical terms. We turn now to different translations of effect sizes and show how they can be used to describe effect sizes. We start with outcomes that are scaled continuously, then discuss translations for binary outcomes.

19.3.1 Continuous Outcomes

For continuous outcomes, we assume that the primary meta-analytic results are in the form of a standardized mean difference, *d*. You know by now that we are skeptical that many people have a good feel for how to interpret these effect sizes, and will benefit from additional ways of describing that effect. In the sections below we discuss the correlation coefficient and its square (the proportion of variance explained), the binomial effect-size display, Cohen's *U* metrics, and the common language effect size.

19.3.1.1 Correlation Coefficient Robert Rosenthal argues that because few consumers of research are familiar with standardized mean difference effect sizes but tend to be familiar with correlation coefficients, it is better to express study results in terms of correlation coefficients instead of standardized mean differences (1984). Furthermore, simple randomized experiments can be analyzed using the correlation coefficient as the test statistic. In chapter 11 of this book, Borenstein and Hedges provide a straightforward formula for converting a standardized mean difference to a correlation coefficient.

In our experience, converting the standardized mean difference to a correlation coefficient can be a helpful aid to understanding. But we also think that you should be aware of two potential areas of concern. The first is that for many readers (and some producers) of research, using the correlation coefficient implies that the study did not examine potentially causal relationships. In reality though, causal statements are more closely related to research design than to statistical analysis. Further, many are not aware that the correlation coefficient can be used to analyze the results of a simple randomized experiment. Therefore, if you choose to present results in terms of a correlation coefficient, you should be aware of this common confusion and consider ways of proactively reducing it (for example, by reminding your audience of the tight link between the standardized mean difference and the correlation coefficient).

A second area of concern is that many readers will square the correlation coefficient to yield a proportion of variance explained. This is a valid thing to do and results in an accurate description of the study's results.[1] However, relative to other equally accurate ways of describing study results, the psychological impact of the proportion of variance explained is that it will lead readers to believe that the effect is smaller. For example, "the treatment explained 2 percent of the variance in pain symptoms" and "62 percent of treatment patients scored better than the typical control patient on the pain scale" both describe the same underlying effect size ($d = +0.30$, indicating that patients receiving the treatment scored 0.30 standard deviations better on the pain scale than comparison patients), but the proportion of variance explained formulation feels like a smaller effect size. As a result, the proportion of variance explained should always be presented with another effect-size translation (if it is used at all).

19.3.1.2 Cohen's *U* Metrics In his book *Statistical Power Analysis for the Behavioral Sciences*, Jacob Cohen introduces three effect-size measures based on the extent to which two hypothetical distributions (one treatment and one control) overlap with one another (1988). He somewhat mysteriously called these metrics U_1, U_2, and U_3. U_1 expresses the percentage of population distribution *non*-overlap. That is, imagine that you have two population distributions T (treatment) and C (control). The distributions are normal, are equally large, and have the same variance.

That assumed, to interpret U_1, imagine that the standardized mean difference effect size describing the mean difference between the two distributions is $d = 0.00$. If you superimpose distribution T on distribution C, you will see that the distributions overlap perfectly (that is, 100 percent of distribution T overlaps with distribution C). Stated differently, the percentage of nonoverlap is zero. Now imagine that the standardized mean difference effect size describing the mean difference between the two distributions is $d = +1.00$ (the means are one standard deviation apart). If you superimpose distribution T on distribution C, you will see that the extent of nonoverlap is about half. More precisely, $U_1 = 55.4$ percent if $d = 1.0$ (that is, 44.6 percent of the total area of the two distributions overlaps, and the nonoverlap is 55.4 percent)[2] (see figure 19.1).

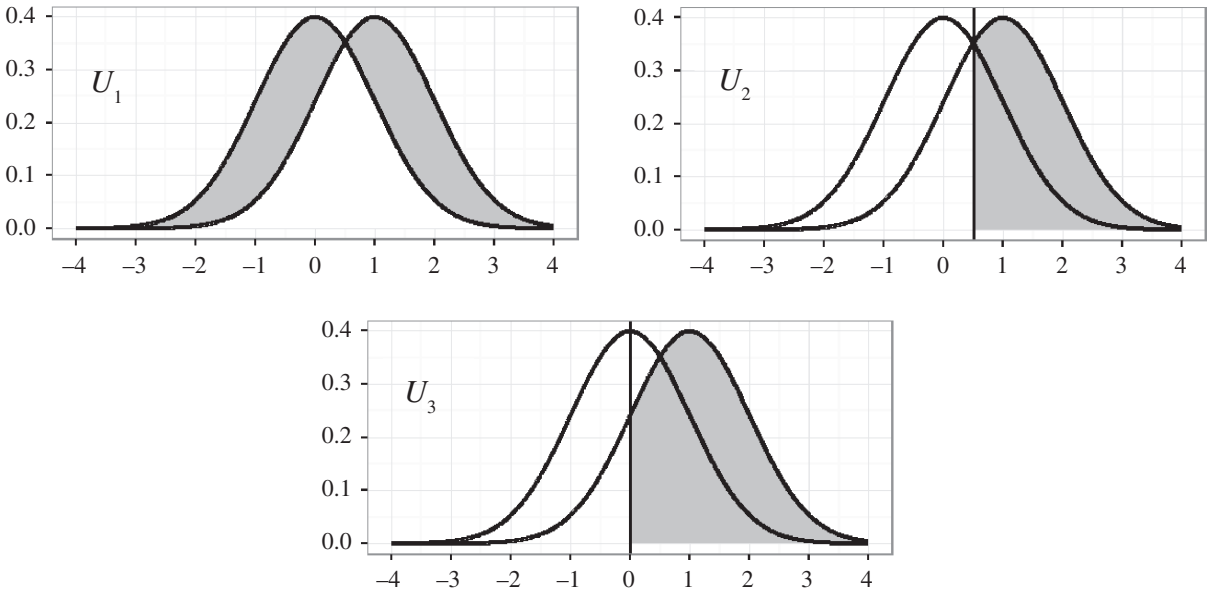

Figure 19.1 Cohen's *U* Metrics

SOURCE: Authors' tabulation.

NOTE: All three graphs illustrate a standardized mean difference effect size (*d*) of +1.0. The x-axis represents the effect size (in standard deviation units) and the y-axis represents probability density.

U_1 (top left) represents the percentage of the total area in the treatment and control distributions that do not overlap with each other (the grey shading illustrates the nonoverlap). For $d = +1.00$, U_1 is 55.4 percent (that is, 55 percent of the area of the two distributions do not overlap). See also figure 19.2.

U_2 (top right) represents the percentage of scores in one distribution that exceed the same percentage of scores in the other distribution. For $d = +1.00$, U_2 is 69.1%: 69 percent of scores in the treatment distribution (the distribution on the right) exceed 69 percent of the scores in the control distribution (on the left). The vertical line in this figure represents the point at which the two distributions intersect. The shaded area represents scores in each distribution that are larger than the value at which the distributions intersect. Here, approximately 69 percent of the scores in the treatment distribution exceed the intersection point. And, approximately 31 percent of the scores in the control distribution exceed the intersection point (hence, approximately 69 percent of the scores in the control distribution do not exceed the intersection point). See also figure 19.2.

U_3 (bottom) represents the proportion of scores in the treatment group (the distribution on the right) that exceed the mean score in the control group (the distribution on the left). This illustrates the $1\text{-}P_{cd}$ formulation of equation (19.1). The control group's mean is represented with a vertical line running down the y-axis at that distribution's apex. Because this figure represents an effect size of $d = +1.0$, and the corresponding U_3 for that effect size is 84.1 percent, we can see that about 84 percent of the scores in the treatment group are larger than the control group's mean.

U_2 expresses the percentage in distribution T that exceeds the same percentage in distribution C. Stated differently, U_2 tells us the percentage of scores in one distribution that exceed the same percentage of scores in the other distribution. If $d = 0.00$ then, as was the case with U_1, the distributions overlap perfectly, and $U_2 = 50$ percent (50 percent of the scores in distribution T exceed 50 percent of scores in distribution C). If $d = +1.0$, then $U_2 = 69.1$ percent. That is, 69.1 percent of scores in distribution T exceed 69.1 percent of scores in distribution C; do not be alarmed if this one is still a little baffling. It might help to look at figure 19.2.

U_3 expresses the percentage of scores in distribution C that are exceeded by the mean of distribution T. Alternatively, and perhaps more intuitively, U_3 can be thought of as expressing the percentage of scores in distribution T that exceed the mean of distribution C. If $d = 0.00$, then 50 percent of the scores in distribution exceed the mean of distribution C (just like 50 percent of the scores in distribution C exceed the mean of distribution T). If

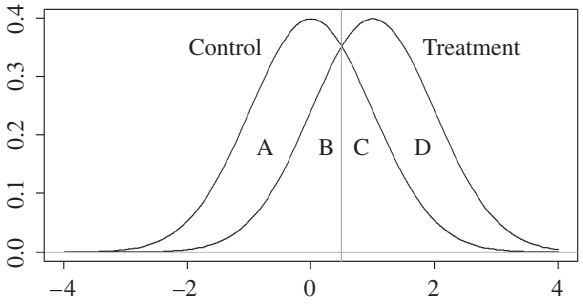

Figure 19.2 The Relationship Between U_1 and U_2.

SOURCE: Authors' tabulation.

NOTE: The letters A, B, C, and D represent different areas in two distributions: a control distribution and a treatment distribution. The area of the control distribution is given by A+B+C and the area of the treatment distribution is given by B+C+D. The letter A represents the proportion of the area of the control distribution that does not overlap with the treatment distribution. B and C represent the proportions of the areas of the two distributions that do overlap. D represents the proportion of the area in the treatment distribution that does not overlap with the control distribution. The vertical line that bisects B and C divides the shared area exactly in half, hence B = C. This property can be inferred from equation (19.2). U_1 (the total area of the two distributions that does not overlap) is defined as (A+D)/(A+B+C+D). U_2 (the percentage of scores in one distribution that exceed the same percentage of scores in the other distribution) is defined as C+D (or equivalently, A+B).

As an example, assume that $d = 1.0$ and that therefore using equation (19.2) and equation (19.3), $U_1 = 0.55$ and $U_2 = 0.69$. This information can be used to determine the relevant area proportions. Because B+C+D = 1 and U_2 = C+D = 0.69 we know that B + 0.69 = 1 and therefore that B = 0.31. Because B = C, C = 0.31, and because B+C+D = 1 we know that D = 0.38 (and also, because A+B+C = 1, and B = 0.31 and C = 0.31, A = 0.38).

To complete the example, U_1 is defined as (A+D)/(A+B+C+D) so U_1 = (0.38+0.38)/(0.38+0.31+0.31+0.38) = 0.55), the same value as that yielded by equation (19.3). U_2 is defined as C+D (or equivalently, A+B) so U_2 = 0.31+0.38 = 0.69, the same value as that yielded by equation (19.2).

d = +1.00 then 84.1 percent of scores in distribution T exceed the mean of distribution C. Out of all of Cohen's U metrics, U_3 is by far the most commonly used. For example, the U.S. Department of Education's What Works Clearinghouse (WWC) routinely reports an "improvement index," which is defined as U_3—50 (so if d = +1.00, then the improvement index is equal to 84.1–50 = 34.1) (What Works Clearinghouse 2017). One way to think about the WWC's improvement index is that it expresses the expected percentile gain if the typical student in distribution C received the intervention. Larry Hedges and Ingram Olkin outline a suite of distribution overlap measures related to U_3, and developed sampling variances for these indices (2016). And similarly, Jeffrey Valentine, Ariel Aloe, and Timothy Lau introduce a variation on U_3 that they call descriptive U_3 (2015). This can be used when one has access to the underlying data.

Figure 19.1 shows each of Cohen's U metrics expressing the same underlying effect size (d = +1.0). In addition, seeing how these effect-size translations are computed might also help you understand what each version is expressing. Like Cohen in 1988, we start with U_3 because it is the easiest to think about. U_3 can be computed using equation (19.1):

$$U_3 = P_d = 1 - P_{-d} \tag{19.1}$$

where P_d is the cumulative distribution function of d. Therefore, P_d represents the percentage of scores in the control distribution that are exceeded by the mean of the treatment distribution, and $1 - P_{-d}$ represents the percentage of scores in the treatment distribution that exceed the mean score in the control distribution. Either formulation yields the same value, so use the version that you think will make the most sense to you and your readers. We used the latter formulation to illustrate U_3 in figure 19.1. Here, if d = +1.0 then 84.1 percent of the scores in the treatment distribution will exceed the mean of the control distribution. Many spreadsheet applications will compute this value. For example, in Google Sheets, the command =normsdist(d)*100 will yield U_3 for the value of d that you choose. In R, the command pnorm(d, 0, 1) can be used to compute U_3 for the value of d that you enter.

U_2 is computed using equation (19.2):

$$U_2 = P_{d/2} \tag{19.2}$$

where all terms are defined as in equation (19.1). In many spreadsheet applications, the command =normsdist(d/2)*100 will yield U_2 for the value of d that you choose. In R, the command pnorm(d/2, 0, 1) can be used to compute U_2.

Finally, U_1 is computed using equation (19.3):

$$U_1 = \frac{2U_2 - 1}{U_2} \tag{19.3}$$

with U_2 defined as in equation (19.2). In many spreadsheet applications, U_1 is easiest to compute by first computing U_2 then computing U_1 as suggested by equation (19.3).[3]

As a reminder, R code for computing all the U transformations is presented in the appendix.

Summary of Cohen's U translations. It is not an accident that U_3 has been the most widely used of Cohen's U translations. Of the three, this translation is the easiest to understand. For example, assume that a psychological intervention results in an improvement of +0.30 standard deviation units in pain symptoms. To use U_3 to communicate to a parent what this effect size means, we would say that about 62 percent of patients who receive the psychological intervention report less pain than the average patient in the control condition. That said, U_1 and U_2 may be easier to understand if they come into common usage.

19.3.1.3 Common Language Effect Size The common language effect size (CLES), or the probability of superiority (McGraw and Wong 1992), represents the probability that a randomly selected observation from one group will be larger than a randomly selected observation from another group. Like Cohen's U_3, the CLES will be 0.50 (50 percent probability) when $d = 0$ and will approach 1.0 as d increases. The CLES can be written using equation (19.4):

$$CLES = P_{d/\sqrt{2}} \quad (19.4)$$

where P_d is defined as in equation 19.1 (the cumulative distribution function for d). In many spreadsheet applications, the CLES can be computed by =normdist((d/sqrt(2)))), where d is the standardized mean difference effect size you are translating. In R, the command pnorm(d/sqrt(2)) can be used to compute the CLES. As an example, assume that a good systematic review and meta-analysis concludes that the effect of a pain treatment is d +0.30 (the pain scores are reverse scaled, so higher numbers represent less pain, and a positively signed effect size means that the intervention patients are doing better than control patients). The CLES for this standardized mean difference is 0.58. The CLES of 0.58 means that the probability that a patient receiving the treatment will score lower than a control patient is 0.58.

19.3.1.4 Binomial Effect-Size Display As we describe more fully further on, binary dependent variables are much easier to explain than continuous dependent variables. When the outcome is continuous, one way to make results easier to understand is to artificially dichotomize it. While doing this has well-known undesirable properties for analysis (see, for example, chapter 15) as a technique for describing and interpreting effect sizes it can be a helpful approach; in a sense this is what Cohen's U_3 does (recall that you can think of U_3 as categorizing treatment group scores as exceeding the control group's mean or not). Rosenthal and Donald Rubin introduce the binomial effect-size display (BESD) as another way to do this (1982). Essentially, the BESD asks "Assume the treatment and control group scores are put together in a single distribution. What proportion of treatment group scores will be above the median? What proportion of the control group's scores will be above the median?" The proportion of the treatment group scoring above the median can be found by using equation (19.5):

$$\text{Treatment proportion above median} = .50 + r/2 \quad (19.5)$$

and the proportion of the control group scoring above the median can be found using equation (19.6):

$$\text{Control proportion above median} = .50 - r/2 \quad (19.6)$$

where r is the point-biserial correlation between the outcome and the treatment condition (scored, for example, as 0 for the control group and 1 for the treatment group). For example, assume that the standardized mean difference effect size for an intervention is d = +0.30. Using the formula for translating d to r given in chapter 11 of this book, we know that if d = +0.30, then r = +.15. Therefore applying equation (19.5), the patients in the treatment condition scoring above the median (that is, experiencing less pain than the average patient) is 0.50 + 0.148/2 = 0.574, or 57.4 percent. Rounding to 57 percent, this implies that 43 percent of treatment patients are scoring below the mean on the pain scale (meaning that they are experiencing more pain than the average patient). Applying equation (19.6) to the control group yields percentages of 43 and 57, and yielding the data for the BESD (see table 19.1).

In a study involving equal sample sizes and a perfectly normal distribution, BESD works out exactly. That is, if you applied the BESD formula to a set of normally distributed observations from a single study in which the treatment and control groups had the same sample size, you would get the same result as would be obtained by dummy coding whether an observation was above or below the median and figuring the percentages that way. But setting aside the clear benefits to communication, the main benefit of the BESD is that it can be used by people, like synthesists, who do not have access to the underlying

Table 19.1 Binomial Effect Size Display for a Continuous Outcome

	% Scoring Above the Median (Less Pain Than Average)	% Scoring Below the Median (More Pain Than Average)
Treatment patients	57	43
Control patients	43	57

SOURCE: Authors' tabulation.
NOTE: Binomial effect size display for a meta-analysis yielding a standardized mean difference effect size of $d = +0.30$. The dependent variable is a pain scale.

data and therefore cannot carry out the dummy coding exercise. We would be remiss not to note that the BESD has its critics (for example, Thompson and Schumacker 1997). However, as Rosenthal suggests, the BESD is reasonably accurate whenever the assumptions of the general linear model are met and as such, results in generally credible translations of the effect size (1991).

19.3.2 Binary Measures of Effect Size

Binary outcomes such as graduated or did not graduate are much easier than continuous outcomes for most people to understand. One reason for this is that the raw data are easily understandable. Consider table 19.2, which shows data for a fictitious study of physical therapy. These data can be readily understood by most people, and that it is pretty clear that physical therapy "worked" in the sense that many more participants who received physical therapy were pain free than participants who did not receive physical therapy. One reason this effect is easy to comprehend is that it is large. Another is that the sample sizes were equal in the physical therapy and non-physical therapy groups. As can be seen in table 19.3, often presenting the results in percentages can help if group sizes are unequal and in general, and this is probably the best way to present results.

Table 19.2 Effects of Physical Therapy for a Binary Outcome

	Pain-Free After Six Months	Not Pain-Free After Six Months
Received physical therapy	100	50
Did not receive physical therapy	60	90

SOURCE: Authors' tabulation.
NOTE: Raw data for a binary outcome. Cell values are the number of participants in each condition-outcome combination.

19.3.2.1 Correlation Coefficient Another option is to express the results as a correlation coefficient (ε). Here, the correlation between assignment condition and being pain free after six months was $r = +0.27$, and its square is 0.07. Recall our earlier concern about proportion of variance explained as an effect size, which can easily be seen here. This effect is pretty dramatic, but it explains only 7 percent of the variance in treatment outcome.

19.3.2.2 Odds Ratio Binary outcomes can also be expressed as odds ratios, which can be defined as

$$OR = \frac{\frac{a}{b}}{\frac{c}{d}} \tag{19.7}$$

where a, b, c, and d refer to cells in a 2×2 table (reading from top left, to top right, to bottom left, to bottom right). Here, the odds ratio is 3.0 (see table 19.2). As can be seen in the equation, the odds ratio is well named—it is comprised of the odds of being pain free in the physical therapy group (the numerator of the fraction), and the odds of being pain free in the non-physical therapy group (the denominator of the fraction). Hence, the odds ratio is literally the ratio of the odds in the treatment and control groups.

Despite the advantage of having a name that is literally true, the odds ratio is difficult to interpret. Most people do

Table 19.3 Effects of Physical Therapy for a Binary Outcome (Based on Percentages)

	Pain-Free After Six Months	Not Pain-Free After Six Months
Received physical therapy	66.7	33.3
Did not receive physical therapy	40	60

SOURCE: Authors' tabulation.
NOTE: Outcomes of a study using the binary outcome from table 19.2, expressed as percentages. The cell varies are the percentages of participants in each condition-outcome combination

not think in terms of odds, and even fewer are comfortable thinking about more than one set of odds at a time. Furthermore, it has the inconvenient form of being centered on 1.0 (that is, when the null hypothesis is true the odds ratio is 1.0). It also ranges from 0 to positive infinity. As a result, an odds ratio of 0.50 is of the same magnitude, but is in the opposite direction, as an odds ratio of 2.0. Due in part to these complexities, analysis are often carried out on the natural log of the odds ratio, and we do not recommend relying too much on the odds ratio if your goal is to make readers understand your results.

19.3.2.3 Risk Ratio The risk ratio is an alternative (and perhaps more understandable) expression of binary study results. A risk ratio is defined as

$$\mathrm{RR} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} \tag{19.8}$$

where a, b, c, and d are defined as in equation (19.7). For the fictitious physical therapy study, the risk ratio is 1.7. Note that the designation of risk and success is completely arbitrary. That is, the risk ratio could be called the success ratio with no changes to its computation. Because the physical therapy intervention was attempting to reduce pain, we will refer it as a success ratio, and interpret it this way: The success rate in the physical therapy group was 1.7 times greater than the success rate in the control group.

19.3.2.4 Risk Difference Another way to express the size of the effect for binary outcomes is the risk (or success) difference, which is defined as

$$\mathrm{RD} = \frac{a}{a+b} - \frac{c}{c+d} \tag{19.9}$$

where a, b, c, and d are defined as in equation (19.7). As can be seen, the success difference is just the difference in success rates between the experimental and control groups. Here, the success difference is 0.27.

19.3.2.5 Number Needed to Treat Finally, the number needed to treat (NNT), describes the number of participants who would have to be in the experimental condition in order to generate one more success—for example, one additional individual is pain free after six months. NNT can be defined as

$$\mathrm{NNT} = \frac{1}{RD} \tag{19.10}$$

where RD is the risk difference. For the physical therapy study, NNT is 1/.27 = 3.75. This implies that for approximately every four individuals who participate in physical therapy, one additional person is pain free after six months, over and above the base rate of being pain free at six months.

Absolute versus relative measures of risk. The risk difference and the number needed to treat are absolute measures of risk; odds ratios and risk ratios are relative measures. The distinction is important for at least two reasons. First, relative measures of risk are largely insensitive to differences in the base rate of events, whereas absolute measures are very sensitive to these differences. In addition, people perceive relative risk and absolute risk differently; relative risk often seems to have a bigger psychological impact (Baron 1997; Covey 2007). As a result, we suggest always presenting absolute and relative measures of risk together along with the original risks (for example, present the risk in the treatment group, the risk in the control group, the risk ratio, and the risk difference).

19.3.2.6 Using Binary Translations in Conjunction with a Meta-Analysis For continuous outcomes, the effect-size translations we discussed earlier can be used without any additional information. For binary outcomes, additional information is needed. The reason for this is that binary effect sizes vary as a function of the underlying base rate. To illustrate this point, assume that the effect of a treatment is to increase survival rates from 45 percent to 55 percent (that is, survival rates are 45 percent in the control group and 55 percent in the treatment group). The odds ratio associated with this treatment effect is 1.5 (the odds of survival were 1.5 times greater in the treatment group than they were in the control group). If instead the survival rates were 75 percent and 85 percent in the control and treatment groups, this same 10 percentage point increase would be associated with an odds ratio of 1.9.

There are multiple ways to arrive at a base rate for the effect-size translations. In our experience, base rates are often available in studies, either in the form of a 2×2 table of frequencies or of reported percentages in the treatment and control groups. One option then is to use this information to estimate a typical base rate (that is, the control group's rate of success), and base the effect-size translations off of that base rate. In the rare case of a study that does not report enough information for you to derive the 2×2 table, the base rate can be based on the studies that do provide that information. However, this means that the effect-size translations will be partly a function of the chosen base rate and should be interpreted in that light. Thus it might be helpful to present the effect-size translations for a reasonable range of base rates so that readers

can get a feel for what the effect size looks like at those different base rates.

19.3.3 Using Effect-Size Translations with Confidence Intervals

So far, we have addressed how to take an effect size from a single study, or from a meta-analysis of multiple studies, and convert that effect size into different metrics as an understanding aid. By providing a concrete expression of the precision of the estimated effect, confidence intervals can also aid understanding (see Cumming 2014; Valentine and Aloe 2016). They can also be used in conjunction with effect-size translations. As most readers will be aware, all common theoretical confidence intervals are formed by multiplying a standard error by a test statistic (for example, in a simple randomized experiment, multiplying the standard error for the difference between two means by the appropriate critical value of t will yield a confidence interval for a mean difference). Although theoretical standard errors have been developed for many of the effect-size translations we provide, they have not been derived for all of them. However, a relatively straightforward strategy that can be easily applied to meta-analytic results is to first compute the lower and upper bounds of the confidence interval using the metric of the meta-analysis (usually the standardized mean difference, the Fisher's z transformed correlation coefficient, or the natural log of the odds ratio), then apply the methods we describe to the lower and upper bounds of the confidence interval.

For example, recall that the meta-analysis of pain interventions described earlier resulted in a mean effect size of $d = +0.55$. Assume that the 95 percent confidence interval for this effect size was +/– 0.50. This implies that the lower bound of the confidence interval was +0.05 and that the upper bound was +1.05. If you were interested in discussing these results in terms of Cohen's U_3 translation, you might report this:

> On average, patients in the treatment condition had better pain scores than patients in the control condition, $d = +0.55$, $p = .03$. The lower bound of the 95 percent confidence interval was $d = +0.05$ and the corresponding upper bound was $d = +1.05$. The U_3 metric expresses study effects in terms of the percentage of one group that exceeds the mean of another group (Cohen 1988). Here, the standardized mean difference of +0.55 suggests that about 71 percent of treatment patients had better pain scores the average control patient, with 95 percent confidence interval ranging from a 52 percent to 85 percent of treatment patients scoring better than the typical control patient.

This can also be done using binary outcomes. For example, assume that a good systematic review and meta-analysis investigates the effects of an intervention designed to encourage employees to participate in a higher retirement savings category. The results suggest an odds ratio of 2.06, with a corresponding log odds ratio of .724 and a standard error of the log odds ratio of 0.154 (for a description of how this standard error is computed, see chapter 13). If you were interested in discussing these results in terms of the risk difference, you might report as follows:

> On average, employees in the treatment condition were more likely to select the higher savings category (OR = 2.06, logOR = 0.724, $SE_{logOR} = 0.154$, $p < .001$). Overall about 12 percent of employees in the control condition opted into the higher savings rate plan, and about 22 percent of treatment employees did, a success (or "risk") difference of 0.10. Assuming a constant success rate in the control group of 12 percent, the effect size translates to a 95 percent confidence interval ranging from a low risk difference of .05 and a high risk difference of 0.16.

19.3.4 Suggestions for Researchers

Our discussion to this point has focused on effect sizes and some translations to help make effect sizes more understandable; we have presented these for both continuous and binary outcomes. One thing that should be clear is that presenting effect-size translations can be done a number of ways. Even though they are all technically accurate descriptions of study results, their psychological impacts can differ. Because there is no single correct effect-size translation (Rosenthal 1991), one option is to present several translations to help readers come away with the most complete description possible. We illustrate how this might be done in tables 19.4, 19.5, and 19.6. Table 19.4 contains effect sizes and translations for a continuous outcome. Assume that a good systematic review and meta-analysis suggests that a physical therapy intervention results in improvement of $d = +0.30$ in pain symptoms. Table 19.5 contains effect sizes and translations for a binary outcome. Assume that a good systematic review and meta-analysis suggests that the logged odds ratio describing the effect of a physical therapy intervention on whether patients were pain free after six months was +1.10. Both tables have columns for the effect size or effect-size translation name, its formal interpretation, and a description of how we would write text that explains the metric to readers. Finally, to give you a feel for how the effect-size translations vary as

Table 19.4 Effect-Size Translations and Interpretation for Continuous Outcomes

Effect Size (or Translation)	Formal Interpretation	Value	Interpretation for Pain Study
Standardized mean difference (d)	Mean difference between treatment and control group, expressed in standard deviation units	+0.30	The physical therapy group scored +0.30 standard deviations better on the pain scale than the control group
Point-biserial correlation coefficient (r_{pb})	Correlation between treatment condition and outcome	+.15	The point-biserial correlation between assignment condition (treatment = 1, control = 0) and pain at the posttest was +0.15.
Proportion of variance explained (r^2_{pb})	Proportion of variance shared between treatment condition and outcome	0.02	The treatment explained 2 percent of the variance in pain symptoms.
U_1	Extent of nonoverlap between the treatment and control distributions	21	Twenty one percent of the area in the distribution of treatment scores does not overlap with the distribution of control scores.
U_3	Percentile rank of the typical participant receiving the treatment, relative to the control group	62	The average treatment participant had a better score on the pain inventory than 62 percent of control participants.
Common language effect size	Probability that a randomly selected member of the treatment group will outscore a randomly selected member of the comparison group	0.58	The probability that a randomly selected member of the treatment group would have a better pain score than a randomly selected member of the control group is 0.58.

SOURCE: Authors' tabulation.
NOTE: Effect sizes and effect-size translations for a study examining the effects of an intervention on continuously scaled pain scores. The underlying effect size is $d = +0.30$.

a function of the underlying effect size, table 19.6 contains effect-size translations for a variety of values of d.

19.4 BENCHMARKING EFFECT SIZES

Although not as widely applicable as using effect-size translations, another strategy for describing effect sizes is to benchmark the observed effect against some external reference value or values. Carolyn Hill and her colleagues describe several types of benchmarks that we will elaborate on: comparing the effect size with normative expectations, policy-relevant goals or gaps, and the effects observed in similar interventions (Hill et al. 2008). Benchmarking requires information that is not necessary for most of the translations we described: at least one reference value. Not all outcomes will have an appropriate or available reference value criterion, but a little creativity and some information about the context of the interventions or outcomes at issue can help.

19.4.1 Comparing Effect Sizes with Norms

In education, many standardized tests have published norms that describe expected scores for students at different grade levels. Hill and her colleagues computed grade-to-grade gain effect sizes from such published norms of standardized tests and present those as benchmarks against which individual study authors or synthesists can evaluate their effect sizes (Hill et al. 2008). The normative gains are those that would be expected in the absence of any intervention, given the context in which the intervention is used. For example, according to Hill and her colleagues (2008) the growth in achievement that typical students achieve between first and second grade in read-

Table 19.5 Effect-Size Translations and Interpretation for Binary Outcomes

Effect Size (or Translation)	Formal Interpretation	Value	Interpretation for Pain Study
Odds ratio	The odds of success in the treatment group, divided by the odds of success in the control group	3.0	The odds of being pain free after six months were three times greater in the treatment group than in the control group
Risk ratio	The ratio of rate of success (or rate of failure) in the treatment group divided by the rate of success (or the rate of failure) in the control group	1.67	The pain free rate was 1.67 times greater in the treatment group than in the control group.
Risk difference	The difference is success (or risk) rates between the treatment and control groups	0.27	The difference in pain free rates between the treatment and control groups was 0.27.
Number needed to treat	The number of participants needing to experience the treatment in order to result in one additional success	4	For every four participants who receive the treatment, we expect one additional participant will report being pain free after six months (rounded from 3.75).
Correlation (phi)	The correlation between condition (here, treatment = 1 and control = 0) and outcome (here, pain free after six months = 1, not pain free after six months = 0)	+0.27	The correlation between assignment and being pain free after six months was $r = +0.27$.
Proportion of variance explained	Proportion of variance shared between treatment condition and outcome	0.07	The treatment explained 7 percent of the variance in whether participants reported being pain free after six months.

SOURCE: Authors' tabulation.
NOTE: Effect sizes and effect-size translations for a study examining the effects of a physical therapy intervention. The dependent variable is whether participants were pain free after six months. The underlying effect size is an odds ratio of 3.0.

ing is about 1.0 standard deviations, whereas that between grades six and seven is about 0.25 standard deviations. A study (or a meta-analysis) of a reading intervention for first graders produces a standardized mean differences effect size of $d = +0.15$. Given that typical growth in the first-grade year is relatively larger than the effect of treatment, we might conclude that the intervention's effect is somewhat modest. If, however, our study is of an intervention for sixth graders, an intervention that produces an effect of $d = +0.15$ might be deemed very effective. In the health sciences, large-scale epidemiologic studies that track changes in health behaviors and outcomes over time could be used in a similar manner.

A variation on this type of benchmarking that can be useful with both individual intervention studies and meta-analyses of intervention studies is to use untreated groups to create benchmarks. For example, in an individual intervention study, the effect size indexing the gains experienced by the comparison group can be a benchmark for the effect size indexing the differences between the groups after treatment to answer questions about whether the observed treatment effect is relatively larger or smaller *than typical growth*.

19.4.2 Comparing Effect Sizes with Policy-Relevant Goals or Gaps

Some outcomes are a focus of public policy, and intervention effects can be described in terms of how much progress might be made toward the public policy goal

Table 19.6. Effect-Size Translation Equivalents

d	*r*	r^2	U_1	U_2	U_3	WWC-II	CLES	BESD
+ 4.00	+.89	80.0%	97.7%	97.7%	99.9%	50	99.8%	94.7%
+ 3.00	+.83	69.2	92.8	93.3	99.9	50	98.3	91.6
+ 2.00	+.71	50.0	81.1	84.1	97.7	48	92.1	85.4
+ 1.50	+.60	36.0	70.7	77.3	93.3	43	85.6	80.0
+ 1.00	+.45	20.0	55.4	69.1	84.1	34	76.0	72.4
+ 0.50	+.24	5.9	33.0	59.9	69.1	19	63.8	62.1
+ 0.40	+.20	3.5	27.4	58.0	65.5	16	61.1	59.8
+ 0.30	+.15	2.2	21.3	56.0	61.8	12	58.4	57.4
+ 0.20	+.10	1.0	14.8	54.0	57.9	8	55.6	55.0
+ 0.10	+.05	0.25	7.7	52.0	54	4	52.8	52.5
0.00	.00	0	0	50	50	0	50	50
–0.10	–.05	0.25	7.7	48.0	46	–4	47.2	47.5
–0.20	–.10	1.0	14.8	46.0	42.1	–8	44.4	45.0
–0.30	–.15	2.2	21.3	44.0	38.2	–12	41.6	42.6
–0.40	–.20	3.5	27.4	42.0	34.5	–16	38.9	40.2
–0.50	–.24	5.9	33.0	40.1	30.9	–19	36.2	37.9
–1.0	–.45	20.0	55.4	30.9	15.9	–34	24.0	27.6
–1.50	–.60	36.0	70.7	22.7	6.7	–43	14.4	20.0
–2.00	–.71	50.0	81.1	15.9	2.3	–48	7.9	14.6
–3.00	–.83	69.2	92.8	6.7	0.1	–50	1.7	8.4
–4.00	–.89	80.0	97.7	2.3	0.1	–50	0.2	5.3

SOURCE: Authors' tabulation.

NOTE: Effect-size translations for a variety of values of *d*. WWC-II is the What Works Clearinghouse's improvement index. The value in the BESD column reflects the percentage of scores in the treatment distribution that score above the control distribution (see equation 19.5) and therefore represents the value that would be in the upper left data cell in a BESD table. Although generally the BESD should be presented in a table (like table 19.1), we present this statistic alone to facilitate comparison with other effect-size translations.

assuming that the intervention is widely adopted. For example, many organizations have guidelines for the number of minutes per week that adults should be physically active. The Canadian Society for Exercise Physiology suggests 150 minutes per week (2012). Assume that on average, Canadian adults get 120 minutes of exercise per week. An intervention that results in participants engaging in fifteen minutes of additional physical activity per week can be said to reduce the gap between actual and goal behavior by 50 percent. Similarly, in the United States educators and policymakers often worry about the achievement gaps between students from economically disadvantaged backgrounds relative to students from more economically viable backgrounds. As an example, according to the nationally representative Education Longitudinal Study of 2002, 35 percent of students classified as low socioeconomic status (SES) obtained a postsecondary degree or certificate within six years of exiting high school, compared with 50 percent of students classified as middle SES. Thus, the gap is 15 percentage points; an intervention that increases the postsecondary attainment rate of low-SES students by 5 percentage points can be said to reduce that gap by one-third. In both of these cases, benchmarking an intervention's effect to a valued public policy goal can be a useful way to contextualize the magnitude of that effect.

19.4.3 Comparing Effect Sizes with Other Similar Interventions

Another benchmarking strategy, though potentially more labor intensive, involves comparing an effect size with those observed in other studies. To do this well, the effect sizes against which you are benchmarking should

come from a high-quality systematic review. As an example, Mark Lipsey and his colleagues were interested in benchmarking effects in the educational research in the United States (Lipsey et al. 2012). They carried out a systematic search and found 181 independent estimates of the effect of an educational intervention on academic achievement generated from randomized experiments. They then categorized the effects according to a number of dimensions, including the level of schooling (elementary, middle, and high school; roughly equivalent to the age of the students) and the type of achievement measure. The type of achievement measure was subdivided into researcher-developed tests, narrow-scope standardized tests, and broad-scope standardized tests. Their findings indicate the most data at the elementary school level, and as a result their conclusions about the different types of tests rest on the firmest ground. Researcher-developed tests tend to yield larger effect sizes ($d = +0.40$) than either narrow-scope standardized tests ($d = +0.25$) or broad-scope standardized tests ($d = +0.08$). Imagine two interventions aimed at elementary school students, one that uses a researcher-developed test as the dependent variable, and one that uses a broad-scope standardized test. The effect size for both interventions is $d = +0.20$. Many researchers would cite Cohen and label both as small effect sizes (1988). The Lipsey work, however, suggests that the intervention assessed with a researcher-developed test is actually about half the magnitude of the typical intervention aimed at elementary school students, and that the intervention assessed using the broad-scope standardized test yielded effects that were more than twice as large as the typical intervention (Lipsey et al. 2012).

19.4.4 Benchmarking Effect Sizes: Summary

We hope we have convinced you that benchmarking effect sizes has the potential to shed light on the practical importance of effect sizes. You will not be able to benchmark in all situations, however. To benchmark, you need well-accepted policy goals, or for gaps, estimates that can be treated in essence as population parameters (either actual population data, or really good, precise estimates—simple cutoffs such as "educationally significant effect size" or "medium effect size" will not do). To compare observed effects with what might be expected by a similar intervention, really good estimates are again needed. Despite these caveats, when the conditions are right, benchmarking has the potential to highlight the relevance of, and add context and nuance to, the interpretation of effect sizes.

19.5 COMBINING THE TRANSLATION AND BENCHMARKING STRATEGIES

We present the translation and benchmarking strategies separately; they can in fact be used together. One natural way to do this is with the BESD. Although the traditional BESD applied to continuous outcomes involves dichotomizing at the median, benchmarks can provide additional options for dichotomizing that can facilitate the interpretation of an effect size. For example, an important benchmark might be the percentage of students who score above some policy-relevant level on a state or national exam. In other situations, outcomes may have clinical thresholds associated with them. For example, scores higher than 70 on the Child Behavior Checklist are commonly used to identify children in need of further diagnostic assessment. Valentine and Aloe show how research synthesists can use these clinical thresholds in conjunction with meta-analytic results to produce a table like table 19.1, except instead of dichotomizing at the median, the dichotomization occurs at a clinically meaningful threshold or thresholds (2016). Using yet another alternative, a synthesist can use a standardized mean difference effect size from a meta-analysis, translate that effect size into an odds ratio (see chapter 11), select a normative proportion from the literature for the comparator, and produce an expected percentage change associated with the treatment effect. For continuous outcomes, estimates of the normative proportion could come from published statistics on the proportion of students in a school district who are proficient in math, when the outcome is continuous scores on math achievement, or the proportion of students in a school who were suspended for aggressive behavior, when the outcome is continuous scores on measures of aggressive behavior (for an example, see Wilson and Lipsey 2007). The most important point is that in many cases, researchers working with continuous outcomes face challenges in communicating their results. Creating easy to digest tables such as the binomial effect-size display can facilitate communication, and benchmarks will also often provide a sensible and useful way of doing so.

19.6 CONCLUSION

In this chapter, we discuss the importance of thoughtful interpretation of effect sizes, present several effect sizes and translations, and show how they might be used. Two

related strategies that we did not discuss are cost effectiveness and benefit-cost analyses. Although we are optimistic about the potential for these types of analyses to contribute to decision making, in our experience most studies are not reported in enough detail to make even basic cost analyses possible at the meta-analytic level. For example, even implementation costs are difficult to estimate from the information presented in most studies (for an example, see Shemilt et al. 2012). We are hopeful that the increasing use of online repositories to accompany journal articles and the increased attention being given to cost issues will improve this situation.

Still, this hints at an important point. We believe that effect-size interpretation can be facilitated by presenting the underlying results in multiple ways. Merely presenting multiple interpretations will often not be enough, however. In most cases, additional contextual information will be needed beyond the statistical results and their translations. We wrote this chapter in part because of our skepticism about the value of the most commonly used effect-size interpretation (Cohen's guidelines). This skepticism stems from our belief that no fixed interpretation of an effect size can be applied regardless of context. This is where strategies such as benchmarking are particularly helpful: they help us bring information about the study's context into our judgments about the meaning of effect sizes. Other dimensions of context will surely be relevant. Even an informal reckoning of the benefits and costs of an intervention is one such consideration. Any positive effect size, no matter how small, might be worthwhile if it is essentially free (that is, requires almost no resources or effort). Another dimension is the acceptability of the intervention, both to those implementing it and to those experiencing it. For an example of this latter point, assume that a school is considering adopting one of two reading curricula. Evidence suggests both have positive effects relative to the existing curriculum. If the effect of the intervention depends on the quality of implementation, and individuals responsible for implementing the curriculum were really excited about one and relatively unenthusiastic about the other, we would likely recommend adoption of the favored curriculum unless the weight of the evidence was pretty strongly against it.

The commonality behind our discussion is that people likely interpret different effect-size translations differently, even when the underlying effect is the same across translations. This likely happens in two ways: two different readers may see the same translation and react to it differently, and the same judge might read two translations of the same underlying effect and react differently to them. Some evidence exists on the latter point (Baron 1997; Covey 2007), and our assertion that the proportion of variance explained leads people to believe that the effect is smaller than other equally valid translations is probably uncontroversial. That said, it turns out that we actually know very little about how people interpret effect sizes and effect-size translations. For example, consider the number needed to treat translation. The following two statements describe the same underlying effect size: treating ten individuals results in one fewer death; treating one hundred individuals results in ten fewer deaths. Because many people have a cognitive heuristic that more is better, we suspect that people reading the effect if it is framed in terms of larger numbers will perceive the effect as being larger than if it is framed in terms of the smaller numbers. Examples of this sort are numerous, yet strikingly little related empirical work has been undertaken. It is for this reason that we suggest that multiple effect sizes and effect-size translations be used. That is, we recommend that researchers present their readers with a suite of relevant effect sizes and effect-size translations, and benchmark these when possible. We also believe that researchers should routinely present results in a form of an easy to digest table (for example, a binomial effect-size display like table 19.1 for continuous outcomes, or a table with the percentages of successes and failures by group for binary outcomes like table 19.3). And finally, we believe that researchers should also present results in the metric of the most original scales used in their analyses. Our hope is that by providing readers with a rich description of the observed effect from multiple perspectives, they will arrive at a fuller understanding of how important that effect might be.

19.7 APPENDIX

19.7.1 R Code to Compute Effect-Size Translations (Continuous Outcomes)

The following code is a function written for R that will take a user-defined standardized mean difference (d) and its standard error, and will compute the effect-size translations discussed in this chapter (see table 19.1 and table 19.4). Users can enter the code into R, and it will generate the translations for $d = +0.30$ (SE $= 0.18$).[4] Users can change these values as desired. The function will also produce effect-size translations at the lower and upper limits of the effect size's confidence interval.

```
#-------------------------
# Transforming Across Effect-Size Metrics
# Valentine, Aloe, Wilson
#-------------------------
# Starting with the standardized mean
   difference (d)
# and its standard error
#-------------------------

 # Read both functions first

 es.trans <- function(d, se){
 # Point-biserial correlation coefficient
 A <- 4 # assumes equal sample sizes
 across groups
 # if groups are not equal in size,
 comment off the line above,
  # uncomment the next lines of code
  (beginning nt, nc, and A) and
  # enter sample sizes for treatment (nt)
  and control (nc) groups
 # nt <- enter treatment group sample size
 here
 # nc <- enter control group sample size
 here
 # A <- ((nt + nc)^2)/ (nt*nc)
   r <- d/sqrt(d^2 + A)
 # Proportion of variance explained
 r2 <- r^2
 # Cohen's u3
 u3 = pnorm(d)*100
 # Cohen's u2
 u2 = pnorm(d/2)*100
 # Cohen's u1
 u1 = (2*(pnorm(abs(d)/2)) - 1) /
 (pnorm(abs(d)/2))* 100
 cles <- pnorm(d/sqrt(2))
  res.es <- c(d = d, r = r, r2 = r2,
  u1 = u1, u2 = u2, u3 = u3, cles = cles)
   # BESD
   a <- (.5 + r/2) * 100 # percent
   treatment above median
   b <- (1 - a/100) * 100 # percent
   treatment below median
   c <- (.5 - r/2) * 100 # percent control
   above median
   d <- (1 - c/100) * 100 # percent
   control below median
    # 2 by 2 table for BESD
    mytab <- matrix(c(a,b,c,d), ncol = 2,
    byrow = TRUE)
      colnames(mytab) <- c("% Above the
      Median", "% %Below the Median")
      rownames(mytab) <- c("Treatment",
      "Control")
      res_mytab <- as.data.frame(round
      (mytab, 2))
      res_mytab <- as.data.frame(round
      (mytab, 2))
    res.all <- list(res.es, res_mytab)
 return(res.all)
 }

 es.com <- function(d, se, conf_level =
 .95, dist_ci = 'qnorm', . . . ){
 args <- list(p = (1 - conf_level)/2,
 lower.tail = FALSE, . . . )
 value <- do.call(eval(parse(text =
 dist_ci)), args)
 d.ci <- d + c(-1, 1) * value*se
 Estimate <- round(es.trans(d = d,
 se = se)[[1]],2)
 Lower <- round(es.trans(d = d.ci[1],
 se = se)[[1]],2)
   Lower['u1'] <- ifelse(Lower['u1'] < 0,
   0, Lower['u1'])
 Upper <- round(es.trans(d = d.ci[2],
 se = se)[[1]],2)
   Upper['u1'] <- ifelse(Upper['u1'] > 100,
   100, Upper['u1'])
 Names.es <- c("d", "r", "r^2", "U1",
 "U2", "U3", "CLES")
 res <- data.frame(cbind(Names.es,
 Estimate, Lower, Upper))
   l <- paste0(conf_level*100,"%")
   lower <- paste(l, "CI", "Lower")
   upper <- paste(l, "CI", "Upper")
  names(res) <- c("Effect Size",
  "Estimate", lower, upper)
   rownames(res) <- NULL
return(list(res, es.trans(d = d, se = se)
[[2]]))
 }

 # Enter the mean effect size and its
 standard error here
 d <- .30 # standardized mean difference (d)
 se <- .18 # standard error of d
```

```
es.com(d = d, se = se, conf_level = .95)
# Note that the desired confidence level
can be changed
# Note that users can adopt a different
distribution but if so
# other arguments for specific distribution
must be specified.
# In the example below the t distribution
is used
# thus the degrees of freedom are
included
es.com(d = d, se = se, dist_ci = 'qt',
df = 9, conf_level = .99)
```

19.7.2 R Code to Compute Effect-Size Translations (Binary Outcomes)

The following code is a function written for R that will take a user-defined log odds ratio and its standard error, and will compute the effect-size translations discussed in this chapter (see table 19.3 and table 19.5). Users can enter the code into R, and it will generate the translations for log OR = 0.6466 (se = 0.12) and assuming a base rate of 40 percent. Users can change these values as desired. The function will also produce effect-size translations at the lower and upper limits of the effect size's confidence interval.

```
#--------------------------
# Transforming Across Categorical Metrics
# Valentine, Aloe, Wilson
#--------------------------
# Starting with the natural log odds
ratio (LOR),
# its standard error (seLOR)
# and the success rate for control
group (SRC)
#--------------------------

# Read both functions first

es.cat <- function(LOR, seLOR, SRC){
 OR <- exp(LOR) # transform ln(OR) to OR
 c <- SRC # enter the base rate of success
 (or events) in the control group (for
 example, 10, 50, 70)
 d <- 100-c # determines the percentage of
 failures (or non-events) in the control
 group
OddsC <- c/d
OddsT <- OddsC * OR # computes the odds
in the treatment group (a/b)
b = 100/(OddsT+1) # computes the percent-
age of failures (or non-events) in the
treatment group
a = 100-b # computes the percentages of
successes (or events) in the treatment
group
# Risk Ratio
 RR = (a/(a+b))/(c/(c+d))
# Risk Difference
 RD = (a/(a+b))-(c/(c+d))
# Number Needed to Treat
 NNT <- 1/RD
# Correlation coefficient (phi)
 r <- ((a*d)-(b*c))/
 sqrt((a+b)*(c+d)*(a+c)*(b+d))
# Proportion of variance explained
 r2 <- r^2
    res.es <- c(LOR = LOR, OR = OR,
    RR = RR, RD = RD, NNT = NNT, r = r,
    r2 = r2)
    return(res.es)
}

es.com <- function(LOR, seLOR, SRC , conf_
level = .95, dist_ci = 'qnorm', . . . ){
args <- list(p = (1 - conf_level)/2,
lower.tail = FALSE, . . . )
value <- do.call(eval(parse(text =
dist_ci)), args)
LOR.ci <- LOR + c(-1, 1) * value*seLOR
Estimate <- round(es.cat(LOR = LOR,
seLOR = seLOR, SRC = SRC),2)
Lower <- round(es.cat(LOR = LOR.ci[1],
seLOR= seLOR, SRC =SRC),2)
Upper <- round(es.cat(LOR = LOR.ci[2],
seLOR = seLOR, SRC = SRC),2)
Names.es <- c("lnOR", "Odds Ratio",
"Risk Ratio", "Risk Difference",
"Number Needed to Treat", "r", "r^2")
res <- data.frame(cbind(Names.es,
Estimate, Lower, Upper))
  l <- paste0(conf_level*100,"%")
  ll <- paste(l, "CI", "Lower")
  uu <- paste(l, "CI", "Upper")
 names(res) <- c("Effect Size",
 "Estimate", ll, uu)
```

```
rownames(res) <- NULL
return(res)
}

# Enter the mean effect size and its
standard error here
LOR <- 1.10 # enter the meta-analytic log
odds ratio
seLOR <- .12 # enter the standard error
of the meta-analytic log odds ratio
SRC <- 40 # enter the base rate of
success (or events) in the control
group (for example, 10, 50, 70)

es.com(LOR = LOR, seLOR = seLOR, SRC = SRC)

# Note that the desired confidence level
can be changed
# Note that users can adopt a different
distribution but if so
# other arguments for specific distribution
must be specified.
# In the example below the t distribution
is used
# thus the degrees of freedom are included
es.com(LOR = LOR, seLOR = seLOR, dist_ci =
'qt', SRC = SRC, df = 9, conf_level = .99)
```

19.8 NOTES

1. Note that in a meta-analysis, this assertion is not true. Some of the unexplained variance is random sampling error and is therefore unexplainable, and the proportion of variance explained understates the "true" variance explained (see Aloe, Becker, and Pigott 2010).
2. It is probably best to think of U_1 as a measure of nonoverlap than as a measure of overlap (, $1\text{-}U_1$), even though both are possible and overlap may seem a bit more intuitive. The reason to prefer non-overlap is that as the underlying effect size increases, non-overlap also increases (if expressed as overlap, the opposite will be true).
3. If d is negative, compute U_1 by first computing U_2 using the absolute value of d, then use that result to compute U_1.
4. Users can download the code to copy and paste into R at http://www.russellsage.org/publications/handbook-research-synthesis-and-meta-analysis.

19.9 REFERENCES

Aloe, Ariel M., Betsy J. Becker, and Therese D. Pigott. 2010. "An Alternative to R2 for Assessing Linear Models of Effect Size." *Research Synthesis Methods* 1(3–4): 272–83.

Baron, Jonathan. 1997. "Confusion of Relative and Absolute Risk in Valuation." *Journal of Risk and Uncertainty* 14(3): 301–309.

Canadian Society for Exercise Physiology. 2012. "Canadian Physical Activity, and Sedentary Behaviour Guidelines." Accessed December 13, 2018. http://www.csep.ca/guidelines.

Carter, Susan Payne, Kyle A. Greenberg, and Michael S. Walker. 2016. "The Impact of Computer Usage on Academic Performance: Evidence from a Randomized Trial at the United States Military Academy." SEI discussion paper no. 2016.02. Cambridge, Mass.: Massachusetts Institute of Technology. Accessed December 13, 2018. https://seii.mit.edu/wp-content/uploads/2016/05/SEII-Discussion-Paper-2016.02-Payne-Carter-Greenberg-and-Walker-2.pdf.

Cohen, Jacob. 1962. "The Statistical Power of Abnormal-Social Psychological Research: A Review." *Journal of Abnormal and Social Psychology* 65(3): 145.

———. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, N.J.: Lawrence Erlbaum.

———. 1994. "The Earth Is Round (p<.05)." *American Psychologist* 49(12): 997–1003.

Cooper, Harris M. 2008. "The Search for Meaningful Ways to Express the Effects of Interventions." *Child Development Perspectives,* 2(3): 181–86.

Covey, Judith. 2007. "A Meta-Analysis of the Effects of Presenting Treatment Benefits in Different Formats." *Medical Decision Making* 27(5): 638–54.

Crosby, Ross D., Ronette L. Kolotkin, and G. Rhys Williams. 2003. "Defining Clinically Meaningful Change in Health-Related Quality of Life." *Journal of Clinical Epidemiology* 56(5): 395–407. DOI: 10.1016/S0895–4356(03)00044–1.

Cumming, Geoff. 2014. "The New Statistics: Why and How." *Psychological Science* 25(1): 7–29.

Del Re, A. C. 2014. "compute.es: Compute Effect Sizes." R package version 0.2–4. Accessed December 13, 2018. http://cran.r-project.org/web/packages/compute.es.

Galer, Bradley S., and Mark P. Jensen. 1997. "Development and Preliminary Validation of a Pain Measure Specific to Neuropathic Pain: The Neuropathic Pain Scale." *Neurology* 48(2): 332–38.

Hedges, Larry V., and Ingram Olkin. 2016. "Overlap Between Treatment and Control Distributions as an Effect Size Measure in Experiments." *Psychological Methods* 21(1): 61–68.

Hill, Carolyn J., Howard S. Bloom, Alison R. Black, and Mark W. Lipsey. 2008. "Empirical Benchmarks for Interpreting Effect Sizes in Research." *Child Development Perspectives* 2(3): 172–77.

Jensen, Mark P., Judith A. Turner, and Joan M. Romano. 1994. "What Is the Maximum Number of Levels Needed in Pain Intensity Measurement?" *Pain* 58(3): 387–92.

Lipsey, Mark W., Kelly Puzio, Cathy Yun, Michael A. Hebert, Kasia Steinka-Fry, Mikel W. Cole, Megan Roberts, Karen S. Anthony, and Matthew D. Busick. 2012. "Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms." NCSER 2013–3000. Washington: U.S. Department of Education.

McGraw, Kenneth O., and S. P. Wong. 1992. "A Common Language Effect Size Statistic." *Psychological Bulletin* 111(2): 361–65.

R Core Team. 2016. "The R Project for Statistical Computing." R Foundation for Statistical Computing. Accessed December 13, 2018. https://www.R-project.org.

Rosenthal, Robert. 1984. *Meta-Analytic Procedures for Social Research.* Beverly Hills, Calif.: Sage Publications.

———. 1991. "Effect Sizes: Pearson's Correlation, Its Display via the BESD, and Alternative Indices." *American Psychologist* 46(10): 1086–87.

Rosenthal, Robert, and Donald B. Rubin. 1982. "A Simple, General Purpose Display of Magnitude of Experimental Effect." *Journal of Educational Psychology* 74(2): 166–69.

Shemilt, Ian, Jeffrey C. Valentine, Patrick Pössel, Miranda Mugford, and Don T. Wooldridge. 2012. "Costing Program Implementation Using Systematic Reviews: Interventions for the Prevention of Adolescent Depression." *Research Synthesis Methods* 3(3): 191–201.

Thompson, Kenneth N., and Randall E. Schumacker. 1997. "An Evaluation of Rosenthal and Rubin's Binomial Effect Size Display." *Journal of Educational and Behavioral Statistics* 22(1): 109–17.

Valentine, Jeffrey C., and Ariel M. Aloe. 2016. "How to Communicate Effect Sizes for Continuous Outcomes: A Review of Existing Options and Introducing a New Metric." *Journal of Clinical Epidemiology* 72 (April): 84–89.

Valentine, Jeffrey C., Ariel M. Aloe, and Timothy S. Lau. 2015. "Life After NHST: How to Describe Your Data Without 'p-ing' Everywhere." *Basic and Applied Social Psychology,* 37(5): 260–73.

What Works Clearinghouse. 2017. *Standards Handbook, Version 4.0*. Washington: Institute of Education Sciences. Accessed January 3, 2019. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf.

Wilson, Sandra Jo, and Mark W. Lipsey. 2007. "School-Based Interventions for Aggressive and Disruptive Behavior: Update of a Meta-Analysis." *American Journal of Preventive Medicine,* 33(2S): S130–S143.

20

HETEROGENEITY IN META-ANALYSIS

MICHAEL BORENSTEIN
Biostat, Inc.

CONTENTS

20.1 INTRODUCTION

When we speak about heterogeneity in a meta-analysis, we usually refer to the extent to which the effect size varies from one population to the next. If the meta-analysis assesses the impact of a treatment, a small amount of heterogeneity tells us that the treatment's impact will be reasonably consistent across populations, whereas a large amount of heterogeneity tells us that the impact will be substantially larger in some populations than in others. This, along with the mean effect size, speaks to the potential utility of the treatment.

Unfortunately, this simple fact tends to be lost in reports of heterogeneity that often focus on the wrong questions and then address these with the wrong statistics. Researchers ask whether there is *any* heterogeneity, rather than asking *how much* heterogeneity there is. They use the I^2 statistic to tell us how much the effects vary, when in fact I^2 does not provide this information. They classify heterogeneity as being *small*, *moderate*, or *large*, when these classifications are meaningless in the absence of additional context. They suggest that heterogeneity may hurt the quality of the evidence, when in fact it can be a core strength of the analysis.

My goal in this chapter is to provide clarity. I explain what we mean by heterogeneity and what role heterogeneity plays in a meta-analysis. I introduce the various statistics that quantify specific aspects of heterogeneity and explain the unique role of each. With this as background, I offer a template for reporting heterogeneity. Most meta-analyses report Q, df, p, I^2, and T^2 as indices of heterogeneity, but these provide little (if any) information about how the effects are dispersed. By contrast, if we compute the prediction interval and report (for example) that "the effect size varies from as little as 0.30 in some populations, to as much as 0.70 in others," we have provided the information that readers need, in a clear and concise format.

20.2 OVERVIEW

A meta-analysis is a synthesis, and the goal of a meta-analysis is not simply to report the mean effect size, but instead to understand the pattern of effects. The studies included in the meta-analysis may vary in any number of ways, including the populations, the specific variant of the intervention, and the scale used to assess outcomes, among others. The first step in the analysis is to determine how the effect size varies across the array of studies. If the effect size is reasonably consistent, we would focus on the mean and report that the effect size is consistent despite variation in the populations and methods. On the other hand, if the effect size varies in substantive ways, we would shift our focus to the variation in effects and report, for example, that the intervention increases the mean score by 10 points in some populations and by 90 points in others. Finally, we might want to explain some of the variation—for example, to report that the effect size was higher in studies that enrolled older patients, or in studies that employed a more intensive variant of the intervention. When we look at heterogeneity in this way, it enables us to explore important questions that we could not address if the studies were essentially replicates of each other. In that sense, the heterogeneity becomes a key strength, rather than a potential problem, of the meta-analysis (Althuis, Weed, and Frankenfeld 2014; Berlin 1995; Higgins, Thompson, and Spiegelhalter 2009; Lau, Ioannidis, and Schmid 1998; Sutton and Higgins 2008; Thompson 1994).

20.3 HETEROGENEITY AND THE STATISTICAL MODEL

Most meta-analyses are based on either the fixed-effect or the random-effects model. The fixed-effect model (sometimes called the common-effect model) is appropriate when all studies in the meta-analysis are estimating

the same parameter. Operationally, this means that all studies sample from the same population and are identical in all material respects. The random-effects model is appropriate in all other cases (Borenstein et al. 2010; Hedges and Vevea 1998; Higgins, Thompson, and Spiegelhalter 2009; Nikolakopoulou, Mavridis, and Salanti 2014).

The sampling frame that calls for the fixed-effect model is relatively rare. The overwhelming majority of meta-analyses are based on studies performed with nonidentical populations, and therefore the random-effects model applies. This chapter addresses the issue of heterogeneity in these meta-analyses.

20.4 CONFIDENCE INTERVALS AND PREDICTION INTERVALS

Consider a meta-analysis of studies that assess the impact of tutoring on student scores. In each study, students are randomly assigned to be tutored after school, or to a control condition. The effect size is the difference in mean scores between the two conditions. Based on the meta-analysis, we report that the mean effect size is 50 points with a standard error of 2.5 points and a standard deviation of 10 points.

The standard error is an index of precision, and it tells us how precisely we have been able to estimate the mean effect size. The mean plus or minus 1.96 standard error yields a confidence interval of approximately 45 to 55. If the confidence interval is accurate, then in 95 percent of all analyses, the confidence interval will include the true mean. The mean tells us how well the intervention is performing on average, and the confidence interval tells us how precisely we know the mean. But these statistics tell us nothing about how widely the effect size varies across populations. For that, we turn to the standard deviation and prediction interval.

The standard deviation is an index of dispersion, and it tells us how widely the effect sizes are distributed. The mean plus or minus 1.96 standard deviations yields a prediction interval of approximately 30 to 70. If the prediction interval is accurate, then some 95 percent of populations will have an effect size in this interval. The interval is called a prediction interval because if we were asked to predict the impact of the intervention for a randomly selected population, we would predict that the effect size would fall between 30 and 70, and we would be correct some 95 percent of the time.

Consider three separate meta-analyses, for three separate interventions. In each case, the mean effect size is 50. In one case, the standard deviation is 5 points and the prediction interval extends from 40 to 60—the treatment has approximately the same impact in all populations. In the second case, the standard deviation is ten points and the prediction interval extends from 30 to 70—there are some populations where the effect is weak, some where it is moderate, and some where it is very strong. In the third case, the standard deviation is 20 points and the prediction interval extends from to 10 to 90: in some populations, the impact is trivial, in some moderate, and in some exceptional. These three cases are very different from each other, and the prediction interval is what captures this difference.

20.4.1 Confidence Interval Versus Prediction Interval

Because this is a chapter on heterogeneity, I talk primarily about the prediction interval. The reason I discuss the confidence interval as well is that researchers sometimes confuse the two, and I want to draw a clear distinction between them.

Figure 20.1 presents a fictional set of studies for the meta-analysis to assess the impact of tutoring. At the bottom of the plot are two diamonds. The first shows the confidence interval for the fixed-effect model, the second for the random-effects model. The first diamond has a width of 7.5 points, the second of 17.9 points. Researchers sometimes assume that the span for the random-

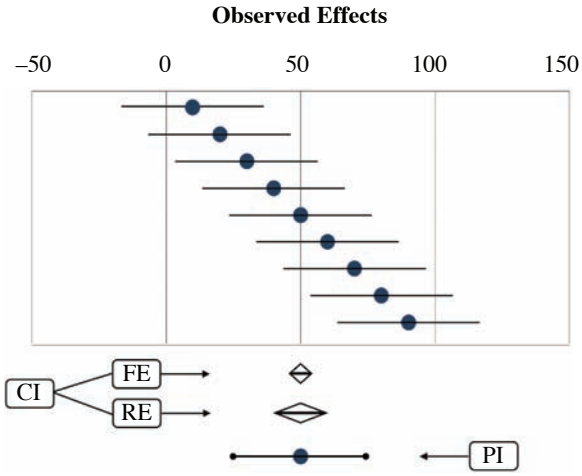

Figure 20.1 Confidence Intervals and Prediction Interval for a Fictional Meta-Analysis

SOURCE: Author's tabulation.

effects model tells us that the effects are dispersed over this (wider) range. This is incorrect—both diamonds speak only to the precision of the estimate for the mean.

The confidence interval labeled FE is based on the standard error for the fixed-effect model. If all studies are sampled from the same population, and we are generalizing from these samples to this one population, then this confidence interval will usually include the true effect size in this population.

The confidence interval labeled RE is based on the standard error for the random-effects model. If the studies are sampled from different populations, and we are generalizing to the universe of similar populations, then this confidence interval will usually include the true mean effect size in this universe.

Critically, the second diamond is wider than the first because it includes an additional source of sampling error. Under the fixed-effect model, the error comes from sampling people from a common population. Under the random-effects model, the error comes from sampling people from each population, and populations from a universe of populations. The additional width in the second diamond reflects additional error in estimating the summary effect. It tells us nothing about how widely the effects are actually dispersed, however.

To address the dispersion of effects, we turn to the prediction interval, which is denoted as PI. The prediction interval is 49.4 points wide. We expect that in some 95 percent of all relevant populations, the treatment will increase scores by at least 25 points to as much as 75 points.

20.5 STATISTICS FOR HETEROGENEITY

The statistics typically reported for heterogeneity are Q, df, p, I^2, T^2, T, and the prediction interval. These are related to each other, but each addresses a specific aspect of heterogeneity. To explain the meaning of each statistic, it is helpful to keep in mind that we are working with two distinct distributions—the distribution of observed effects and the distribution of true effects. Some statistics quantify variation in observed effects, some quantify variation in true effects, and some address the relationship between the two.

20.5.1 Observed Effects Versus True Effects

In a primary study with one level of sampling, we typically treat the observed scores as identical to the true scores. By contrast, in a meta-analysis, we need to distinguish between an observed effect size and a true effect size. The observed effect size is what we see in a study. It serves as the estimate of the effect size in the study's population, but invariably differs from the true effect size in that population due to sampling error. By contrast, the true effect size is the actual effect size. It is the effect size that we would see with an infinitely large sample size, and (it follows) no sampling error.

Figure 20.2 displays a fictional meta-analysis in which the error variance is the same in all studies. The left-hand plot shows the observed effects. This is the plot that is typically included with a published meta-analysis. The

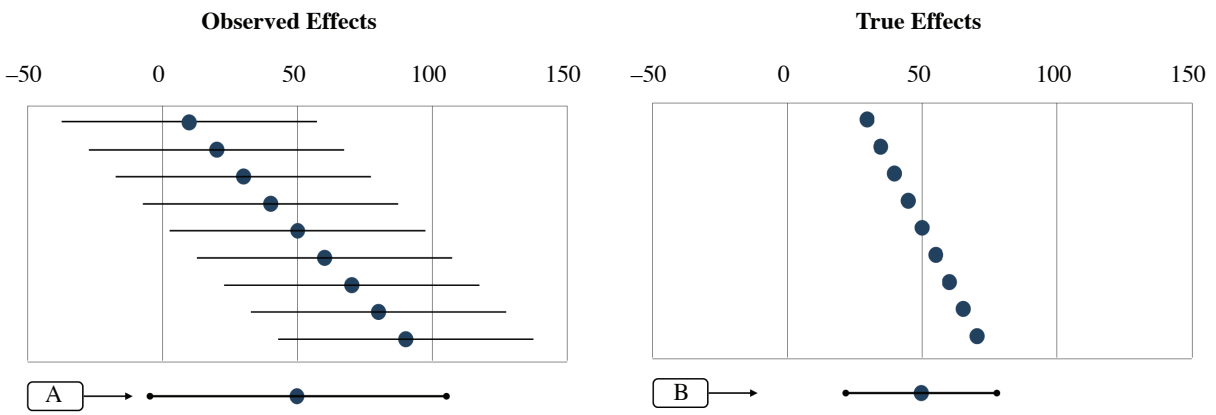

Figure 20.2 Observed Effects and True Effects for a Fictional Meta-Analysis

SOURCE: Author's tabulation.

standard deviation of the observed effects is 27.4, and we expect that some 95 percent of all observed effects will fall within two standard deviations of the mean. This corresponds to a range of 110 points (−0.05 to +1.05) as suggested by line [A].

By contrast, the right-hand plot shows the true effects for the same analysis. This is the plot that we would see if every study had an extremely large sample size, so that the sampling error was close to zero. This is the plot that we care about because it tells us how much the effect size actually varies. The standard deviation of the true effects is 13.69, and we expect that some 95 percent of all true effects will fall within two standard deviations of the mean. This corresponds to a range of 55 points (23 to 77) as suggested by line [B]. This is the prediction interval discussed earlier.

For the present discussion, the key point is that the variance of the observed effects tends to be larger than the variance of true effects. To understand why that is true, consider what would happen if the true effect size were identical in all studies. When the variance of true effects is zero, the expected variance of *observed* effects (V_{OBS}) would be equal to the typical error variance (V_{ERR}). That is,

$$V_{OBS} = V_{ERR}. \tag{20.1}$$

In words, the variance of observed effects is equal to the variance due to sampling error.

The same paradigm applies when the true effects vary, as they do here. If T^2 is the variance of *true* effects, then the expected variance of the *observed* effects is given by

$$V_{OBS} = T^2 + V_{ERR}. \tag{20.2}$$

In words, the variance of observed effects is equal to the variance of true effects plus variance due to sampling error.

The two plots are intended to convey the idea that the variance of true effects tends to be smaller than the variance of observed effects. I do not mean to suggest that any specific study on the right corresponds to a specific study on the left.

20.5.2 Quantifying Variation in Observed Effects

The Q-statistic refers to the left plot in figure 20.2, and is defined as the sum of squared deviations (of each observed effect from the mean effect) on a standardized scale. If all studies share a common true effect size (and all the variance in observed effects is due to sampling error), Q would approximately follow a chi-squared distribution with degrees of freedom equal to the number of studies minus 1. We can use this to obtain a p-value for a test of the null hypothesis that there is no variation in true effects. In this example Q is 10.67 with 8 degrees of freedom. The p-value for a test of the null (that all studies share a common true effect size) is 0.22 (for caveats, see Hoaglin 2016).

20.5.3 Quantifying Variation in True Effects

The statistic called T (tau) is the standard deviation of true effects. As such, it serves the same role as the standard deviation in a primary study. We can use the mean plus or minus 1.96 standard deviations to compute the approximate prediction interval. If the effects are normally distributed, then the true effect size in some 95 percent of all comparable populations will fall within this interval. In this example, T is 13.69. The mean of 50 plus or minus 1.96 standard deviations yield a prediction interval of approximately 23 to 77 (line B in figure 20.2).

This formula assumes that the mean effect size and standard deviation are known precisely. In practice, we use formula (20.7), which takes account of the fact that these statistics are estimated rather than known.

The statistic called T^2 is the variance of true effects. This is simply the standard deviation squared. As is true in a primary study, the variance is not a terribly intuitive measure given that it uses squared rather than linear units. However, it has statistical properties that make it useful in the computations. In particular, the variance is a component in the weight assigned to each study for purposes of computing the mean effect size. In this example, T^2 is 187.50.

20.5.4 Quantifying Relationship Between True and Observed Effects

If the right plot shows the variance of true effects and the left plot shows the variance of observed effects, it might be useful to have a statistic that quantifies the relationship between the two. This statistic is called I^2, defined as the ratio of true to total variance,

$$I^2 = \frac{V_{TRUE}}{V_{TOTAL}} \times 100 = \frac{V_{TRUE}}{V_{TRUE} + V_{ERR}} \times 100 = \frac{T^2}{V_{TOTAL}} \times 100. \tag{20.3}$$

If I^2 is the ratio of true to total variance, then it is also the proportion of the observed variance that would remain if we could somehow remove the sampling error from the plot. Equivalently, it is the ratio of the variance in the right plot to the variance in the left plot.

Here, the variance in the right-hand plot is 187.50 and the variance in the left plot is 750.0, so we can compute I^2 as

$$I^2 = \frac{V_{TRUE}}{V_{Obs}} \times 100 = \frac{T^2}{V_{Obs}} \times 100 = \frac{187.50}{750.00} \times 100 = 25\%. \tag{20.4}$$

We work with I^2 (the ratio of variances) rather than I (the ratio of standard deviations) because variances are additive while standard deviations are not. However, I^2 is not a terribly intuitive index. Consider figure 20.2. Suppose we are presented with the forest plot of observed scores (at left), and told that I^2 is 25 percent. Then we are asked to imagine what the dispersion of true score might look like. This is a difficult task, at best.

By contrast, if we take the square root of I^2 to get I, we have the ratio of standard deviations. This is in linear units, and as such is more intuitive. In this case

$$I = \frac{S_{True}}{S_{Obs}} \times 100 = \frac{T}{S_{Obs}} \times 100 = \frac{13.69}{27.49} \times 100 = 50\%. \tag{20.5}$$

This corresponds to the ratio of the lines under the two plots in figure 20.2. The observed effects are dispersed over a range of 110 points [A]. If we multiply that by 50 percent, we know that the true effects will be dispersed over a range of 55 points [B].

20.5.5 Statistics Are Not Interchangeable

To this point, I have introduced a series of statistics for heterogeneity. These are all mathematically related to each other, but they are not interchangeable—each quantifies a specific aspect of heterogeneity. Thus, if we want to report on some aspect of heterogeneity, we need to choose the statistic that addresses that aspect. This should be an obvious point, but in practice is often overlooked.

Some papers treat Q and its degrees of freedom as a surrogate for the amount of variation. In fact, these are simply an interim step in computing the standard deviation of true effects, and can tell us only that the estimated variance is (or is not) zero. Others focus on the variance, T^2. The variance is a key component in the process of assigning a weight to each study, but does not directly tell us how much the effects vary. Often, papers present I^2 as a surrogate for the amount of dispersion. This approach is so widespread that we discuss it in the common mistakes section of this chapter. As explained there, I^2 does not tell us how much the effects vary.

Rather, if we want to know how much the effect size varies, the only relevant statistic is the standard deviation of true effects, T, and the prediction interval.

20.6 STANDARD DEVIATION AND PREDICTION INTERVAL

In a primary study, if we want to report how much the scores vary, we invariably use the standard deviation. The standard deviation is a useful index because we intuitively understand what it says about the distribution of scores. If we are told that the mean score is 50 with a standard deviation of 10, we immediately understand that most scores fall in the range of 30 to 70. This process is so automatic that we do not think about it. But I want to be explicit about it here, and make the point that the standard deviation is not intrinsically useful—rather, it is useful because it yields a direct link to the prediction interval. The relevance of this point will become obvious shortly.

Researchers invariably report the standard deviation (S) in a primary study but rarely report the standard deviation (T) in a meta-analysis. Today, the reason for this practice is simply that researchers follow the common template. Originally, the reason was that although the standard deviation is an intuitive statistic for some effect-size indices, it is not for others.

Specifically, when we are working with a mean, mean difference, standardized mean difference, or a risk difference, the standard deviation is in the same metric as the effect size itself. In these cases, the standard deviation could serve the same function that it does in a primary study. If we are told that the mean effect size is a standardized mean difference of 0.50 and that the standard deviation is 0.10, we know that most effects will fall in the range of 0.30 to 0.70.

By contrast, when we work with a risk ratio or an odds ratio, the standard deviation is reported in log units, and few of us know what a standard deviation of (for example) 0.10 log units means. Additionally, we cannot use the mean risk ratio plus or minus 2 standard deviations to compute a prediction interval. Because the risk ratio is

reported in ratio units and the standard deviation in log units, the results of this computation would be meaningless. We face a similar problem with correlations (where the standard deviation may be reported in Fisher's *Z* units), prevalence (where it may be reported in logit units) and other indices.

This is why the standard deviation is not widely used as an index of heterogeneity in meta-analyses, but there is a simple way to bypass the problem. The key is that the standard deviation is not intrinsically important—instead, it is important because it allows us to intuit the prediction interval. If we can report the prediction interval in an intuitive metric, then the fact that the standard deviation may be in another metric becomes irrelevant (Borenstein et al. 2009, 2017; Higgins, Thompson, and Spiegelhalter 2009; Riley, Higgins, and Deeks 2011; Roth 2009).

In fact, it is possible to compute the 95 percent prediction interval in an intuitive metric for *all* effect sizes using

$$Interval = M \pm 1.96T, \tag{20.6}$$

if we simply convert M and T to the same units. For example, if T is in log units and M is in ratio units, we convert M to log units, compute the interval in log units, and then convert the limits back to ratio units. The same idea applies if we are working with correlations (where we might convert M to Fisher's Z units), prevalence (where we might convert M to logit units), or other indices where we employ a transformation.

This simple solution entirely solves the problem outlined. However, we still need to address an entirely separate problem, which is that formula (20.6) works well only if M and T are known precisely. To address the fact that they are estimated with error, we can use

$$Interval = M \pm t_{(df)}\sqrt{V_m + T^2}. \tag{20.7}$$

As before, we would convert all values to the same metric before applying the formula. This formula includes three adjustments to (20.6), as follows:

First, we have replaced T with the square root of T^2. This is the identical value, but this format allows us to combine two variance components in the next step.

Second, we have added the variance of the mean (V_M) to account for the fact that the true mean may be lower or higher than M.

Third, we have replaced the factor of 1.96 (which is the critical z-value for a 95 percent interval) with the critical *t*-value for *df*, to account for the fact that the standard deviation of effects is estimated, rather than known.

An Excel spreadsheet to perform these computations is available at www.Meta-Analysis/Prediction. To use it, we need only four values—the mean effect size (M), the upper limit of the confidence interval, the between-study variance (T^2), and the number of studies (k). Virtually all computer programs report these values, and so the spreadsheet can be used regardless of what program is used for the basic analysis.

Here, I present four examples that show how we would report and interpret the prediction interval for four effect-size indices. The computations for these examples are provided in (Borenstein et al. 2017).

20.6.1 Standardized Mean Difference

Xavier Castells and his colleagues performed a meta-analysis of seventeen studies that assessed the impact of methylphenidate on cognitive function in adults with ADHD (2011). The effect size is the standardized mean difference (d) between the treated and control groups on a cognitive task. The mean effect size is 0.51. We expect that in some 95 percent of all populations, the true effect size will fall in the approximate range of 0.06 to 0.95. The take-home message here might be that the treatment has a trivial effect in some cases, a moderate effect in others, and a substantial effect in others. Because the impact is always positive (not harmful) we might recommend that this treatment be employed immediately.

20.6.2 Risk Ratio

Alexander Tsertsvadze and his colleagues performed a meta-analysis of nineteen studies that evaluated the impact of Viagra on sexual function (2009). The outcome was the patient's report that he was (or was not) satisfied, and the effect-size index was the risk ratio. The mean effect size is 2.5, and we expect that in some 95 percent of all populations, the true effect size will fall in the approximate range of 1.8 to 3.5. The take-home message here might be that this treatment works well, and consistently. From a substantive perspective, if the drug increases the likelihood of success by 180 percent in some populations, and by 350 percent in other populations, it would probably be worth trying for everyone.

20.6.3 Prevalence

Mariana Cabizuca and her colleagues performed a meta-analysis to synthesize data from eleven studies that reported prevalence of post-traumatic stress disorder (PTSD) in mothers of children with chronic illness or undergoing invasive procedures (2009). The mean prevalence is 18 percent, and we expect that in some 95 percent of all populations, the true prevalence will fall in the approximate range of 5 percent to 47 percent. The take-home message here might be that the prevalence of PTSD varies so widely, that the mean prevalence is of little relevance. We need to understand where the risk of PTSD is relatively low, and where it is relatively high, which will allow us to target the populations with the higher risk.

20.6.4 Correlation

Thomas Wright and Douglas Bonett performed a meta-analysis to synthesize data from twenty-seven studies that reported the correlation between attitudinal commitment and job performance (2002). The mean effect size is 0.17. We expect that in some 95 percent of all populations, the true correlation will fall in the approximate range of –0.08 to 0.41. The take-home message here might be that the correlation varies so widely, that the mean correlation is of little relevance. We need to understand where the correlation is trivial (or negative), and where it is modest.

20.6.5 In Sum

My goal here is to provide examples of how the prediction interval provides context for understanding the results of the analysis. My evaluation that the effect size ranges from trivial to substantial (for the treatment of ADHD) is clearly subjective. Others will have a different opinion. My point is that the prediction interval is what we have in mind when we ask how much the effect size varies across studies, and the evaluation should be based on this rather than some other statistic.

20.7 HOW HETEROGENEITY AFFECTS MEAN AND STANDARD ERROR

To this point, I have focused primarily on the issue of heterogeneity itself. A separate issue is that heterogeneity will have an impact on our estimate of the mean effect, and also the precision with which we can estimate the mean. These issues are addressed here.

Figure 20.3 shows a fictional meta-analysis of six studies as displayed by the computer program Comprehensive Meta-Analysis Version 3 (Borenstein et al. 2014). The individual studies are displayed at the top, followed

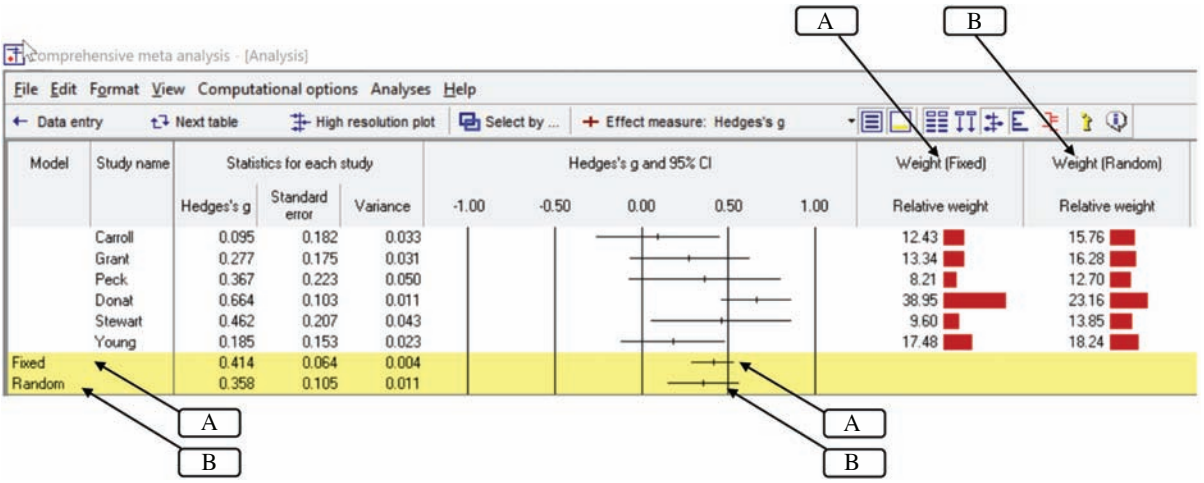

Figure 20.3 Meta-Analysis Showing Relative Weights for Fixed-Effect and Random-Effects Models

SOURCE: Author's tabulation.

by two rows that display the summary effect. The first [A] shows the summary effect for the fixed-effect model; the second [B] shows the summary effect using the random-effects model.

Under the fixed-effect model, T^2 is 0.000, and under the random-effects model, T^2 is estimated as 0.037. Therefore, by comparing the two rows, we can see what happens as the heterogeneity increases. First, the mean effect size shifts. Second, the confidence interval becomes wider, indicating that our estimate of the mean is less precise. I address each of these in turn.

20.7.1 How the Statistical Model Affects Estimates of the Mean

The mean effect size in a meta-analysis is computed as the weighted mean of the effect size in all studies. Larger studies (more accurately, studies with less error variance) always get more weight than smaller studies, but the extent of the difference depends on the statistical model. Under the fixed-effect model, the difference between large and small studies can be relatively extreme, whereas under the random-effects model the differences are more moderate. Therefore, as we move from a fixed-effect model to a random-effects model (that is, when we take account of between-study variance), large studies lose influence and small studies gain influence. This follows the logic of the two statistical models, as follows.

Consider a meta-analysis where our goal is to estimate the mean score for all students in a school. We draw ten random samples of various sizes, compute the mean score in each sample, and then synthesize the results. The fixed-effect model applies here because all studies are estimating the same parameter (the school mean). Because all studies are estimating the same value, if one sample has one hundred students and another has one thousand students we would assign ten times as much weight to the second sample.

By contrast, consider a meta-analysis for which our goal is to estimate the mean score for all schools in a district. We randomly sample ten schools, with the sample size in each school ranging from one hundred to one thousand students. In this case, the random-effects model applies. Here, we do not want to assign too much weight to the sample of one thousand. Although we know the mean in that school precisely, there is no reason to think that the mean in this particular school falls any closer to the district mean than the mean in any other school. Conversely, we also do not want to assign too little weight to a sample of one hundred. Although we do not know the mean in that school precisely, that sample provides the only estimate we have of that particular school's mean. So, we might assign (for example) twice as much weight (rather than ten times as much weight) to the larger sample as compared to the smaller one.

This logic finds expression in the weight assigned to each study, which is

$$W_i = \frac{1}{V_i + T^2}, \tag{20.8}$$

where V_i is the error variance for the *ith* study, and T^2 is the between-study variance. Critically, V_i is unique for each study but T^2 is the same for all studies. When T^2 is small relative to the typical V, the weights will be driven primarily by V. At the extreme, when T^2 is 0.00, a study with V of 10 will get ten times the weight of a study with V of 100. Conversely, when T^2 is large relative to the typical V, the weights will be driven primarily by T^2. At the extreme, if T^2 approached infinity, all studies would get essentially the same weight.

In figure 20.3, we can see why the summary effect shifted to the left when we moved to the random-effects model. In this example, the largest study (Donat) happens to have a high effect size, which would tend to pull the mean to the right. Under the fixed-effect model, this study was given 39 percent of the weight in the analysis, and so pulled the mean all the way to 0.414. Under the random-effects model, this study was given only 23 percent of the weight, and so pulled the mean only to 0.358. Also, as we moved from the fixed-effect to the random-effects model, the smaller studies, which tended to have smaller effects, gained influence, and were able to pull the mean further to the left. That is why the effect size shifted to the left in this example when the heterogeneity increased, but the shift can be either in direction. If Donat happened to have a small effect size, the mean would have shifted to the right as heterogeneity increased.

20.7.2 How the Statistical Model Affects Confidence Interval Width

The second difference was that as we moved from the fixed-effect model to the random-effects model, the confidence interval expanded. This will always be true (provided that T^2 is estimated as greater than zero), and the logic is reasonably straightforward.

Under the fixed-effect model, the true effect size is the same in all studies, so the only source of error is the fact that the observed effect size in each study differs from the true effect size in that study. The standard error of the common effect size is given by

$$SE_M = \sqrt{\frac{S^2}{N}}, \tag{20.9}$$

where S^2 is the common within-study error variance, and N is the total number of subjects, accumulated across studies.

By contrast, under the random-effects model there are two sources of error. One is that the observed effect size in each study differs from the true effect size in that study. The second is that the true mean in each study differs from the mean of all studies in the relevant universe of studies. The standard error of the mean effect size is given by

$$SE_M = \sqrt{\frac{S^2}{N} + \frac{T^2}{k}}. \tag{20.10}$$

As before, S^2 is the typical within-study error variance, and N is the total number of subjects, accumulated across studies. But we have added a new term to account for the second source of sampling error. In this term, T^2 is the between-study variance, k is the number of studies. It is what allows us to generalize from the studies in the analysis to the universe from which the studies were sampled.

As we move from formula (20.9) to (20.10), the first component of the error variance remains the same but the second component is additional. Therefore, unless T^2 is estimated as zero, the standard error will always be higher using (20.10) than using (20.9). The extent of the increase depends on the amount of heterogeneity and the number of studies in the analysis.

Researchers sometimes assume that a meta-analysis with a large number of subjects will yield a precise estimate of the mean effect size, but this is not necessarily true. There are two components to the error term—within-study error and between-study error. Whereas S^2 is divided by the total number of subjects, T^2 is divided by the number of studies. Therefore, if the between-study variance is substantial, the only way to obtain a precise estimate of the mean effect size is to include a substantial number of studies in the analysis. Increasing the number of subjects within these studies will reduce the within-study error but have no impact whatsoever on the between-study error.

Formulas (20.9) and (20.10) are useful for didactic purposes, but not in practice because they require that the within-study variance (S^2) be the same for all studies. In practice, we use a formula that allows S^2 to vary from study to study. Here, I used S^2 to denote within-study variance, to highlight the parallel to T^2, which denotes between-study variance. Normally, we use V rather than S^2 for thus purpose, and I return to that designation later.

In this example, I compared the mean effect for the fixed-effect model versus the random-effects model. Although our interest in this chapter is limited to the random-effects model, the example is nevertheless instructive because the fixed-effect model is computationally identical to a random-effects model when T^2 is zero. Therefore, in comparing the two models we can see what happens as T^2 increases. In this example, T^2 increases from zero to 0.037, but the same idea applies more generally. As was true in this example, as T^2 increases,

- the impact of small studies on the mean will increase, while the impact of larger studies will decrease; and
- the standard error of the mean will increase.

20.8 MISTAKES

In published meta-analyses, it is not unusual to see reports that interpret heterogeneity statistics incorrectly. These issues are explored elsewhere and summarized here (for more, see Borenstein 2019).

20.8.1 Using a Test for Heterogeneity to Choose a Statistical Model

As discussed earlier, most meta-analyses are based on either of two statistical models. The fixed-effect model is appropriate when all studies in the analysis are essentially replicates of each other. The random-effect model is appropriate in all other cases. By this criterion, when studies for the meta-analysis are culled from the published literature, the random-effects model will almost always apply. Some researchers have adopted the practice of starting the analysis with the fixed-effect model and then switching to the random-effects model if the test for heterogeneity meets the criterion for statistical significance. This practice is strongly discouraged (for a more extensive discussion and differing approaches, see Borenstein et al. 2010; Cooper and Hedges 2009; Cooper, Hedges, and Valentine 2009; Higgins, Thompson, and Spiegelhalter 2009; Viechtbauer 2007).

20.8.2 Assuming That I^2 Tells Us How Much Effects Vary

In some fields of research, I^2 has become the statistic most frequently cited as an index of heterogeneity in meta-analysis. A low value of I^2 is taken to mean that the effect size varies little across populations, and a high value that the effect size varies substantially. The use of I^2 as an index of absolute dispersion may be ubiquitous, but is nevertheless a fundamental mistake. I^2 is a proportion, not an absolute value. It tells us what proportion of the observed variance reflects variation in true effects, it does not tell us how much variation there is. Because I^2 is a proportion, by definition, it can never tell us how much the effects actually vary.

Consider the ADHD analysis introduced earlier. The researchers performed a meta-analysis of seventeen studies that assessed the impact of methylphenidate on cognitive function in adults with ADHD. The mean effect size is a standardized mean difference (d) of 0.50 and I^2 is 47 percent. If we are asked how widely the effects vary, we cannot answer based on this information. Before proceeding, take a moment and ask yourself how you would get from an I^2 of 47 percent to an estimate of the actual dispersion. It cannot be done.

Not only does I^2 not tell us the absolute amount of dispersion, it also does not reliably tell us the relative amount. If we know that I^2 in one meta-analysis is 25 percent and in a second is 75 percent, we might assume that there is more variation in the second analysis than the first. However, that is true only if the observed variance was comparable in the two analyses. Consider the two fictional analyses shown in figure 20.4. The first row shows the dispersion of observed effects (left) and true effects (right) for one intervention. The second row shows the dispersion of observed effects (left) and true effects (right) for a different intervention.

In the top row, the variance of observed effects is 750, I^2 is 25 percent and the prediction interval is 55 points wide. In the bottom row, the variance of observed effects is 187.5, I^2 is 50 percent and the prediction interval is 39 points wide. Thus, the larger value of I^2 corresponds to the smaller range of effects.

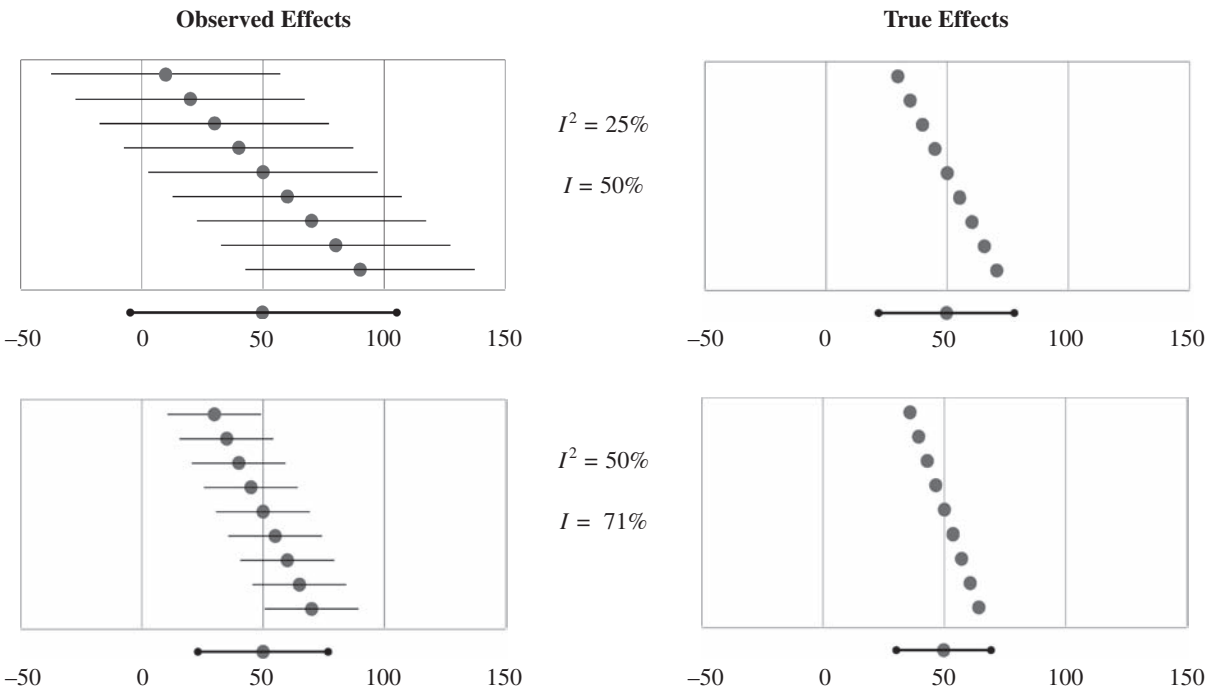

Figure 20.4 Observed Effects and True Effects for Two Fictional Meta-Analyses

SOURCE: Borenstein 2019.

The problem of using I^2 as a surrogate for the amount of dispersion is analogous to the well-known problem that researchers tend to use the *p*-value as a surrogate for the effect size. A significant *p*-value is assumed to reflect a large effect, and a nonsignificant *p*-value is assumed to reflect a smaller (or null) effect. It is true that significant *p*-values do tend to be associated with larger effects, but for any single study this fact is not terribly useful. Because the *p*-value is a function of the effect size and the sample size, a significant *p*-value *may* reflect a large effect but could also reflect a trivial effect size in a large study. Conversely, although a nonsignificant *p*-value can reflect a small effect size, it can also reflect a large effect size in a small study. Perhaps more to the point, even if a significant *p*-value does reflect a large effect size in a given case, if we are told the *p*-value, we do not know how large the effect size is. By contrast, if we are simply told the effect size, then we know the effect size, and no speculation is needed.

By analogy, in meta-analyses, the I^2 statistic is being pushed into service as a surrogate for the amount of dispersion. A high value of I^2 is assumed to mean that the effects vary widely, while a small value of I^2 is assumed to mean that the effects are consistent. Although it is true that high values of I^2 to tend to be associated with more dispersion, for any single meta-analysis this fact is not terribly useful, as is evident in figure 20.4. We need to multiply I^2 by the variance of observed effects to get the variance of true effects. Therefore, if I^2 is high the true effects could vary substantially but could also fall within a narrow range. Conversely, if I^2 is low the true effects could fall within a narrow range but could also vary widely. Perhaps more to the point, even if a high value of I^2 does reflect substantial dispersion in a given case, it we are told the value of I^2, we still don't know how much the effects vary. By contrast, if we are simply told the prediction interval, then we know the extent of dispersion, and no speculation is needed.

The correct way to use I^2 is to provide context for the forest plot. If I^2 is high, then the plot of true effects would look very similar to the plot of observed effects. If I^2 is low, then much of the dispersion would disappear, and the points would all move substantially toward the mean. At the extreme, if I^2 is 0 percent, then all of the dispersion would disappear, and all true effects would fall precisely at the mean. So, if we are reading a meta-analysis and the only information we have is the forest plot and I^2, then we could use I^2 to get a general sense of the dispersion. On the other hand, if we are reporting a meta-analysis, we should report the prediction interval.

The original papers on I^2 are by Julian Higgins and his colleagues (Higgins and Thompson 2002; Higgins et al. 2003; for more detailed discussion of the issues raised in this section, see Borenstein et al. 2017; see also Borenstein 2019; Coory 2010; Higgins 2008; Huedo-Medina et al. 2006; Ioannidis 2008; Patsopoulos, Evangelou, and Ioannidis 2008; Rücker et al. 2008).

20.8.3 Classifying Heterogeneity

Researchers sometimes classify heterogeneity as being small, moderate, or large based on the value of I^2. This is wrong for two reasons. First, it is based on a statistic (I^2) that does not reflect the amount of variance. Second, it is not useful without additional context to describe variance as being small, moderate, or large. Dispersion that might be trivial in some contexts will be large in others. The idea of using these classifications was proposed for a specific context (to compare the amount of dispersion among studies within the Cochrane Database of Systematic Reviews) and the authors never intended for these designations to be extended beyond that context (Higgins et al. 2003).

20.8.4 Conflating Confidence and Prediction Intervals

A key theme in this chapter is that the prediction interval is the preferred way to report the dispersion in effect sizes. Researchers sometimes confuse the prediction interval with the confidence interval (Higgins, Thompson, and Spiegelhalter 2009). These are two entirely separate indices, and it is important not to confuse one with the other (Borenstein et al. 2009; Guddat et al. 2012; Higgins, Thompson, and Spiegelhalter 2009; Roth 2009).

20.8.5 Failure to Report the Prediction Interval

When someone asks whether the effect size varies, they usually intend to ask how much it varies. The prediction interval is what captures this information and yet the prediction interval is rarely reported. The most common reason is probably that researchers are simply not familiar with this interval, but some who are familiar with it have raised concerns.

Some suggest that the prediction interval sometimes covers a wide swath of effects and may include the null.

That is true, but is also precisely why it is important to report this information. If the mean effect is positive but in some populations the effect will be nil (or even harmful), the details speak to the utility of the intervention and are critical information. An old joke tells of a statistician who drowned in a creek with a mean depth of three feet. If the creek is six feet deep in places, we need to know that.

20.8.6 When T^2 Is Estimated as Zero

As explained earlier, the fixed-effect model assumes that the true effect size is the same in all studies, whereas the random-effects model allows that the true effect size may vary from study to study. Under the fixed-effect model, we are estimating a common effect size, whereas under the random-effects model we are estimating a mean effect size.

If we are using the random-effects model, T^2 will sometimes be estimated as zero. Because the weight assigned to each study under the random-effects model is

$$W_i = \frac{1}{V_i + T^2}, \tag{20.11}$$

and the weight assigned to each study under the fixed-effect model is

$$W_i = \frac{1}{V_i}, \tag{20.12}$$

when T^2 in (20.11) is estimated as zero, this equation becomes functionally identical to (20.12) and the two models yield identical estimates for all statistics. In cases where someone has elected to use the random-effects model and it turns out that T^2 is estimated as zero, it is common for the researcher to report that the analysis is based on the fixed-effect model. That, however, is not correct. The analysis is based on a random-effects model, which in this case happens to yield the same values as a fixed-effect model.

This is an academic point, but an important one. The selection of a statistical model must conform to our understanding of the inference population. If the fixed-effect model applies, then our results can be generalized only to the one population studied. If the random-effects model applies, then our results can be generalized to the universe of populations sampled. In this case, if T^2 is indeed near zero, we would report that the effect size is consistent across the populations in this universe.

Parenthetically, if the studies are based on multiple populations and τ^2 is estimated as zero, we can assume that we have underestimated the true value, since the actual variance is almost certain to be positive. Indeed, we will underestimate τ^2 in about half of all cases, and if the correct value is small we can easily end up with a negative value (which is then set to zero).

20.9 WHEN WE HAVE ONLY A FEW STUDIES

When a meta-analysis includes only a few studies, our estimate of the variance, standard deviation, prediction interval, and I^2, will be unreliable. The extent of the problem is not generally recognized because researchers tend to look at the number of subjects and assume that if we have hundreds of subjects in the analysis, our estimate of these statistics must be reasonably precise. Again, that is simply not true. Our ability to estimate the amount of dispersion is driven primarily by the number of studies, not the number of subjects. Our ability to estimate the variance of effects based on five studies is no better (and is indeed worse) than our ability to estimate the variance of scores in a primary study based on five subjects.

The importance of this problem will vary by field of study. If we are working with studies that draw on similar populations, employ similar methods, and assess the impact of an intervention that tends to have a consistent effect, the concern may be mitigated. Conversely, if we are working with studies that draw from a universe where the populations and methods vary, and assess the impact of an intervention whose effect varies, the concern may be potentiated.

A related problem is the fact that we want to generalize from the studies in our analysis to a wider universe. We need to think carefully about how we define that universe, and the extent to which the studies in our sample are representative of that universe. When the analysis includes only a few studies, it might not be plausible to assume that these studies are representative of the larger universe.

As noted, this problem affects all heterogeneity statistics. If T^2 is unreliable then I^2, T, and the prediction interval will all be unreliable. Ironically, the practical impact of the problem is more pronounced for the prediction interval. An incorrect estimate of I^2 or T^2 will have little practical impact because researchers do not actually use those values (except to report them). By contrast, the prediction interval does address important information, and an incorrect estimate will have practical import.

20.10 CONCLUSION

In the second edition of this book, Harris Cooper and Larry Hedges wrote,

> for yesterday's synthesist, the variation among studies was a nuisance. It clouded interpretation. The old methods of synthesis were handicapped severely when it came to judgments of whether a set of studies revealed consistent results and, if not, what might account for the inconsistency. . . . For today's synthesist, variety is the spice of life. The methods described in this handbook make such analyses routine. When the outcomes of studies prove too discrepant to support a claim that one estimate of a relationship underlies them all, current techniques provide the synthesist with a way of systematically searching for moderating influences, using consistent and explicit rules of evidence. (Cooper and Hedges 2009, 563)

In the relatively brief span since that edition was published, the field has continued to evolve. Researchers are more likely to take account of heterogeneity when computing the mean effect and its confidence interval. Most meta-analyses report indices of heterogeneity, and many attempt to explain some of that heterogeneity using such tools as subgroups analysis or meta-regression as discussed elsewhere in this volume (see also Borenstein et al. 2019; Borenstein and Higgins 2013).

However, although the goal posts have shifted somewhat, the way the goals are being addressed is seriously problematic.

One issue is that researchers often report the wrong statistics for heterogeneity and then interpret them incorrectly. Researchers might report a *p*-value for heterogeneity or the value of I^2. Then they discuss the implications of the heterogeneity when in fact these statistics tell us nothing about how much heterogeneity there actually is. When this is done in a single meta-analysis, we lose a lot of potentially useful information. When this is done consistently, the entire field suffers.

When we speak about heterogeneity in a meta-analysis, what we usually have in mind is how widely the effect size varies. The way to address this question is to report the prediction interval and then discuss the substantive implications of this interval. For example, the prediction interval will allow us to report that

- The effect size is essentially the same in all studies, or
- The effect size varies somewhat, but the impact is non-trivial in all cases, or
- The effect size varies from trivial to exceptional, but is never harmful, or
- The intervention is harmful in some cases, and helpful in others.

One can argue about whether a given effect size is trivial or substantive, but it should be clear that the discussion should focus on the information captured by this interval, rather than statistics that quantify other aspects of heterogeneity.

20.11 Excel Spreadsheet

An Excel spreadsheet to perform all the computations is available at www.Meta-Analysis.com/Prediction or from the author at Biostat100@GMail.com.

20.12 Acknowledgments

The ideas discussed in this chapter are the result of my many years of collaboration with Larry Hedges, Julian Higgins, and Hannah Rothstein, although the opinions are mine. I would also like to express my appreciation to Emily Tanner-Smith and Jack Vevea for their thoughtful comments on an earlier draft of this paper.

20.13 REFERENCES

Althuis, Michelle D., Douglas L. Weed, and Cara L. Frankenfeld. 2014. "Evidence-Based Mapping of Design Heterogeneity Prior to Meta-Analysis: A Systematic Review and Evidence Synthesis." *Systematic Reviews* 3: 80. DOI: 10.1186/2046-4053-3-80.

Berlin, Jesse A. 1995. "Invited Commentary: Benefits of Heterogeneity in Meta-Analysis of Data from Epidemiologic Studies." *American Journal of Epidemiology* 142(4): 383–87.

Borenstein, Michael. 2019. *Common Mistakes in Meta-Analysis and How to Avoid Them*. Englewood, N.J.: Biostat, Inc.

Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hanna R. Rothstein. 2009. *Introduction to Meta-Analysis*. Chichester, UK: John Wiley & Sons.

———. 2010. "A Basic Introduction to Fixed-Effect and Random-Effects Models for Meta-Analysis." *Research Synthesis Methods* 1(2): 97–111. DOI: 10.1002/jrsm.12.

———. 2014. Comprehensive Meta-Analysis, Version 3. Englewood, N.J.: Biostat, Inc.

———. 2019. *Meta-Regression: Multiple Regression in Meta-Analysis*. Unpublished manuscript, Biostat, Inc.

Borenstein, Michael, and Julian P. T. Higgins. 2013. "Meta-Analysis and Subgroups." *Prevention Science* 14(2): 134–43. DOI: 10.1007/s11121-013-0377-7.

Borenstein, Michael, Julian P. T. Higgins, Larry V. Hedges, and Hanna R. Rothstein. 2017. "Basics of Meta-Analysis: I^2 Is Not an Absolute Measure of Heterogeneity." *Research Synthesis Methods* 8(1): 5–18. DOI: 10.1002/jrsm.1230.

Cabizuca, Mariana, Carla Marques-Portella, Mauro V. Mendlowicz, Evandro S. Coutinho, and Ivan Figueira. 2009. "Posttraumatic Stress Disorder in Parents of Children with Chronic Illnesses: A Meta-Analysis." *Health Psychology* 28(3): 379–88. DOI: 10.1037/a0014512.

Castells, Xavier, Josep Antoni Ramos-Quiroga, David Rigau, Rosa Bosch, Marianna Lima Nogueira, Xavier Vidal, and Miquel Casas. 2011. "Efficacy of Methylphenidate for Adults with Attention-Deficit Hyperactivity Disorder: A Meta-Regression Analysis." *CNS Drugs* 25(2): 157–69. DOI: 10.2165/11539440-000000000-00000.

Cooper, Harris M., and Larry V. Hedges. 2009. "Potentials and Limitations." In *The Handbook of Research Synthesis and Meta-Analysis*, 2nd ed., edited by Harris M. Cooper, Larry V. Hedges, and Jeffrey C. Valentine. New York: Russell Sage Foundation.

Cooper, Harris M., Larry V. Hedges, and Jeffrey C. Valentine. 2009. *The Handbook of Research Synthesis and Meta-Analysis*, 2nd ed. New York: Russell Sage Foundation.

Coory, Michael D. 2010. "Comment on: Heterogeneity in Meta-Analysis Should Be Expected and Appropriately Quantified." *International Journal of Epidemiology* 39(3): 932–33. DOI: 10.1093/ije/dyp157.

Guddat, Charlotte, Ulrich Grouven, Ralf Bender, and Guido Skipka. 2012. "A Note on the Graphical Presentation of Prediction Intervals in Random-Effects Meta-Analyses." *Systematic Reviews* 1: 34. DOI: 10.1186/2046-4053-1-34.

Hedges, Larry V., and Jack L. Vevea. 1998. "Fixed and Random-Effects Models in Meta-Analysis." *Psychological Methods* 3(4): 486–504.

Higgins, Julian P. T. 2008. "Commentary: Heterogeneity in Meta-Analysis Should Be Expected and Appropriately Quantified." *International Journal of Epidemiology* 37(5): 1158–60. DOI: 10.1093/ije/dyn204.

Higgins, Julian P. T., and Simon G. Thompson. 2002. "Quantifying Heterogeneity in a Meta-Analysis." *Statistics in Medicine* 21(11): 1539–58. DOI: 10.1002/sim.1186.

Higgins, Julian P. T., Simon G. Thompson, Jonathan J. Deeks, and Douglas G. Altman. 2003. "Measuring Inconsistency in Meta-Analyses." *British Medical Journal* 327(7414): 557–60. DOI: 10.1136/bmj.327.7414.557.

Higgins, Julian P. T., Simon G. Thompson, and David J. Spiegelhalter. 2009. "A Re-Evaluation of Random-Effects Meta-Analysis." *Journal of the Royal Statistical Society. Series A* 172(1): 137–59. DOI: 10.1111/j.1467-985X.2008.00552.x.

Hoaglin, David C. 2016. "Misunderstandings About *Q* and 'Cochran's *Q* Test' in Meta-Analysis." *Statistics in Med* 35(4): 485–95. DOI: 10.1002/sim.6632.

Huedo-Medina, Tania B., Julio Sanchez-Meca, Fulgencio Marin-Martinez, and Juan Botella. 2006. "Assessing Heterogeneity In Meta-Analysis: *Q* Statistic or I^2 Index?" *Psychological Methods* 11(2): 193–206. DOI: 10.1037/1082-989X.11.2.193.

Ioannidis, John P. 2008. "Interpretation of Tests of Heterogeneity and Bias in Meta-Analysis." *Journal of Evaluation in Clinical Practice* 14(5): 951–57. DOI: 10.1111/j.1365-2753.2008.00986.x.

Lau, Joseph, John P. Ioannidis, and Christopher H. Schmid. 1998. "Summing Up Evidence: One Answer Is Not Always Enough." *Lancet* 351(9096): 123–27. DOI: 10.1016/S0140-6736(97)08468–7.

Nikolakopoulou, Adriani, Dimitris Mavridis, and Georgia Salanti. 2014. "Demystifying Fixed and Random Effects Meta-Analysis." *Evidence-Based Mental Health* 17(2): 53–57. DOI: 10.1136/eb-2014-101795.

Patsopoulos, Nikolaos A., Evangelos Evangelou, and John P. Ioannidis. 2008. "Sensitivity of Between-Study Heterogeneity in Meta-Analysis: Proposed Metrics and Empirical Evaluation." *International Journal of Epidemiology* 37(5): 1148–57. DOI: 10.1093/ije/dyn065.

Riley, R. D., Julian P. T. Higgins, and Jonathan J. Deeks. 2011. "Interpretation of Random Effects Meta-Analyses." *British Medical Journal* 342: 549. DOI: 10.1136/bmj.d549.

Roth, Jonathan V. 2009. "Prediction Interval Analysis Is Underutilized and Can Be More Helpful Than Just Confidence Interval Analysis." *Journal of Clinical Monitoring and Computing* 23(3): 181–83. DOI: 10.1007/s10877-009-9165-0.

Rücker, Gerta, Guido Schwarzer, James R. Carpenter, and Martin Schumacher. 2008. "Undue Reliance on I(2) in Assessing Heterogeneity May Mislead." *BMC Medical Research Methodology* 8: 79. DOI: 10.1186/1471-2288-8-79.

Sutton, Alexander J., and Julian P. T. Higgins. 2008. "Recent Developments in Meta-Analysis." *Stat Med* 27(5): 625–50. DOI: 10.1002/sim.2934.

Thompson, Simon G. 1994. "Why Sources of Heterogeneity in Meta-Analysis Should Be Investigated." *BMJ* 309(6965): 1351–55.

Tsertsvadze, Alexander, Howard A. Fink, Fatemeh Yazdi, Roderick MacDonald, Anthony J. Bella, Mohammed T. Ansari, Chantelle Garritty, Karla Soares-Weiser, Raymond Daniel, Margaret J. Sampson, Steven Fox, David Moher, and Timothy J. Wilt. 2009. "Oral Phosphodiesterase-5 Inhibitors and Hormonal Treatments for Erectile Dysfunction: A Systematic Review and Meta-Analysis." *Annals of Internal Medicine* 151(9): 650–61. DOI: 10.7326/0003-4819-151-9-200911030-00150.

Viechtbauer, Wolfgang. 2007. "Hypothesis Tests for Population Heterogeneity in Meta-Analysis." *British Journal of Mathematical and Statistical Psychology* 60(Pt 1): 29–60. DOI: 10.1348/000711005X64042.

Wright, Thomas A., and Douglas G. Bonett. 2002. "The Moderating Effects of Employee Tenure on the Relation Between Organizational Commitment and Job Performance: A Meta-Analysis." *Journal of Applied Psychology* 87(6): 1183–90. DOI: 10.1037//0021-9010.87.6.1183.

PART VIII

SUMMARY

21

TRANSPARENT REPORTING: REGISTRATIONS, PROTOCOLS, AND FINAL REPORTS

EVAN MAYO-WILSON
Johns Hopkins University

SEAN GRANT
Indiana University

CONTENTS

21.1 INTRODUCTION

Complete and transparent reports of research syntheses are necessary both to understand their results and for other investigators to reproduce their methods and (using the same data) obtain the same results (Goodman, Fanelli, and Ioannidis 2016). Unfortunately, many studies are wasted because they are not described accurately, comprehensively, and transparently in publicly accessible reports (Glasziou, Altman, et al. 2014; Lund, Juhl, and Christensen 2016). Reviews of research syntheses demonstrate that they may use many methods to answer similar questions (Valentine et al. 2010), and many research syntheses do not describe the information that would be required to reproduce their findings (Page, Shamseer, et al. 2016). Reports of research syntheses are more likely to be usable by fellow scientists and decision makers (Moher et al. 2016) by following standards for transparency that apply to all areas of science (Nosek et al. 2015).

Whereas other chapters describe how to conduct research, this chapter focuses specifically on reporting research. Because systematic reviews with meta-analyses that summarize the evidence about predefined research questions are the most common research synthesis method, this chapter provides practical instructions for reporting the design, conduct, analysis, and findings of these studies. It includes three major sections corresponding to important steps in the research synthesis process: registering the study, writing a protocol, and producing a final report.

In the past, it might have been difficult for a single manuscript to include all information described in this chapter. Today, journals will publish both protocols and final reports of research syntheses, and page limits are not barriers to transparency because most journals allow authors to include online materials (for example, describing their methods or results completely), and authors can use free online resources to share materials, code, and data.

21.2 REGISTRATION: BEGINNING A RESEARCH SYNTHESIS

Registration is the first step toward promoting the transparency and reproducibility of research syntheses (Booth et al. 2011b; Straus and Moher 2010). Many journals endorse reporting guidelines that emphasize the importance of registration (Liberati et al. 2009; Manchikanti et al. 2009; Whiting et al. 2011); thus, the prospective registration (that is, before beginning study screening and data collection) of research syntheses is increasingly expected in reports submitted for publication. A research synthesis can be registered several ways. Most notably, PROSPERO is an international prospective register of systematic reviews with health-related outcomes. The database is open to the public; users can create accounts and register reviews quickly and at no cost (Booth et al. 2013). PROSPERO prompts researchers to provide a minimum dataset by answering a series of questions about their research questions and methods (Booth et al. 2012, 2011a). Staff review new entries and, once approved, assign a unique identification number to each review. Because peer reviewers and editors increasingly ask for information about study registration, researchers who register on PROSPERO may find their registration numbers useful when they publish detailed protocols (for example, online or in a journal article) and when they publish their final reports. For those conducting reviews that do not have health-related outcomes, the Open Science Framework (OSF) allows prospective registration of any type of study; users can create transparent and complete registration records for projects by taking "snapshots" (with the completed protocols and data extraction forms) prior to data collection. Authors using the OSF should include all relevant fields of information requested by PROSPERO.

Two major producers of systematic reviews also ask researchers to register their reviews before writing protocols. First, the Cochrane Collaboration publishes systematic reviews and meta-analyses about a range of topics related to biomedicine, behavioral health, and public health. Second, the Campbell Collaboration publishes systematic reviews about social interventions within the policy areas of crime and justice, education, social welfare, and international development. Both also publish reviews about research methods, such as the impact of reporting guidelines to improve research transparency (Turner et al. 2012). Researchers who register a title with either organization commit to publishing a detailed protocol and final report with the organization. Cochrane reviews are registered automatically in PROSPERO, and they are published in the *Cochrane Database of Systematic Reviews*, which is indexed in Medline and other data-

bases. Whatever method one uses to register a research synthesis, the registration should be as complete as possible to demonstrate to future readers whether the questions and methods used were decided a priori.

By registering research syntheses and searching for related registrations, scientists can increase confidence in their findings, prevent duplication of effort, and identify research collaborators. Registration also can have unexpected benefits. For example, researchers in the United Kingdom and Pakistan met when they tried to register a Cochrane review about the same question (Imdad et al. 2010); rather than conduct two separate reviews, the teams worked together to complete the project faster than either could have done alone. More broadly, prospective registration of the hundreds of systematic reviews published each year would make research syntheses more credible and informative for practice (Ioannidis 2016; Page, Shamseer, et al. 2016).

21.3 THE PROTOCOL: DESCRIBING PLANS FOR RESEARCH SYNTHESES

Writing detailed protocols demonstrates researchers' commitment to transparency and reproducibility, and there are several direct benefits to publishing protocols. Because systematic reviews may be biased if they fail to follow specified methods (Kirkham, Altman, and Williamson 2010; Silagy, Middleton, and Hopewell 2002), tools used to assess the quality of research syntheses emphasize the importance of protocols (Shea et al. 2007; Whiting et al. 2016). Moreover, both the Cochrane and Campbell Collaborations require that researchers publish their protocols before collecting and analyzing data. Writing a protocol and tracking amendments can help guide the conduct of a research synthesis and maintain consistency across members of the team for the duration of the study. Moreover, information in a protocol may be used to write the introduction and methods for the final report, leading to faster publication. Finally, referencing a published protocol can reduce the need to include detailed methods in the report, freeing space to report results and discuss conclusions.

Protocols are usually more detailed than registrations. Protocols should typically include a title, introduction (background) and objectives, and methods. Of these, the methods section is likely to be the longest and most detailed, including descriptions of the planned research questions, search strategy, eligibility criteria, data extraction, quality assessment, and analysis. Reporting guidelines—most notably the PRISMA extension for protocols—have been developed by leaders in research synthesis methods to improve the transparency and completeness of systematic review and meta-analysis protocols (Moher et al. 2015; Shamseer et al. 2015). The Cochrane Collaboration has also published two handbooks that describe what to include in review protocols for diagnostic tests (Deeks, Bossuyt, and Gatsonis 2010) and review protocols for intervention reviews (Higgins and Green 2011), and both are available online at no cost.

Several journals publish protocols of research syntheses. Most prominently, *Systematic Reviews* publishes protocols for several types of research syntheses, such as intervention effectiveness reviews with meta-analysis, systematic reviews on research methods and reporting, qualitative evidence syntheses, and realist reviews, amongst others. Additionally, some journals such as *Psychological Science* have begun awarding "badges" for registering study design and analysis plans; some will review reports based on the protocol and agree to publish the results without regard to their direction or statistical significance. Badges were introduced to foster open science, and they may help reduce reporting biases (Kidwell et al. 2016).

Although publication in an open-access journal may be the best way to maximize visibility and access, authors who want to share their protocols without publishing in journals can use institutional or public repositories (for example, figshare, https://figshare.com/) to post dated documents with permanent IP addresses. In addition to registering studies, the OSF provides a comprehensive solution to project management, which also allows researchers to share materials and data.

21.3.1 Title

A clear title improves the likelihood that an interested reader will find, read, and use a research synthesis. Because many journals and reports are published every year, most readers use electronic searches and automated alerts to identify research that is relevant to their needs. To help readers find a research synthesis, the title should state the research question and method, as well as identify the report as a protocol (for example, "a protocol for a systematic review"). Although some journals allow creative titles (such as those using wordplay or allusions), these should be avoided because they reduce the chance that a review would be identified through an electronic search or by scanning titles for relevance.

21.3.2 Introduction

The introduction should explain the rationale and motivation for the research synthesis. For example, common scientific reasons to conduct a systematic review and meta-analysis are (1) to resolve differences in results

across studies of the same question and (2) to estimate an effect or association more precisely by combining the results of individual studies. Research syntheses are used to inform policy decisions and to design new primary studies; if relevant, the introduction might explain that a research synthesis was conducted to develop clinical guidelines or to identify priorities for future research (Li et al. 2012).

Most guidance has been developed specifically for systematic reviews of interventions. Normally, the introduction should describe the problem or association of interest and how the independent variable or variables might be related to the dependent variable or variables. It should summarize what is already known and what the results of the review will add. The introduction should state if previous systematic reviews and meta-analyses have been conducted and, if so, it should explain how this review would differ from them. For example, a new review may be appropriate if the previous review is outdated.

The Cochrane Collaboration provides required and recommended headings used to structure the backgrounds of reviews about interventional studies; these headings and others are required for submitting a review using the Cochrane Collaboration's software (RevMan 5.3. 2014):

- description of the condition
- description of the intervention
- how the intervention might work
- why it is important to do this review

For a review of diagnostics test accuracy, these are (RevMan 5.3, 2014):

- target condition being diagnosed
- index test(s)
- clinical pathway
- prior test(s)
- role of index test(s)
- alternative test(s)
- rationale
- objectives

Authors should provide a brief statement about the objectives of a research synthesis. For systematic reviews, this typically involves an operationalization of the question or questions to be addressed, including: the eligibility criteria, independent and dependent variables, comparisons, and eligible study types (Squires, Valentine, and Grimshaw 2013). Each of these items should be described in detail in the methods section.

21.3.3 Methods

Several guidelines for conducting systematic reviews have been published, and authors of review protocols should consult them when writing their methods sections. For example, the Institute of Medicine (IOM) guidelines are widely used, and many federal funders in the United States require or anticipate that researchers will follow them (Institute of Medicine 2011). The Cochrane Collaboration has produced a checklist, Methodological Expectations of Cochrane Intervention Reviews, or MECIR (Chandler et al. 2013), which is used internationally. Although reporting guidelines for protocols ask researchers to state whether they will use specific methods (Moher et al. 2015), IOM and Cochrane guidelines state that reviews are expected to use specific methods associated with rigor and transparency. Thus, they are useful resources when deciding which methods to include in a protocol for a systematic review.

21.3.3.1 Inclusion and Exclusion Criteria To be transparent and reproducible, the criteria used to decide which studies are included and excluded in a research synthesis must be stated clearly and unambiguously. The protocol should state, as precisely as possible, the eligibility criteria, defining each relevant term as needed. Eligibility criteria should typically address the following:

- study designs (for example, randomized trials)
- participants or places (for example, young people age twelve through eighteen diagnosed with major depressive disorder)
- independent variables (for example, group cognitive behavioral therapy)
- eligible comparisons (for example, usual care)

Some research syntheses are limited to a specific study design (for example, randomized controlled trials) but others include multiple study designs; whatever approach is used, the protocol should describe exactly which types of studies will be included and excluded.

Interventions are often described according to their goals or aims, and many reports do not include enough information to reproduce interventions described in research reports (Glasziou, Macleod, et al. 2014). The protocol should describe the specific content and structure of eligible interventions as objectively as possible (Montgomery et al.

2013), including any restrictions by format, duration, or other aspects of intervention implementation.

In addition to the inclusion criteria, the protocol should state whether any exclusion criteria will be applied. For example, a research synthesis might be limited to English-language reports, or participants might be excluded because of comorbid problems (for example, substance misuse, serious mental illness) or because people cannot receive an intervention (for example, because of language or geographic restrictions).

Finally, protocols should anticipate handling studies that include a mix of eligible and ineligible participants or variables. For example, a researcher interested in adolescents between the ages of fifteen and eighteen might find a study that assessed adolescents between fourteen and seventeen. The protocol should explain if disaggregated data would be sought and, if the researchers cannot obtain data for each group, the protocol should state which studies would be included or excluded (for example, depending on the proportion of participants that meet the inclusion criteria).

21.3.3.2 Dependent Variables (Outcomes) Although often described with eligibility criteria, researchers should normally consider the dependent variables (outcomes) of interest separately from the inclusion criteria. Many primary research reports do not include all of the outcomes measured. Moreover, most studies are not registered prospectively and most registered studies do not define their outcomes in enough detail to determine which outcomes they actually assessed (Cybulski, Mayo-Wilson, and Grant 2016; Zarin et al. 2011). Thus, excluding studies based on the outcomes or time-points reported in a journal article or other research report may lead to the inadvertent exclusion of eligible studies. If a study is eligible but does not report the outcomes of interest, researchers may contact the study authors to request information that would be needed to include the study in the review.

Protocols should include the following information, or state explicitly why this information is not included:

- dependent variables (for example, school attendance)
- moderators or mediators of interest (for example, sex, changes in depression)
- eligible time-points (for example, after one year)

Guidelines for registering clinical trials recommend describing several elements to define each outcome of interest (Zarin et al. 2011). Protocols for systematic reviews, however, rarely include all of the information required to define their outcomes (Saldanha et al. 2014). For example, researchers should anticipate in the protocol how they will handle multiple related outcomes (for example, multiple questionnaires for measuring depression).

21.3.3.3 Searching for Studies A complete description of the literature search is needed to critically appraise or reproduce the search strategy (Atkinson et al. 2015). Important items to describe include information about electronic databases, journals, bibliographies of identified studies, forward citation searches, searching for unpublished documents, and direct contact with authors and experts.

In particular, systematic reviews use highly sensitive search strategies to identify relevant studies. A search of electronic databases (for example, PsycINFO, MEDLINE), done in collaboration with an information scientist, is often the centerpiece of this process. Because all results of the review will depend on the studies identified, it is essential to document electronic search strategies clearly. Documentation should include the exact search terms used, the Boolean operators used to combine them, and the fields searched.

Because the results of unpublished research often differ from the results of published research (Chan et al. 2004; Cuijpers et al. 2010; Dwan et al. 2013), many research syntheses include additional methods to identify unpublished studies and outcomes (Agency for Healthcare Research and Quality 2014; Higgins and Green 2011; Institute of Medicine 2011). Researchers increasingly have access to databases and to unpublished research reports; if relevant, protocols should describe plans to search for unpublished data, study materials, and analytic code. These might include contacting authors, requesting data from companies or regulators (such as the Food and Drug Administration), or reviewing reports to specific funders or agencies.

After the search has been conducted and citations have been retrieved, researchers decide which reports merit further consideration. The screening process is normally conducted in duplicate, and the methods for screening studies and determining eligibility should be described in the protocol. The results should be documented clearly so they can be included in the final report.

21.3.3.4 Data Extraction The protocol should describe which information will be extracted and how it will be recorded. Researchers should publish a dated copy of the final form used for data collection (for example, as an online supplement) with the protocol or the final report.

Like study screening, data extraction is usually conducted in duplicate. If done by two researcher, discrepancies

during the data collection should be documented and the methods to resolve them should be described (for example, discussion). These might include simple errors (for example, typos) as well as disagreements about subjective coding decisions.

Data collection has been conducted historically using a paper form and then entering the information into a spreadsheet or database. Today, several systems allow researchers to extract data directly into online databases. For example, researchers can follow guidance for extracting data using Systematic Review Data Repository (SRDR), a free online system that supports form development and data sharing (Li, Vedula, et al. 2015), or EPPI-Reviewer, an online software program for managing and analyzing data in research syntheses.

21.3.3.5 Risk of Bias in Individual Studies Assessing risk of bias, that is, threats to internal validity, is a hallmark of systematic reviews and may be especially important for reviews with subjective outcomes (Page, Higgins, et al. 2016). Several scales have been used to rate study quality, but methods to derive a single summary score have been unsuccessful (Valentine and Cooper 2008). They are consequently discouraged.

The protocol should identify how the review will rate risk of bias in each included study. The Cochrane risk of bias tools for randomized studies and nonrandomized studies are commonly used (Higgins and Green 2011; Higgins et al. 2013; Sterne et al. 2016). The Design and Implementation Assessment Device (DIAD) also helps researchers assess the specific and overall quality of intervention research (Valentine and Cooper 2008). Other scales have been developed for critical appraisal and for syntheses of observational (Stroup et al. 2000), epidemiological (Sanderson, Tatt, and Higgins 2007; Shamliyan, Kane, and Dickinson 2010), case-control (Wells et al. 2014), diagnostic (Whiting et al. 2011, 2013), and qualitative studies (Tong et al. 2012).

Many systematic reviewers assess risk of bias but fail to consider bias in their synthesis and conclusions (Katikireddi, Egan, and Petticrew 2015). The protocol should describe how the review will incorporate the assessment of bias in their data analysis (for example, by conducting sensitivity analyses) and when summarizing the overall results and conclusions.

21.3.3.6 Summary Measures The protocol should describe the planned methods for describing study results, including continuous and categorical data from each included study. For example, results from included studies might be expressed using odds ratios, relative risks, or risk differences. Precision could be expressed as confidence intervals or standard errors (Valentine, Aloe, and Lau 2015). When possible, expressing information in the same way across studies will make it easier to compare them.

21.3.3.7 Unit of Analysis Systematic reviews may include studies with different units of analysis. The protocol should specify if studies with different units will be included and, if so, how studies will be compared in an unbiased manner.

Many studies of interventions compare groups (clusters) rather than individuals. For example, studies of educational interventions could compare students, teachers, classrooms, schools, states, or countries. An intervention might be delivered to schools and outcomes might be measured for students; the statistical analysis should address correlation among students in a classroom. If the review will include clustered studies, the protocol should describe how clustering will be addressed in the risk of bias assessment, statistical analysis, and conclusions (Richardson, Garner, and Donegan 2016).

Studies of diagnostic tests might include individuals who receive more than one test. The unit of analysis should also be considered for crossover studies of intervention, which are often reported incompletely and incorrectly. For example, studies in which participants receive more than one intervention should specify plans for handling missing data (for example, participants who complete some but not all periods), report results for each period, and include the period effect in their analysis (Li, Yu, et al. 2015).

21.3.3.8 Statistical Synthesis The protocol should explain how the results of individual studies will be presented and, if appropriate, combined. A statistical synthesis of results (meta-analysis) may be possible if the included studies assess similar outcomes in similar ways. In some cases, it would be inappropriate to combine studies making different comparisons or measuring different outcomes. Whether meta-analysis is used or not, researchers should consider how they will organize the review to address different comparisons, outcomes, time-points, and statistical information across studies.

If a meta-analysis may be conducted, the protocol should describe the methods that will be used to combine studies, including the effect metric (such as standardized mean difference or relative risk), the methods that will be used to calculate effects (such as Hedges *g* or Cohen's d), and the methods that will be used to combine studies (for example, random effects or fixed effect, or Bayesian

methods). Missing data is common, and the methods should also describe plans for handling missing data (such as assumptions about dropouts) and for selecting results from included trials that report estimates for different analysis populations (Li et al. 2014; Liu, Wei, and Zhang 2006). Formulas need not be included in the protocol, but the statistical packages or procedures that will be used should be referenced clearly so that another analyst could identify and reproduce the analysis.

21.3.3.9 Subgroup and Sensitivity Analyses Subgroup analyses in research syntheses allow researchers to consider potential differences across populations or other variables that could differ within studies or between studies. For example, subgroup analyses might explore whether an intervention has similar effects in different countries or whether the effects differ for boys and girls. The first analysis might simply organize the overall results of each study according to the country in which the study was conducted. For the second analysis, a researcher might need to extract two results for each trial (that is, one result for boys and one for girls). Within-study comparisons are likely to be more informative than between-study comparisons, and the protocol should describe if one or both types of comparisons would be made.

Sensitivity analyses might be used to explore differences across studies or to explore the impact of certain methods (for example, inclusion and exclusion criteria). For example, a researcher might compare the results of studies at high risk of bias with studies at low risk of bias. A sensitivity analysis might also be performed by repeating an analysis using a different method (such as fixed effect versus random effects).

For subgroup and sensitivity analyses, the protocol should specify both the type of variables that will be used for analysis (for example, country) and the exact categories that will be compared (for example, all countries in Asia compared with all countries in Africa, China versus Japan versus Korea). To limit false positives and to prevent bias, most reviews should include few planned subgroup and sensitivity analyses, and there should be few categories in most subgroup and sensitivity analyses. The protocol should specify which analyses will be conducted, the reasons one might expect a difference between studies or groups, and the anticipated direction of effects.

21.3.3.10 Heterogeneity Studies included in a systematic review or meta-analysis might find different effects because of chance or because of clinical or methodological heterogeneity. For example, results for the same intervention might differ across places that have different standards of care (that is, normal services that all participants receive in addition to the intervention being investigated). The presentation of study characteristics and subgroup analyses can identify potential sources of heterogeneity. Furthermore, it may be important to identify clinical and methodological heterogeneity even in the absence of statistical heterogeneity; the protocol should describe planned qualitative methods for assessing heterogeneity.

If a meta-analysis might be conducted, the protocol should describe statistical tests for heterogeneity and how the results of the review will be interpreted if there is more heterogeneity than one would expect by chance (Higgins et al. 2003). In an analysis with a great amount of heterogeneity, the average effect might be uninformative; the protocol should anticipate under what circumstances, if any, the results of individual studies would not be combined.

21.3.3.11 Risk of Reporting Bias Publication bias refers to the selective reporting of entire studies based on their results. Selective outcomes reporting refers to the partial reporting of results from studies based on the magnitude or significance of the results. Both types of reporting bias have similar consequences and can result in meta-bias in systematic reviews and meta-analyses (Goodman and Dickersin 2011).

There are several methods for identifying reporting bias. Some authors directly compare multiple reports for evidence of underreporting (Mathieu et al. 2009). Most statistical tests use a funnel plot (Sterne and Egger 2001) to determine whether the effects of small studies differ from larger studies; some procedures have been developed for "correcting" for publication bias (Duval and Tweedie 2000), though these tests may perform poorly when heterogeneity between studies is substantial (Peters et al. 2007).

Review protocols should describe planned methods to minimize the risk of reporting bias on the review (for example, searching for grey literature and contacting authors), or state that no such methods were used. Protocols should also describe planned methods to assess publication bias and to "correct" for publication bias, or state that no statistical tests will be conducted. Researchers should describe any circumstances under which planned tests would not be conducted (for example, because of the number of included studies). Finally, the protocol should describe how the researchers will evaluate the potential impact of reporting biases on their results and conclusions.

21.4 THE FINAL REPORT: CLEAR AND COMPLETE REPORTING

21.4.1 Incorporating Information from the Protocol

When a researcher has a complete protocol, many sections in the protocol can be incorporated easily in the final report, either directly or by reference. The title may be nearly identical and, as appropriate, should use words that identify the type of review (for example, systematic review, research review, research synthesis) and the methods used (for example, meta-analysis, network meta-analysis).

21.4.2 Reporting Guidelines

Reporting guidelines can help authors write reports that include all the information needed to understand their methods and results. The most commonly used reporting guideline for systematic reviews focuses on interventional studies (Liberati et al. 2009). Extensions to the PRISMA statement describe specific information needed for interventional reviews related to health equity (Welch et al. 2012) and harms (Zorzela et al. 2016) and for reviews using specific methods such as network meta-analysis (Hutton et al. 2015) and individual participant data meta-analysis (Stewart et al. 2015).

Reporting guidelines other than PRISMA have been underused (Fleming, Koletsi, and Pandis 2014). Some are broad and relate to multiple types of reviews, including the meta-analysis reporting standards (MARS) guidelines (APA 2008). Others provide guidance for reviews combining epidemiological (Manchikanti et al. 2009), diagnostic accuracy (Whiting et al. 2011), or qualitative studies (Tong et al. 2012). Some journals require that reviews not only conform to reporting guidelines but also include a checklist documenting where each piece of information is included in the report.

21.4.3 The Abstract

Many readers will read only the abstract, particularly in non-open-access journals, so this may be the most important part of the final report. Therefore, it is essential that the abstract provide an accurate and complete summary of the review. If possible, the abstract should be structured and follow relevant reporting guidelines (Beller et al. 2013).

21.4.4 Introduction

Journals have highly variable expectations for introductions. Some journals allow a few hundred words. Other journals require or expect much more comprehensive description of existing literature. Consequently, the introduction might span several pages, drawing much from the study protocol, or the introduction might provide a brief summary of information in the protocol.

21.4.5 Methods

If no changes were made to the researchers' plans, the objective and methods in the final report may be essentially the same as the protocol. Journals typically see published protocols as strengths (for example, to identify revisions in research synthesis methods or assess the extent of selective outcome reporting in a meta-analysis) and do not consider them duplicate publications. Changes to planned methods are extremely common, and researchers should not be abashed about explaining that plans changed during the conduct of their studies. Instead, deviations from protocols should be documented and explained in final reports.

21.4.6 Results

Study Selection The study selection process should be reported in a flowchart (Liberati et al. 2009), and the final report should describe the qualifications and training of people who conducted each step in the review process. Researchers should report the date on which each search was conducted and the number of citations retrieved from each source (for example, electronic database). The review should report the number of citations remaining after duplicate citations have been identified and removed.

Once a list of unique citations has been identified, all or some of the selection process may be conducted in duplicate. Each researcher may identify studies that are potentially eligible based on the title or abstract. The report should describe the number of potentially relevant citations identified by at least one researcher and the number of citations that were selected for full-text review. To report this information, it is often helpful for each researcher to keep a record of their work by retaining copies of reference libraries sorted by each reviewer at each stage of the process.

The number of full-text reports assessed for eligibility should be described. Following full-text review, the number of included studies as well as the number of reports

about the included studies should be stated. Researchers should also report the number of excluded studies, and for studies that were similar to those included, the reasons for excluding each study (for example, did not include an eligible comparator).

In addition to articles including enough information to assess eligibility, researchers often encounter reports (such as conference abstracts) that do not provide enough information to assess eligibility. Researchers might contact authors for more information, and the final report should describe how any reports were handled when the researchers could not obtain enough information to determine eligibility.

It might not be necessary to report agreement statistics if two researchers discussed each potentially relevant citation and reached consensus about the included studies. Because of resource constraints, some researchers double-code only a sample of citations; in these cases, levels of agreement or inter-rater reliability for double-coded citations should be reported.

If authors of primary studies were contacted for unpublished data, these methods and the data received should be described.

Data Extraction Procedures for data extraction should be described in the protocol, and any deviations from the protocol should be described in the final report, including dated copies of changes to the extraction form and related documents (such as guidance for using the form and codebooks).

The final report should indicate who extracted data, their qualifications, and any training provided. As with study selection, some reviews report the level of initial agreement between independent data extractors; in reviews that extract all data in duplicate and resolve differences through discussion, the level of initial agreement may be unimportant. If only some of the data were extracted in duplicate and other data were extracted by a single researcher, the final report should describe agreement between researchers for the data extracted in duplicate. Agreement statistics may be especially important for reporting the consistency of subjective assessments (for example, risk of bias assessment).

Characteristics of Included Studies Before describing the results of the included studies, it is often useful to summarize the study characteristics (for example, study design, participants, interventions, comparisons, and settings). The final report review should include data for the variables specified a priori in the protocol, which is often done in a table of included studies (Higgins and Green 2011). Commonly, researchers also discover information about participants or differences across studies that were not anticipated in the protocol; these can also be described, and the report should explain why these variables were added.

Excluded Studies The report should describe the number of studies excluded for each reason, and citations to excluded studies may be provided in a table or appendix.

Studies not meeting all inclusion criteria for a systematic review or meta-analysis may, nonetheless, include important information about the topic under review. It can be tempting to describe the results of excluded studies in the interpretation of findings. Researchers should take caution in discussing studies that did not meet the inclusion criteria; for example, the search strategy might not have identified all studies that would have been excluded for the same reason, so the results of excluded studies may not be representative of the overall evidence.

Results of Individual Studies The results of individual studies should be organized and presented following methods described in the protocol. For each outcome, the report should present the results of each individual study including its average result (such as mean difference or relative risk) and a measure of precision (usually a 95 percent confidence interval).

If it is not possible to follow the protocol because studies made unanticipated comparisons, researchers should describe the changes from protocol and their potential effects on the overall evidence and conclusions. Researchers should also describe any included studies that measured important outcomes but failed to report adequate statistical information for meta-analysis.

Synthesis of Results Research syntheses add value by summarizing evidence and providing overviews of what is known, how much confidence can be placed in the available evidence, and what is not known. Particularly in large research syntheses, reporting the results of all included studies individually provides little benefit to the readers. The synthesis of extracted data may be done with or without meta-analysis, and results in final reports should be organized consistent with the outcomes defined in protocols.

In reviews with meta-analyses, results should be reported using the pre-specified summary measures, including the overall results (for example, mean difference), precision (for example, confidence interval), heterogeneity (for example, I^2), and statistical significance of the heterogeneity. For each outcome, the results of the individual studies and the overall synthesis may be reported together

in a forest plot. In reviews without meta-analyses, final reports should present the individual findings clearly (for example, in tables) and provide summaries of the overall results. In summarizing the results, researchers should be careful to consider the size and precision of the individual studies, as one would in a meta-analysis.

Subgroup analyses and sensitivity analyses should be reported for predefined categories. Any outcomes or analyses added post hoc should be identified and reported as exploratory analyses. Results for subgroup and sensitivity analysis should focus on the interactions between variables; reviews should not report effects for individual subgroups without also reporting results from statistical analyses examining differences between subgroups (that is, interactions).

Just as results in a study might vary across participants as a result of chance, some heterogeneity between studies would be expected by chance. The final report should describe the amount of observed heterogeneity in the results, including any statistical tests, and the potential sources and importance of those differences.

Risk of Bias The final report should describe threats to internal validity in each included study and an overall summary of the risk of bias (figure 21.1). This may be accomplished by providing a table showing the risk of bias for each item assessed in each included study and a figure showing an overall summary for each item.

The specific methods related to managing risk of bias may differ by type of intervention; for example, pharmacological and nonpharmacological interventions might use different methods to mask participants (Boutron et al. 2007). Moreover, in reviews with multiple outcomes, risk of bias might differ across them (for example, masking participants might affect their rating of quality of life but not mortality). The final report should describe the ways in which potential biases were addressed and the ways in which bias could have affected the results, including the likely direction of effects.

Finally, the report should describe potential biases in the review itself. For example, if some studies could not be included in the meta-analysis because eligible results were not published, the review might be vulnerable to reporting bias.

Overall Quality of the Evidence Some producers, including the Cochrane Collaboration, now ask researchers to rate their overall confidence in the body of evidence for key outcomes using the GRADE system (Brozek, Akl, Alonso-Coello, et al. 2009; Brozek, Akl, Jaeschke, et al. 2009). GRADE provides a framework for incorporating evidence about the precision and consistency of evidence, and it prompts researchers to consider risk of bias in the individual studies and risk of reporting bias. GRADE also helps researchers consider the *directness* of evidence (for example, if the outcomes relate to clinically important events or surrogates). For reviews of qualitative data, the Confidence in the Evidence from Reviews of Qualitative research (CERQual) provides a transparent method for assessing how much confidence to place in findings from a qualitative evidence synthesis (Lewin et al. 2015). If GRADE was used, the final report should include a summary of findings table.

21.4.7 Discussion

The discussion section provides an opportunity to consider the totality of the evidence from the research synthesis and

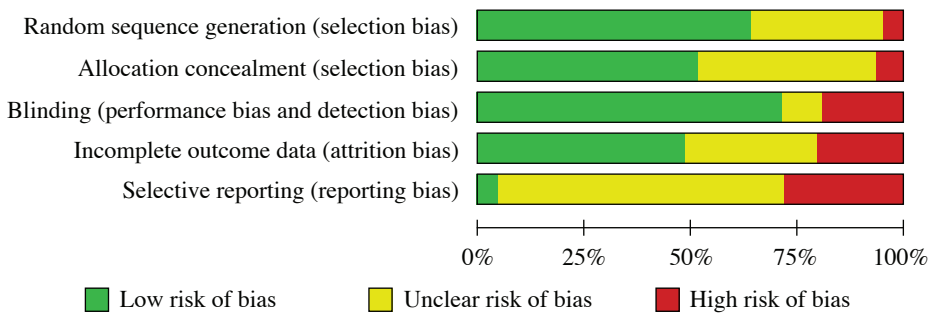

Figure 21.1 Summary of Risk of Bias in Included Studies

SOURCE: Authors' tabulation.

what it contributes to previous knowledge about the subject. It should normally begin with a summary of the review's findings, including the main results, a summary of confidence in the findings, and a description of their importance. For example, researchers might consider whether the results are clinically meaningful and whether they are likely to be affected by ongoing research in the field.

The discussion should identify to whom the results are likely to apply and under what conditions they are likely to be relevant (Nasser et al. 2012). For example, if the included studies focused on a particular subset of the eligible population, then the findings of the research synthesis might be more limited than the inclusion criteria would otherwise suggest.

Research syntheses about the same topic sometimes come to different conclusions because of differences in methods (for example, inclusion criteria) or because different researchers interpret the same evidence differently (Grant et al. 2016). The discussion should compare the research synthesis with similar research syntheses, including their methods, results, and inferences.

The discussion should identify any important limitations in the research synthesis. For example, the completeness of the review might be limited by restriction to English-language reports or by the dates of the searches. If the research synthesis sought to include unpublished data but did not acquire those data, or if it aimed to investigate mediators or moderators that were not reported consistently across studies, these limitations should be described. Researchers should also consider the potential effects of changes to the protocol.

Finally, the discussion should address the implications of the research synthesis for policy, practice, and future research. Many research syntheses conclude that more research is needed, which can be a vacuous claim. Given the results, the discussion should identify what specific types of research are needed to answer the most pressing questions that remain (Brown et al. 2006). For instance, use of the GRADE approach can identify whether future research needs to target important biases in the current literature (for example, investing resources to minimize participant attrition) or provide more data for certain subgroups to explain heterogeneity in current meta-analyses.

21.5 SHARING DATA AND CODE

Researchers should make the analysis plan and the data used for analysis publicly available to ensure the reproducibility (and validity) of reported analyses and to facilitate future research (Christensen 2016). Data sharing is rapidly becoming a scientific norm, and efforts are under way to facilitate sharing primary study data used in research syntheses (Wolfenden et al. 2016). Several resources exist to help researchers maintain documentation of the *workflow* of a research synthesis: that is, the steps for collecting, coding, organizing, and analyzing data (Gandrud 2013; Gentzkow and Shapiro 2014; Long 2008). For instance, it is often helpful to keep a record of each researcher's initial extraction and to create a third, reconciled record for each study or report. A file or database with the final reconciled data should be included with the final report. The exact code and dataset used for reporting analyses should be made freely and publicly available on the journal website or using services such as GitHub or the Open Science Framework—both of which support version control, or stored versions of all created files to facilitate easy comparison of file updates. The final code used to clean and analyze data should be annotated well enough to allow other researchers to clean and merge the original datasets and to reproduce the reported analyses. Programs such as RMarkdown can provide a single, shareable output (that is, HTML, PDF, or MS Word file) containing statistical code, comments describing each line of code, and the output from analyses through which another researcher could reproduce analyses in one click.

21.6 Additional Resources

This chapter focused on systematic reviews with meta-analyses that investigate the effects of interventions, however, there any many resources for other types of research syntheses as well. For instance, the meta-analysis of observational studies in epidemiology (MOOSE) guidelines describe methods for reporting syntheses of observational studies (Stroup et al. 2000) and the RAMESES (realist and meta-narrative evidence syntheses: evolving standards) project has produced reporting standards for realist syntheses (Wong et al. 2016) and meta-narratives (Wong et al. 2013). To facilitate user-friendliness, these checklists are organized in a format similar to the PRISMA statement and other established reporting guidelines. In addition, reporting standards for meta-ethnography, a specific type of qualitative synthesis, are currently under development (France et al. 2015). Last, several methodological standards may be useful for those conducting research syntheses, such as the Transparency and Openness Promotion guidelines, which include standards for the

production-to-dissemination cycle for any type of study (Nosek et al. 2015).

Reporting guidelines are useful for authors and peer-reviewers. Readers may also use related tools to assess the quality of research syntheses (Shea et al. 2007; Whiting et al. 2016).

Methods for conducting research syntheses are evolving rapidly, and new methods are emerging for conducting different types of reviews (for example, scoping reviews, rapid reviews). Before beginning a report or submitting a report for publication, researchers should check the library of reporting guidelines on the EQUATOR website for the most relevant and current guidelines.

21.7 CONCLUSION

Like the primary studies on which they rely, research syntheses may be more or less useful depending on their completeness and transparency. Editors increasingly demand that all reports follow guidelines for transparency in their final reports as well as earlier reports such as registrations and protocols (see box 21.1). Researchers

Box 21.1 Checklist for Research Synthesis Transparency

- Prospectively register and publish the protocol for the research synthesis, including at least following information:
  - Review title (stating the study is a research synthesis)
  - Anticipated or actual start and completion dates
  - Stage of research synthesis
  - Authors, their organizational affiliations, and their contact information
  - Funders, sponsors, and collaborators for research synthesis
  - Declarations of interests
  - Research questions, including a clear description of the participants, interventions, comparators, outcomes, and other items as appropriate
  - Eligibility criteria for studies and the process for assessing studies for eligibility
  - Information sources (such as electronic databases, contacting study authors, and bibliographies of identified studies) and a reproducible search strategy
  - Process for extracting data from included studies and the data to be extracted (such as primary and secondary outcomes, information related to risks of bias, and descriptive information)
  - Describe planned method of extracting data from reports (such as piloting forms, done independently, in duplicate), any processes for obtaining and confirming data from investigators
  - Primary and secondary outcomes
  - Process for synthesizing data, including quantitative and/or qualitative syntheses, sub-group and sensitivity analyses, and assessing the certainty of the evidence
  - Processes for managing records and data throughout the research synthesis
  - Identify updates to the registration and protocol as such, identifying changes from the last version
- Publish the completed report, containing at least the following information:
  - Reference to the registration and protocol, indicating any changes to methods described in the last version
  - Give numbers of studies screened, assessed for eligibility, and included in the review, ideally using a flow diagram
  - Characteristics of included studies, including risks of bias and individual study results
  - Results of all planned research syntheses
  - Summary of the main findings of the research synthesis, including the certainty of the evidence for each finding
  - Limitations of the individual studies, bodies of evidence for each finding, and research synthesis methods
  - Implications for future research, policy, and practice
- Post data and analytic code to a trusted, publicly and freely accessible repository

SOURCE: Authors' tabulation.

Box 21.2 Key Resources for Transparent Research Syntheses

- APA Meta-Analysis Reporting Standards: http://www.apastyle.org/manual/related/JARS-MARS.pdf
- Berkeley Initiative for Transparency in the Social Sciences (BITSS) Educational Resources: http://www.bitss.org/resource-tag/education/
- Cochrane Handbook for Systematic Reviews of Interventions: http://training.cochrane.org/handbook
- Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy: http://methods.cochrane.org/sdt/handbook-dta-reviews
- EQUATOR Network (reporting guidelines): http://www.equator-network.org/
- figshare (for sharing data): www.figshare.com
- GitHub (for sharing code): www.github.com
- Methodological Expectations of Cochrane Intervention Reviews (MECIR): http://editorial-unit.cochrane.org/mecir
- Open Science Framework: https://osf.io/
- PROSPERO (title registration): http://www.crd.york.ac.uk/PROSPERO/index.php
- RMarkdown: http://rmarkdown.rstudio.com/

SOURCE: Authors' tabulation.

should use relevant guidelines and tools for reporting their work at every stage in the research process (see box 21.2). Publishing protocols, and including information in supplemental files (that is, online), may help researchers describe their research transparently while adhering to limits on manuscript length.

21.8 REFERENCES

Agency for Healthcare Research and Quality. 2014. "Methods Guide for Effectiveness and Comparative Effectiveness Reviews." *AHRQ* publication no. 10*(*14*)*-EHC063-EF. Rockville, Md.: Agency for Healthcare Research and Quality.

APA Publications, and Communications Board Working Group on Journal Article Reporting Standards (APA). 2008. "Reporting Standards for Research in Psychology: Why Do We Need Them? What Might They Be?" *American Psychology* 63(9): 839–51. DOI: 10.1037/0003-066X.63.9.839.

Atkinson, Kayla M., Alison C. Koenka, Carmen E. Sanchez, Hannah Moshontz, and Harris M. Cooper. 2015. "Reporting Standards for Literature Searches and Report Inclusion Criteria: Making Research Syntheses More Transparent and Easy to Replicate." *Research Synthesis Methods* 6(1): 87–95. DOI: 10.1002/jrsm.1127.

Beller, Elaine M., Paul P. Glasziou, Douglas G. Altman, Sally Hopewell, Hilda Bastian, Iain Chalmers, Peter C. Gøtzsche, Toby Lasserson, and David Tovey. 2013. "PRISMA for Abstracts: Reporting Systematic Reviews in Journal and Conference Abstracts." *PLoS Medicine* 10(4): e1001419. DOI: 10.1371/journal.pmed.1001419.

Booth, Alison, Mike Clarke, Gordon Dooley, Davina Ghersi, David Moher, Mark Petticrew, and Lesley Stewart. 2012. "The Nuts and Bolts of PROSPERO: An International Prospective Register of Systematic Reviews." *Systematic Reviews* 1: 2. DOI: 10.1186/2046-4053-1-2.

———. 2013. "PROSPERO at One Year: An Evaluation of Its Utility." *Systematic Reviews* 2: 4. DOI: 10.1186/2046-4053-2-4.

Booth, Alison, Mike Clarke, Davina Ghersi, David Moher, Mark Petticrew, and Lesley Stewart. 2011a. "Establishing a Minimum Dataset for Prospective Registration of Systematic Reviews: An International Consultation." *PloS One* 6(11): e27319. DOI: 10.1371/journal.pone.0027319.

———. 2011b. "An International Registry of Systematic-Review Protocols." *Lancet* 377(9760): 108–109. DOI: 10.1016/S0140-6736(10)60903–8.

Boutron, Isabelle, Lydia Guittet, Candice Estellat, David Moher, Asbjørn Hrobjartsson, and Philippe Ravaud. 2007. "Reporting Methods of Blinding in Randomized Trials

Assessing Nonpharmacological Treatments." *PLoS Medicine* 4(2): e61. DOI: 10.1371/journal.pmed.0040061.

Brown, Polly, Klara Brunnhuber, Kalipso Chalkidou, Iain Chalmers, Mike Clarke, Mark Fenton, Carol Forbes, Julie Glanville, Nicholas J. Hicks, Janet Moody, Sara Twaddle, Hazim Timimi, and Pamela Young. 2006. "How to Formulate Research Recommendations." *British Medical Journal* 333(7572): 804–06. DOI: 10.1136/bmj.38987.492014.94.

Brozek, Jan L., Elie A. Akl, Paul Alonso-Coello, David M. Lang, R. Jaeschke, John W. Williams, et al. 2009. "Grading Quality of Evidence and Strength of Recommendations in Clinical Practice Guidelines. Part 1 of 3. An Overview of the Grade Approach and Grading Quality of Evidence About Interventions." *Allergy* 64(5): 669–77. DOI: 10.1111/j.1398–9995.2009.01973.x.

Brozek, Jan L., Elie A. Akl, R. Jaeschke, David M. Lang, Patrick Bossuyt, Paul P. Glasziou, et al. 2009. "Grading Quality of Evidence and Strength of Recommendations in Clinical Practice Guidelines: Part 2 of 3. The GRADE Approach to Grading Quality of Evidence About Diagnostic Tests and Strategies." *Allergy* 64(8): 1109–16. DOI: 10.1111/j.1398–9995.2009.02083.x.

Chan, An-Wen, Asbjørn Hrobjartsson, Mette T. Haahr, Peter C. Gøtzsche, and Douglas G. Altman. 2004. "Empirical Evidence for Selective Reporting of Outcomes in Randomized Trials: Comparison of Protocols to Published Articles." *Journal of the American Medical Association* 291(20): 2457–65. DOI: 10.1001/jama.291.20.2457.

Chandler, Jackie, Rachel Churchill, Julian P. T. Higgins, Toby Lasserson, and David Tovey. 2013. "Methodological Standards for the Conduct of Cochrane Intervention Reviews." Version 2.3. London: The Cochrane Library.

Christensen, Garrett. 2016. *Manual of Best Practices in Transparent Social Science Research*. Berkeley Initiative for Transparency in the Social Sciences. Berkeley: University of California. Accessed December 17, 2018. https://www.alnap.org/system/files/content/resource/files/main/manual.pdf.

Cuijpers, Pim, Annemieke van Straten, Ernst Bohlmeijer, Steven D. Hollon, and Gerhard Andersson. 2010. "The Effects of Psychotherapy for Adult Depression Are Overestimated: A Meta-Analysis of Study Quality and Effect Size." *Psychological Medicine* 40(2): 211–23. DOI: 10.1017/S0033291709006114.

Cybulski, Lukasz, Evan Mayo-Wilson, and Sean Grant. 2016. "Improving Transparency and Reproducibility Through Registration: The Status of Intervention Trials Published in Clinical Psychology Journals." *Journal of Consulting Clinical Psychology*. DOI: 10.1037/ccp0000115.

Deeks, Jonathan J., Patrick M. Bossuyt, and Constantine Gatsonis. 2010. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*, Version 1.0. London: The Cochrane Collaboration.

Duval, Sue, and Richard Tweedie. 2000. "Trim and Fill: A Simple Funnel-Plot-Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis." *Biometrics* 56(2): 455–63.

Dwan, Kerry M., Carrol Gamble, Paula R. Williamson, Jamie J. Kirkham, and Reporting Bias Group. 2013. "Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias—An Updated Review." *PloS One* 8(7): e66844. DOI: 10.1371/journal.pone.0066844.

Fleming, Padhraig S., Despina Koletsi, and Nikolaos Pandis. 2014. "Blinded by PRISMA: Are Systematic Reviewers Focusing on PRISMA and Ignoring Other Guidelines?" *PloS One* 9(5): e96407. DOI: 10.1371/journal.pone.0096407.

France, Emma F., Nicola Ring, Jane P. Noyes, Margaret Maxwell, Ruth Jepson, Edward Duncan, Ruth L. Turley, Derek Jones, and Isa Uny. 2015. "Protocol-Developing Meta-Ethnography Reporting Guidelines (eMERGe)." *BMC Medical Research Methodology* 15: 103. DOI: 10.1186/s12874-015-0068-0.

Gandrud, Christopher. 2013. *Reproducible Research with R and R Studio*. Boca Raton, Fl.: CRC Press.

Gentzkow, Matthew, and Jesse M. Shapiro. 2014. "Code and Data for the Social Sciences: A Practitioner's Guide." Mimeo, University of Chicago. Accessed December 17, 2018. https://web.stanford.edu/~gentzkow/research/CodeAndData.pdf.

Glasziou, Paul P., Douglas G. Altman, Patrick M. Bossuyt, Isabelle Boutron, Mike Clarke, S. Julious, et al. 2014. "Reducing Waste from Incomplete or Unusable Reports of Biomedical Research." *Lancet* 383(9913): 267–76. DOI: 10.1016/S0140-6736(13)62228-X.

Glasziou, Paul P., Malcolm Macleod, Iain Chalmers, John P. Ioannidis, Rustam Al-Shahi Salman, and An-Wen Chan. 2014. "Research: Increasing Value, Reducing Waste—Authors' Reply." *Lancet* 383(9923): 1126–27. DOI: 10.1016/S0140–6736(14)60563–8.

Goodman, Steven N., and Kay Dickersin. 2011. "Metabias: A Challenge for Comparative Effectiveness Research." *Annals of Internal Medicine* 155(1): 61–62. DOI: 10.7326/0003-4819-155-1-201107050-00010.

Goodman, Steven N., Daniele Fanelli, and John P. Ioannidis. 2016. "What Does Research Reproducibility Mean?" *Science Translational Medicine* 8(341): 341ps312. DOI: 10.1126/scitranslmed.aaf5027.

Grant, Sean, Eric R. Pedersen, Karen Cha Osilla, Magdelena Kulesza, and Elizabeth J. D'Amico. 2016. "Reviewing and Interpreting the Effects of Brief Alcohol Interventions: Comment on a Cochrane Review About Motivational Interviewing for Young Adults." *Addiction* 111(9): 1521–27. DOI: 10.1111/add.13136.

Higgins, Julian P. T., and Sally Green. 2011. *Cochrane Handbook for Systematic Reviews of Interventions.* Version 5.1.0 [updated March 2011]. London: The Cochrane Collaboration. Accessed December 17, 2018. http://handbook-5-1.cochrane.org.

Higgins, Julian P. T., C. Ramsay, Barnaby C. Reeves, Jonathan J. Deeks, Beverley Shea, Jeffrey C. Valentine, Peter Tugwell, and George A. Wells. 2013. "Issues Relating to Study Design and Risk of Bias When Including Non-Randomized Studies in Systematic Reviews on the Effects of Interventions." *Research Synthesis Methods* 4(1): 12–25. DOI: 10.1002/jrsm.1056.

Higgins, Julian P. T., Simon G. Thompson, Jonathan J. Deeks, and Douglas G. Altman. 2003. "Measuring Inconsistency in Meta-Analyses." *British Medical Journal* 327(7414): 557–60. DOI: 10.1136/bmj.327.7414.557.

Hutton, Brian, Georgia Salanti, Deborah M. Caldwell, Anna Chaimani, Christopher H. Schmid, Chris Cameron, John P. Ioannidis, Sharon Straus, Kristian Thorlund, Jeroen P. Jansen, Cynthia Mulrow, Ferrán Catalá-López, Peter C. Gøtzsche, Kay Dickersin, Isabelle Boutron, Douglas G. Altman, and David Moher. 2015. "The PRISMA Extension Statement for Reporting of Systematic Reviews Incorporating Network Meta-Analyses of Health Care Interventions: Checklist and Explanations." *Annals of Internal Medicine* 162(11): 777–84. DOI: 10.7326/M14–2385.

Imdad, Aamar, K. Herzer, Evan Mayo-Wilson, Mohammad Y. Yakoob, and Zulfiqar A. Bhutta. 2010. "Vitamin A Supplementation for Preventing Morbidity and Mortality in Children from 6 Months to 5 Years of Age." *Cochrane Database of Systematic Reviews* 12: CD008524. DOI: 10.1002/14651858.CD008524.pub2.

Institute of Medicine. 2011. *Finding What Works in Health Care: Standards for Systematic Reviews*. Washington, D.C.: National Academies Press.

Ioannidis, John P. 2016. "The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-Analyses." *Milbank Q* 94(3): 485–514. DOI: 10.1111/1468–0009.12210.

Katikireddi, Srinivasa V., Matt Egan, and Mark Petticrew. 2015. "How Do Systematic Reviews Incorporate Risk of Bias Assessments into the Synthesis of Evidence? A Methodological Study." *J Epidemiol Community Health* 69(2): 189–95. DOI: 10.1136/jech-2014–204711.

Kidwell, Mallory C., Ljilijana B. Lazarevic, Erica Baranski, Tom E. Hardwicke, Sarah Piechowski, Linba S. Falkenberg, Curtis Kennett, Agnieszka Slowik, Carina Sonnleitner, Chelsey Hess-Holden, Timothy M. Errington, Susann Fiedler, and Brian A. Nosek. 2016. "Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency." *PLoS Biol* 14(5): e1002456. DOI: 10.1371/journal.pbio.1002456.

Kirkham, Jamie J., Douglas G. Altman, and Paul R. Williamson. 2010. "Bias Due to Changes in Specified Outcomes During the Systematic Review Process." *PloS One* 5(3): e9810. DOI: 10.1371/journal.pone.0009810.

Lewin, Simon, Claire Glenton, Heather Munthe-Kaas, Benedicte Carlsen, Christopher J. Colvin, Metin Gülmezoglu, Jane Noyes, Andrew Booth, Ruth Garside, and Arash Rashidian. 2015. "Using Qualitative Evidence in Decision Making for Health and Social Interventions: An Approach to Assess Confidence in Findings from Qualitative Evidence Syntheses (Grade-Cerqual)." *PLoS Medicine* 12(10): e1001895. DOI: 10.1371/journal.pmed.1001895.

Li, Tianjing, Susan Hutfless, Daniël O. Scharfstein, Michael J. Daniels, Joseph W. Hogan, Roderick J. Little, Jason Roy, Andrew H. Law, and Kay Dickersin. 2014. "Standards Should Be Applied in the Prevention and Handling of Missing Data for Patient-Centered Outcomes Research: A Systematic Review and Expert Consensus." *Journal of Clinical Epidemiology* 67(1): 15–32.

Li, Tianjin, Swaroop Vedula, Nira Hadar, Christopher Parkin, Joseph Lau, and Kay Dickersin. 2015. "Innovations in Data Collection, Management, and Archiving for Systematic Reviews." *Annals of Internal Medicine* 162(4): 287–94. DOI: 10.7326/M14-1603.

Li, Tianjin, Swaroop Vedula, Roberta Scherer, and Kay Dickersin. 2012. "What Comparative Effectiveness Research Is Needed? A Framework for Using Guidelines and Systematic Reviews to Identify Evidence Gaps and Research Priorities." *Annals of Internal Medicine* 156(5): 367–77. DOI: 10.7326/0003-4819-156-5-201203060-00009.

Li, Tianjin, Tsung Yu, Barbara S. Hawkins, and Kay Dickersin. 2015. "Design, Analysis, and Reporting of Crossover Trials for Inclusion in a Meta-Analysis." *PloS One* 10(8), e0133023. DOI: 10.1371/journal.pone.0133023.

Liberati, Alessandro, Douglas G. Altman, Jennifer Tetzlaff, Cynthia Mulrow, Peter C. Gøtzsche, John P. A. Ioannidis, Mike Clarke, Patrick J. Devereaux, Jos Kleijnen, and David Moher. 2009. "The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That

Evaluate Health Care Interventions: Explanation and Elaboration." *PLoS Medicine* 6(7): e1000100. DOI: 10.1371/journal.pmed.1000100.

Liu, Min, Longxing Wei, and Jin Zhang. 2006. "Review of Guidelines and Literature for Handling Missing Data in Longitudinal Clinical Trials with a Case Study." *Pharmaceutical Statistics* 5(1): 7–18. DOI: 10.1002/pst.189.

Long, J. Scott. 2008. *The Workflow of Data Analysis Using Stata*. College Station, Tex.: Stata Press.

Lund, Hans, Carsten Juhl, and Robin Christensen. 2016. "Systematic Reviews and Research Waste." *Lancet* 387(10014): 123–24. DOI: 10.1016/S0140–6736(15)01354–9.

Manchikanti, Laxmaiah, Sukdeb Datta, Howard S. Smith, and Joshua A. Hirsch. 2009. "Evidence-Based Medicine, Systematic Reviews, and Guidelines in Interventional Pain Management: Part 6. Systematic Reviews and Meta-Analyses of Observational Studies." *Pain Physician* 12(5): 819–50.

Mathieu, Sylvaine, Isabelle Boutron, David Moher, Douglas G. Altman, and Philippe Ravaud. 2009. "Comparison of Registered and Published Primary Outcomes in Randomized Controlled Trials." *Journal of the American Medical Association* 302(9): 977–84. DOI: 10.1001/jama.2009.1242.

Moher, David, Paul P. Glasziou, Iain Chalmers, Mona Nasser, Patrick M. Bossuyt, Daniël A. Korevaar, Ian D. Graham, Philippe Ravaud, and Isabelle Boutron. 2016. "Increasing Value and Reducing Waste in Biomedical Research: Who's Listening?" *Lancet* 387(10027): 1573–86. DOI: 10.1016/S0140-6736(15)00307-4.

Moher, David, L. Shamseer, Mike Clarke, Davina Ghersi, Alessandro Liberati, Mark Petticrew, Paul Shekelle, Lesley Stewart, and PRISMA-P Group. 2015. "Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015 Statement." *Systematic Reviews* 4:1. DOI: 10.1186/2046-4053-4-1.

Montgomery, Paul, Kristen Underhill, Frances Gardner, Don Operario, and Evan Mayo-Wilson. 2013. "The Oxford Implementation Index: A New Tool for Incorporating Implementation Data into Systematic Reviews and Meta-Analyses." *Journal of Clinical Epidemiology* 66(8): 874–82. DOI: 10.1016/j.jclinepi.2013.03.006.

Nasser, Mona, Chris van Weel, Jaap J. van Binsbergen, and Floris A. van de Laar. 2012. "Generalizability of Systematic Reviews of the Effectiveness of Health Care Interventions to Primary Health Care: Concepts, Methods and Future Research." *Fam Pract* 29 *Suppl* 1: i94–i103. DOI: 10.1093/fampra/cmr129.

Nosek, Brian A., G. Alter, George C. Banks, Denny Borsboom, Sara D. Bowman, Steven J. Breckler, et al. 2015. "Scientific Standards. Promoting an Open Research Culture." *Science* 348(6242): 1422–25. DOI: 10.1126/science.aab2374.

Page, Matthew J., Julian P. T. Higgins, Gemma Clayton, Jonathan A. Sterne, Asbjørn Hrobjartsson, and Jelena Savovic. 2016. "Empirical Evidence of Study Design Biases in Randomized Trials: Systematic Review of Meta-Epidemiological Studies." *PloS One* 11(7): e0159267. DOI: 10.1371/journal.pone.0159267.

Page, Matthew J., Larissa Shamseer, Douglas G. Altman, Jennifer Tetzlaff, Margaret Sampson, Andrea C. Tricco, Ferrán Catalá-López, Lun Li, Emma K. Reid, Rafael Sarkis-Onofre, and David Moher. 2016. "Epidemiology and Reporting Characteristics of Systematic Reviews of Biomedical Research: A Cross-Sectional Study." *PLoS Medicine* 13(5): e1002028. DOI: 10.1371/journal.pmed.1002028.

Peters, Jasime L., Alex J. Sutton, David R. Jones, Keith R. Abrams, and Lesley Rushton. 2007. "Performance of the Trim and Fill Method in the Presence of Publication Bias and Between-Study Heterogeneity." *Statistics in Medicine* 26(25): 4544–62. DOI: 10.1002/sim.2889.

RevMan 5.3. 2014. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration.

Richardson, Marty, Paul Garner, and Sarah Donegan. 2016. "Cluster Randomised Trials in Cochrane Reviews: Evaluation of Methodological and Reporting Practice." *PloS One* 11(3): e0151818. DOI: 10.1371/journal.pone.0151818.

Saldanha, Ian J., Kay Dickersin, Xue Wang, and Tianjing Li. 2014. "Outcomes in Cochrane Systematic Reviews Addressing Four Common Eye Conditions: An Evaluation of Completeness and Comparability." *PloS one* 9(10): e109400. DOI: 10.1371/journal.pone.0109400.

Sanderson, Simon, Iaian D. Tatt, and Julian P. T. Higgins. 2007. "Tools for Assessing Quality and Susceptibility to Bias in Observational Studies in Epidemiology: A Systematic Review and Annotated Bibliography." *International Journal of Epidemiology* 36(3): 666–76. DOI: 10.1093/ije/dym018.

Shamliyan, Tatuama, Robert L. Kane, and Stacy Dickinson. 2010. "A Systematic Review of Tools Used to Assess the Quality of Observational Studies That Examine Incidence or Prevalence and Risk Factors for Diseases." *Journal of Clinical Epidemiology* 63(10): 1061–70. DOI: 10.1016/j.jclinepi.2010.04.014.

Shamseer, L., David Moher, Mike Clarke, Davina Ghersi, Alessandro Liberati, Mark Petticrew, Paul Shekelle, Lesley A. Stewart, and the PRISMA-P Group. 2015. "Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015: Elaboration and Explanation." *British Medical Journal* 349: g7647. DOI: 10.1136/bmj.g7647.

Shea, Beverley, Jeremy M. Grimshaw, George A. Wells, Maarten Boers, Neil Andersson, Candyce Hamel, Ashley C. Porter, Peter Tugwell, David Moher, and Lex M. Bouter. 2007. "Development of AMSTAR: A Measurement Tool to Assess the Methodological Quality of Systematic Reviews." *BMC Medical Research Methodology* 7: 10. DOI: 10.1186/1471-2288-7-10.

Silagy, Chris A., Philippa Middleton, and Sally Hopewell. 2002. "Publishing Protocols of Systematic Reviews: Comparing What Was Done to What Was Planned." *Journal of the American Medical Association* 287(21): 2831–34.

Squires, Janet E., Jeffrey C. Valentine, and Jeremy M. Grimshaw. 2013. "Systematic Reviews of Complex Interventions: Framing the Review Question." *Journal of Clinical Epidemiology* 66(11): 1215–22. DOI: 10.1016/j.jclinepi.2013.05.013.

Sterne, Jonathan A. C., and Matt Egger. 2001. "Funnel Plots for Detecting Bias in Meta-Analysis: Guidelines on Choice of Axis." *Journal of Clinical Epidemiology* 54(10): 1046–55.

Sterne, Jonathan A. C., Miguel A. Hernan, Barnaby C. Reeves, Jelena Savovic, Nancy D. Berkman, Meera Viswanathan, et al. 2016. "ROBINS-I: A Tool for Assessing Risk of Bias in Non-Randomised Studies of Interventions." *British Medical Journal* 355: i4919. DOI: 10.1136/bmj.i4919.

Stewart, Lesley A., Mike Clarke, Maroeska M. Rovers, Richard D. Riley, Mark Simmonds, Graema Stewart, and Jayne F. Tierney. 2015. "Preferred Reporting Items for Systematic Review and Meta-Analyses of Individual Participant Data: The PRISMA-IPD Statement." *Journal of the American Medical Association* 313(16): 1657–65. DOI: 10.1001/jama.2015.3656.

Straus, Sharon, and David Moher. 2010. "Registering Systematic Reviews." *Canadian Medical Association Journal* 182(1): 13–14. DOI: 10.1503/cmaj.081849.

Stroup, Donna F., Jesse A. Berlin, Sally C. Morton, Ingram Olkin, G. David Williamson, Drummon Rennie, David Moher, Betsy J. Becker, Theresa Ann Sipe, and Stephen B. Thacker. 2000. "Meta-Analysis of Observational Studies in Epidemiology: A Proposal for Reporting. Meta-Analysis of Observational Studies in Epidemiology (Moose) Group." *Journal of the American Medical Association* 283(15): 2008–12.

Tong, Allison, Kate Flemming, Elizabeth McInnes, Sandy Oliver, and Jonathan Craig. 2012. "Enhancing Transparency in Reporting the Synthesis of Qualitative Research: ENTREQ." *BMC Medical Research Methodology* 12: 181. DOI: 10.1186/1471-2288-12-181.

Turner, Lucy, Larissa Shamseer, Douglas G. Altman, Kenneth F. Schulz, and David Moher. 2012. "Does Use of the Consort Statement Impact the Completeness of Reporting of Randomised Controlled Trials Published in Medical Journals? a Cochrane Review." *Systematic Reviews* 1: 60. DOI: 10.1186/2046-4053-1-60.

Valentine, Jeffrey C., Ariel M. Aloe, and Timothy S. Lau. 2015. "Life After NHST: How to Describe Your Data Without 'p-ing' Everywhere." *Basic and Applied Social Psychology* 37(5): 260–73. DOI: 10.1080/01973533.2015.1060240.

Valentine, Jeffrey C., and Harris M. Cooper. 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(2): 130–49. DOI: 10.1037/1082-989X.13.2.130.

Valentine, Jeffrey C., Harris M. Cooper, Erika A. Patall, Diana Tyson, and Jorgianne C. Robinson. 2010. "A Method for Evaluating Research Syntheses: The Quality, Conclusions, and Consensus of 12 Syntheses of the Effects of After-School Programs." *Research Synthesis Methods* 1(1): 20–38. DOI: 10.1002/jrsm.3.

Welch, Vivian, Mark Petticrew, Peter Tugwell, David Moher, Jennifer O'Neill, Elizabeth Waters, and Howeard White. 2012. "PRISMA-Equity 2012 Extension: Reporting Guidelines for Systematic Reviews with a Focus on Health Equity." *PLoS Medicine* 9(10): e1001333. DOI: 10.1371/journal.pmed.1001333.

Wells, George A., Beverly J. Shea, Dianne O'Connell, J. Peterson, Vivian Welch, M. Losos, and Peter Tugwell. 2014. "The Newcastle-Ottawa Scale (NOS) for Assessing the Quality of Nonrandomised Studies in Meta-Analyses." Presentation. Ottawa: The Ottawa Hospital Institute. Accessed December 17, 2018. http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp.

Whiting, Penny F., Anne W. Rutjes, Marie E. Westwood, Susan Mallett, Jonathan J. Deeks, and Johannes B. Reitsma, Mariska Leeflang, Jonathan A. C. Sterne, Patrick M. M. Bossuyt, and QUADAS-2 Steering Group. 2011. "QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies." *Annals of Internal Medicine* 155(8): 529–36. DOI: 10.7326/0003-4819-155-8-201110180-00009.

Whiting, Penny F., Anne W. Rutjes, Marie E. Westwood, Susan Mallett, and QUADAS-2 Steering Group. 2013. "A Systematic Review Classifies Sources of Bias and Variation in Diagnostic Test Accuracy Studies." *Journal of Clinical Epidemiology* 66(10): 1093–104. DOI: 10.1016/j.jclinepi.2013.05.014.

Whiting, Penny F., Jelena Savovic, Julian P. T. Higgins, Deborah M. Caldwell, Barnaby Reeves, Beverley J. Shea, Philippa A. Davies, Jos Kleijnen, and Rachel Churchill. 2016. "ROBIS: A New Tool to Assess Risk of Bias in

Systematic Reviews Was Developed." *Journal of Clinical Epidemiology* 69 (January): 225–34. DOI: 10.1016/j.jclinepi.2015.06.005.

Wolfenden, Luke, Jeremy Grimshaw, Christopher M. Williams, and Sze Lin Yoong. 2016. "Time to Consider Sharing Data Extracted from Trials Included in Systematic Reviews." *Systematic Reviews* 5(1): 185. DOI: 10.1186/s13643-016-0361-y.

Wong, Geoff, Trish Greenhalgh, Gill Westhorp, Jeanette Buckingham, and Ray Pawson. 2013. "RAMESES Publication Standards: Meta-Narrative Reviews." *BMC Medicine* 11: 20. DOI: 10.1186/1741-7015-11-20.

Wong, Geoff, Gill Westhorp, Ana Manzano, Joanne Greenhalgh, Justin Jagosh, and Trish Greenhalgh. 2016. "RAMESES II Reporting Standards for Realist Evaluations." *BMC Medicine* 14(1): 96. DOI: 10.1186/s12916-016-0643-1.

Zarin, Deborah A., Tony Tse, Rebecca J. Williams, Robert M. Califf, and Nicholas C. Ide. 2011. "The ClinicalTrials.gov Results Database—Update and Key Issues." *New England Journal of Medicine* 364(9): 852–60. DOI: 10.1056/NEJMsa1012065.

Zorzela, Liliane, Yoone K. Loke, John P. Ioannidis, Su Golder, Pasqualina Santaguida, Douglas G. Altman, David Moher, Sunita Vohra, and PRISMA Harms Group. 2016. "PRISMA Harms Checklist: Improving Harms Reporting in Systematic Reviews." *British Medical Journal* 352: i157. DOI: 10.1136/bmj.i157.

22

THREATS TO THE VALIDITY OF GENERALIZED INFERENCES FROM RESEARCH SYNTHESES

GEORG E. MATT
San Diego State University

THOMAS D. COOK
Northwestern University

CONTENTS

22.1 INTRODUCTION

This chapter focuses on the major rationales for research synthesis and on the validity of the inferences that individual meta-analyses claim. We contend that meta-analysis provides a set of tools to facilitate the evolution of knowledge about the direction and strength of associations, especially causal ones, and about the conditions on which such associations depend and across which they continue to be robust. Relying on the specific formal meta-analytic assumptions detailed in previous chapters, we use a falsificationist framework that stresses the extent to which secure knowledge depends on ruling out alternative interpretations as opposed to amassing exact replication of the same association under the same conditions and using the same methods. To this end, we translate violations of the statistical assumptions undergirding meta-analysis into concretely labeled threats to valid inference that have to be ruled out if secure knowledge is to result.

The special promise of meta-analysis is to foster empirical knowledge about valid causal relationships and the conditions under which they are warranted. The hope is for causal knowledge that is general—that is therefore robust across a wide variety of circumstances or that is contingently specified so that the conditions are clear under which the size of the association varies. No rationale for meta-analysis is more important than its ability to identify the realms within, and over which valid causal knowledge holds. Does an association hold with specific populations of persons, settings, times and ways of varying the cause or measuring the effect as well as across different populations of people, settings, times and ways of operationalizing a cause and effect? Can it be extrapolated to other populations of people, settings, times, causes, and effects than those that have been studied to date? These generalization tasks are faced in all research, but perhaps most explicitly in meta-analysis.

Meta-analysis faces a special challenge, though. The past forty years of practice have amply demonstrated that the general inferences meta-analysts seek cannot depend on formal sampling theory alone. The primary studies available rarely present a census, or even a random sample, of all the populations, universes, categories, classes, or entities (terms we use interchangeably) relevant to a particular causal issue. The rare exception is when random sampling occurs from some clearly

designated universe, thus warranting valid generalization to the population from which the sample was drawn, usually a human population in the social sciences. This is a rare occurrence, and even then most studies take place at a single time, in quite restricted settings (a living room, for instance) and the relevant cause and effect constructs are purposively manipulated and measured without question of random selection from the domain of a causal agent or some possible impact. Moreover, many people are not interested in the population a particular random sample represents, asking instead, but does that same association hold with a different kind of person, in a different setting, at a different time, or with a different cause or effect? Such questions concern generalization as extrapolation from particular study samples rather than what the study samples "represent" in more general terms (Cook 1990, 2014). How can we extrapolate from studied populations to populations with many, few or even no overlapping attributes?

Warrants are needed for the general inferences meta-analysists seek other than in formal statistical sampling theory alone. This chapter assumes that ruling out threats to validity can serve as an important (though imperfect) warrant to justify generalized inferences. Doing so is not as simple or as elegant as sampling with known probability from a well-designated universe, but it is more flexible and has been used with success to justify how manipulations or measures are chosen to represent cause and effect constructs (that is, construct validity). If meta-analysis is to deal with generalization understood as using sample-level data to represent more general entities, or extrapolating from these samples to entities that are manifestly different from those sampled, or describing the range over which a relationship holds so as to test robustness or identify the contingencies under which a relationship does or does not hold, then we have to identify ways of using a particular data base to justify conclusions about the populations that its sample-level data represent, the conditions across which any demonstrated associations hold, and how the data can be used to extrapolate to other kinds of persons, settings, times, causes, and effects than those that have been directly studied to date.

This chapter is not the first to propose that a framework of validity threats allows us to probe the validity of research inferences when a fundamental statistical assumption has been violated. Donald Campbell (1957) introduced his internal validity threats for instances when primary studies lack random assignment, creating quasi-experimental design (Campbell and Stanley 1963) as a legitimate extension of the thinking R. A. Fisher had begun. Similarly, this chapter seeks to identify threats to valid inferences about generalization that arise in meta-analyses, particularly those that follow from infrequent random sampling in meta-analysis. Of course, Donald Campbell and Julian Stanley also had a list of threats to external validity, and these also have to do with generalization (1963). But their list was far from complete and was developed more with single primary studies in mind than research syntheses. More recently, frameworks for appraising primary studies have been proposed in the specific context of meta-analyses (see chapter 7, this volume) and systems have been develop to grade the combined evidence of systematic review, for instance, strength of recommendation taxonomy, or SORT (Ebell et al. 2004), and grading of recommendations assessment, development and evaluation, or GRADE (Guyatt et al. 2008). This chapter does not propose another system for appraising the quality of the evidence. Instead, it asks how can one proceed to justify claims about the generality of an association when the within-study selection of persons, settings, times, and measures is almost never random and when it is also not even reasonable to assume that the available sample of studies is itself an unbiased representation of the universe of existing studies about a particular association? The chapter proposes a threats-to-validity approach rooted in a theory of construct validity as one way to throw provisional light on how to justify general inferences.

22.1.1 Why We Conduct Research Syntheses

At the core of every research synthesis is an association about which we want to learn something that cannot be gleaned from a single existing study. These associations can be of many kinds, such as the relationship between a risk factor (such as secondhand smoke exposure) and a disease outcome (such as asthma severity), between a treatment (such as antidepressive medication) and an effect (such as negative affect), between a predictor (such as SAT) and a criterion (such as college GPA), or between a dose (such as hours of psychotherapy) and either a single response (say, psychological distress) or a response change (such as weight loss).

Many primary studies have relatively small sample sizes and thus low statistical power to detect a meaningful population effect. So it is commonly hoped that meta-analysis will increase the precision with which an association is estimated (Cohn and Becker 2003). The relevant intuition here is that combining estimates from multiple parallel studies

(or "exact replications") will increase the total number of primary units for analysis (for example, people), thus reducing the sampling error of an association estimate (Anderson and Maxwell 2016). However, parallel studies are rare and truly exact replications are impossible. More typical are studies of an association that differ in many ways, including how samples of persons were selected so that it is unreasonable to assume they are from the same population. So we need to investigate how inferences about associations are threatened when a meta-analysis is conducted. This is especially important when the topically relevant studies are few in number and heterogeneous in their samples of persons, settings, interventions, outcomes, and times. In our experience, most associations with which meta-analysts deal are causal, of the form that manipulating one entity (that is, treatment) will induce variability in the other (that is, outcome). So a second reason for conducting research syntheses is to examine how causal inference is threatened or enhanced in meta-analyses. A third rationale for research syntheses is to strengthen the generalization of an association (causal or otherwise) that has been examined across the available primary studies, each with its own time frame, ways of selecting persons and settings, and of implementing the treatment and outcome measures.

In meta-analytic practice, conclusions are rarely warranted via statistical sampling theory or even strong explanatory theories. Instead, heavy reliance is placed on less familiar principles for exploring generalizations about specific target universes, for identifying factors that might moderate and mediate an association, and for extrapolating beyond whatever universes are in the available data base (Cook 1990, 1993, 2014). We need to examine these other, less well-known principles to help identify the threats to validity that can be extracted from them.

22.1.2 Validity Threats

In the meta-analytic context, validity threats describe factors that can induce spurious inferences about the existence or size of an association, the likelihood it is causal, or its generalizability. Meta-analytic inferences are more compelling the better the identified threats have been ruled out either because they are empirically implausible or because other evidence indicates they did not operate in the research under consideration. Ruling out alternatives requires a framework of validity threats, some taken from formal statistical models, others from theories about generalization, and others emerging as empirical products born of critical reflection on the conduct of research syntheses. Any list of validity threats so constructed is bound to be subject to change. Indeed, such change is desirable, reflecting improvements in theories of method and in critical discourse about the practice of research synthesis. New threats should emerge as others die out because they are rare in actual research practice.

Some threats relevant to research syntheses also apply to individual primary studies (for example, unreliability in measuring outcomes). Others are extensions to threats that apply to primary studies, but are now reconfigured to reflect how a threat is distributed across studies (for example, the failure to find any studies that assign treatment at random). Other threats are unique to meta-analyses because they depend on how estimates of association are synthesized across single studies (for example, publication bias). These last have higher-order analogs in general principles of research design so that publication bias can be considered a particular instance of the more general threat of sampling bias. However, publication bias is application focused, one of the highly specific ways in which meta-analysts encounter sampling bias. So when two labels are possible we prefer to formulate a threat in the form closest to actual meta-analytic practice.

A list of validity threats can be used in two major ways. Retrospectively, it can help discipline claims emerging from a research synthesis by identifying which threats are plausible in light of current knowledge and by indicating the additional evidence needed to assess the plausibility of a threat and thereby evaluate how well the causal claim is justified. Secondly, validity threats can also be used prospectively for planning a research synthesis. The purpose then is to anticipate as many potential threats as possible and to design the research so as to rule them out. Identifying threats prospectively is better than retrospectively, though each is valuable.

22.1.3 Generalized Inferences

The generalized inferences possible in a research synthesis are different from those in a primary study because inferences in the latter are inextricably limited to the relatively narrow ways in which times, research participants, settings, and cause and effects are usually sampled or measured. In contrast, research syntheses involve multiple samples of diverse persons and settings as well as many treatment implementations and outcome measures, collected at unique times and over varying times. This potential for heterogeneous sampling provides the framework for generalizing to broader classes or universes than

a single study permits. But why is this, because the most secure path to generalization—formal sampling theory—does not generate the heterogeneity that meta-analysts typically find and use for generalization purposes? Interestingly, even advocates of random selection use purposive selection for choosing the items they put into their surveys; in so doing, they rely on theories of construct validity to justify generalizing from the items selected to the more general constructs these measures are intended to represent. Our theory of generalization in the meta-analytic context depends on factors ultimately derived from theories of construct validity rather than formal sampling. Each of these theories uses sampling particulars to justify inferences about abstract entities; construct validity uses systematic sampling to warrant inferences about the constructs a measure or manipulation represents whereas sampling theory uses random sampling. Our theory of generalization also relies on general principles of evolutionary epistemology according to which different studies examine variants of an association that help identify conditions that make and do not make a difference regarding the magnitude and direction of an association and that are being retained if they make a positive difference. How then does construct validity and evolutionary epistemology justify general inferences? Thomas Cook proposes five principles that are potentially useful for this purpose (1990, 1993).

22.1.3.1 Demonstrating Proximal Similarity In the 1980s, Donald Campbell first introduced the principle of proximal similarity in the context of construct validity (1986). For research syntheses, proximal similarity refers to the correspondence in attributes between an abstract entity (for example, an outcome construct, a class of settings) about which inferences are sought and the particular instances of the entity captured by the studies included in the meta-analysis. To make inferences about these abstract target entities, meta-analysts first seek to identify which of the potentially relevant studies are indeed members of the target class. That is, their attributes include those that theoretical analysis indicates are prototypical of the construct being targeted (Rosch 1973, 1978; Smith and Medin 1981). For example, if the target population is people—let us say Asian Americans in particular—then we need to know whether the participants have ancestors who come from certain nations. If we want to generalize to factories, we want to know whether the research actually took place in factory settings. Do the studies all occur in the late twentieth and early twenty-first centuries—the historical period to which generalization is sought? Are the interventions instances of the category labeled *decentralized decision making* in that all affected parties are represented in deliberations about policy? And, are the outcome operations plausible measures of the higher-order construct *productivity* in the sense that a physical output is measured and related to standards about quantity or quality? These kinds of questions about category membership are obvious and the sine qua non of any theory of representation. Alas, though, category membership justified by proximal similarity is not sufficient for inferring that the sampled instances represent the target class. More is needed.

22.1.3.2 Exploring Heterogeneous and Substantively Irrelevant Third Variables In his theory of construct validity, Campbell also required ruling out theoretical and methodological irrelevancies (1986). These are features of the sampled particulars that substantive experts deem irrelevant to a category but that can be part of the sampling particulars used to represent that category. Analysts must therefore have a theory of which aspects of a target entity are prototypical, or central to its understanding, as well as aspects that are peripheral and do not make much of a conceptual difference. If it can be shown that these substantive irrelevancies make no difference, then an association is deemed robust with respect to these features. Evidence concerning the robustness of an association is also of importance from an evolutionary perspective because those features make no positive difference and provide no adaptive advantage. As a result, the generalization is strengthened because the association being tested does not depend on largely or totally irrelevant features that happen to be correlated with the prototypical features. Research syntheses benefit from heterogeneous implementations across primary studies because this makes the irrelevancies associated with each construct heterogeneous. In the relatively rare case when a synthesis depends on primary studies that have the same substantive irrelevancy, then that attribute does not vary and is confounded with the target class. This presents a limiting scenario from an evolutionary perspective because it preempts new contingencies to be tested and new knowledge about robustness to evolve. For instance, if all the Asian Americans in a particular set of studies were Vietnamese or male, then the general Asian American category would be confounded with a single nation of origin or with a single gender. Similarly, if an association is limited to studies of public schools that are all in affluent suburbs with volunteer participants, then public schools are confounded with both the suburban location and the volunteer respondents. If all the measures of productivity tap into the quantity but not quality of what has been produced, then

the productivity concept would be confounded with its quantity component. The obvious requirement for unconfounding the targets of interest and the way they are assessed is ensuring that the studies sampled for meta-analysis vary in the country of origin among Asian Americans, in location among public schools, and in how productivity is measured.

22.1.3.3 Probing Discriminant Validity Even ruling out the role of theoretical irrelevancies is not enough. Any relationship that is not robust across substantively relevant differences invites the analyst to identify the conditions that moderate or mediate the association. Research syntheses include primary studies with their own populations of people, their own categories of setting, their own time period and their own theoretical descriptions of what a treatment or outcome class is and how it should be manipulated or measured. The issue is not then to generalize to a single common population; it is to probe how general the relationship is across populations and to determine the specific conditions under which it varies. This is also tied to general principles of evolutionary epistemology. Only when there is enough variation in the conditions examined in the primary studies is it possible to learn about how they might affect the magnitude and direction of an association. Does the effect of antidepressive medication hold in children, adolescents, adults, and the elderly? Is the effect of secondhand smoke different in pre- and postmenopausal women? Is the effect of psychotherapy conducted in clinical practice settings different from that in university research settings? Questions like these address the generalizability of meta-analytic findings by probing their discriminant validity. From an evolutionary perspective, it is the identification of moderator conditions that make a positive difference that produces general knowledge about the treatment, outcome, setting, and population variants that matter; that is, the conditions that affect the direction and magnitude of an association.

22.1.3.4 Studying Empirical Interpolation and Extrapolation The recent discussion deals only with situations where the synthesized studies contain instances of the classes about which we want to generalize. However, in some cases we seek to extrapolate beyond the available data and draw conclusions about domains not yet studied (Cronbach 1982; Cook 2014). This is an intrinsically difficult enterprise, but even here it is plausible to assert that research syntheses have advantages over primary studies. Consider a relationship that is demonstrably robust across a wide range of conditions. The analyst may then want to induce that the same relationship is likely to hold in other, as-yet unstudied, circumstances given that it has held so robustly over the wide variety of circumstances already examined. Consider, next, the case in which specific causal contingencies have been identified for which the association does and does not hold. This knowledge can then be used to help identify situations for which extrapolation is not warranted, the more so if a general theory can be adduced to explain the variation in effectiveness. Now, imagine a sample of studies that is relatively homogeneous on some or all attributes. There is then no evidence of the association's empirical robustness or of the specific circumstances that moderate when it is observed, making inductive extrapolation to other universes quite insecure. Of course, conceptual examination is always required to detect moderating variables, even with a heterogeneous collection of studies. For instance, many pharmaceutical drugs are submitted to regulatory agencies for approval with adults and may not have been tested with adolescents, young children, or the elderly. Yet individual physicians make off-label prescriptions where they extrapolate from approved groups (for example, adults) to nonstudied groups (such as adolescents, children, elderly). Indeed, some experts estimate that 80 to 90 percent of pediatric patients are prescribed drugs off-label, requiring thoughtful physicians to assume that the relevant physiological processes do not vary by age—an assumption that may or may not be true for a given drug (Tabarrok 2000; Brauner et al. 2016; Arocas Casañ et al. 2016). Extrapolation depends here not just on the empirical robustness of an association across the studied age groups, but also on substantive theory about age differences in physiological processes.

22.1.3.5 Building on Causal Explanations In the experimental sciences, generalization is justified not from sampling theory but instead from attempts to achieve complete understanding of the processes causally mediating an association. The crucial assumption here is that, once these processes have been identified, they can then be set in motion by multiple causal agents, including some not yet studied but plausible on theoretical or pragmatic grounds. Indeed, the general public is often asked to pay for basic science on the grounds that ameliorative processes will be discovered that can be instantiated in many ways, making them viable across a wide range of persons and settings today and tomorrow. Research syntheses have the potential to enhance our understanding of conditions mediating the magnitude and direction of an association because the variety of persons, settings, times and measures available in the data base will be greater than that which any

single study can capture. The crucial issue then becomes what the obstacles (that is, threats) are that arise when using meta-analysis to extrapolate findings by identifying the processes by which a class of interventions produces an effect? This is a difficult issue to pursue because explicit examples of causal explanation remain rare in meta-analysis because the concern has been more with identifying causal relationships and some of their moderating rather than mediating factors (see, for example, Becker 2001; Cook et al. 1992; Harris and Rosenthal 1985; Premack and Hunter 1988; Shadish and Sweeney 1991). But even so, moderation provides clues to mediation through the way results are patterned; and many meta-analyses do include individual studies making claims about causal mediation. It would be stretching the point to claim that detecting causal mediation is a strength of meta-analysis today, and it remains to be seen whether the potential for explanation in meta-analyses can be realized (Cook 1992).

The five principles we have just enumerated for strengthening generalized inference in meta-analyses are not independent; they are invoked simply because they describe current efforts at generalization within a synthesis framework and because they are linked to efforts to warrant generalization via theories of construct validity and evolutionary epistemology rather than formal sampling or well corroborated and well demarcated substantive theory. Taken together, the principles suggest feasible strategies for exploring and justifying conclusions about the generalization of an association. However, they are more relevant for conclusions about what the sampling particulars in a set of studies represent and what the limits of generalization might be than they are for the intrinsically more problematic task of extrapolating to unstudied entities.

The rest of the discussion in this chapter follows from the benefits promised by research synthesis in regard to threats: those relevant to meta-analyses that primarily seek to describe the degree of association between two variables; those to the inference that a relationship is causal in the manipulability or activity theory sense (Collingwood 1940; Gasking 1955; Whitbeck 1977); generalization, beginning with validity threats that apply when generalizing to particular target populations, constructs, or categories; generalization pertinent to moderator variable analyses that concern the empirical robustness of an association and thus affect its strength and even direction (Mackie 1974); and finally generalization threats that apply when meta-analysis is used to extrapolate to novel, as-yet-unstudied universes. Threats discussed earlier almost always apply later in regard to the increasingly more complex inferential tasks of causation and generalization. To avoid repetition, we discuss only the unique threats relevant to the later sections.

22.2 THREATS TO INFERENCES ABOUT THE EXISTENCE OF A TREATMENT-OUTCOME ASSOCIATION

The threats discussed here deal with issues that may lead to erroneous conclusions about the existence of a relationship between two classes of variables, including treatment (that is, cause) and outcome variables (that is, effect). In hypothesis testing language, these threats lead to type I or type II errors arising from deficiencies in either the primary studies or the research synthesis process. A solitary deficiency in a single study is unlikely to jeopardize meta-analytic conclusions in any meaningful way. More problematic is whether a specific deficiency operates across all or most of the studies being reviewed or whether different deficiencies fail to cancel each other out across studies such that one direction of bias predominates, either to over- or underestimate an association (see chapter 7). The need to rule out these possibilities leads to the list of validity threats presented in table 22.1.

22.2.1 Unreliability in Primary Studies

Unreliability in implementing or measuring treatments and in measuring outcomes attenuates the effect-size estimates from primary studies as well as the average effect-size estimates computed across such studies. Attenuation corrections have been proposed and, if their assumptions

Table 22.1 Threats to Inferences About the Existence of a Treatment-Outcome Association

1. Unreliability in primary studies
2. Restriction of range in primary studies
3. Missing effect sizes in primary studies
4. Unreliability of codings in meta-analyses
5. Capitalizing on chance in meta-analyses
6. Biased effect-size sampling
7. Publication bias
8. Bias in computing effect sizes
9. Lack of statistical independence among effect sizes
10. Failure to weight effect sizes proportional to precision
11. Underjustified use of fixed- or random-effects models
12. Lack of statistical power for detecting an association

SOURCE: Authors' compilation.

are accepted, may be used to provide estimates that approximate what would have happened had treatments and outcomes been observed without error, or were the error constant across studies rather than variable (Hunter and Schmidt 2004; Rosenthal 1991; Wolf 1990; Salgado, Moscoso, and Anderson 2016); also see chapter 15). However, there are limits to what can be achieved because the implementation of interventions is less understood than the measurement of outcomes, and may not be documented well in primary studies. Some primary researchers do not report the reliability of the scores on the outcomes they measured. These practical limitations impede comprehensive attempts to correct individual effect-size estimates, and force some meta-analysts either to ignore the impact of unreliability or to rely on (untested) assumptions about what the missing reliability would have been in a given study. For these reasons, attenuation corrections remain a controversial approach to dealing with unreliability in primary studies.

22.2.2 Restriction of Range in Primary Studies

When the range of a treatment or predictor variable is restricted (for example, the treatment and control conditions differ little from each other; low dosage treatment), the restriction reduces effect-size estimates relative to what would have been observed had a wider range been studied. In contrast, if the outcome variable range is restricted, the within-group variability will be reduced relative to a study that had included a wider range of persons unless measured statistical controls accounted for all of this person variability. Restricting range on the outcome will otherwise decrease the denominator of the effect-size estimate, *d*, and thus increase effect-size estimates. For example, if a weight-loss researcher limits participation to subjects who are at least 25 percent but not more than 50 percent overweight, this strategy will likely result in less within-group variability in weight loss than were all overweight subjects included, increasing the effect-size estimate by decreasing its denominator. Thus, range restrictions in primary studies can attenuate or inflate effect-size estimates, depending on which variable is restricted. As Wolfgang Viechtbauer points out, two standardized effect sizes based on the same outcome measures from two studies could be incommensurable if the samples were drawn from populations with different variances (2007). This problem can be avoided if the raw units of outcome measures are consistent across the studies of a meta-analysis, making it unnecessary to standardize the effect size. For instance, this was the case when Jean Twenge and Keith Campbell used the Rosenberg self-esteem scale and the Coopersmith self-esteem inventory to conduct a meta-analysis of birth cohort differences in adults and children, respectively (2001; see also Le and Schmidt 2006).

Given the possibility of counterbalancing range restriction effects within individual studies and across all the studies in a research synthesis, it is not easy to estimate the overall impact of range restriction. Nonetheless, meta-analysts have suggested adjustments to mean estimates of correlations (and their standard error) to control for the attenuation due to different types of range restriction (Hunter and Schmidt 2004; Hunter, Schmidt, and Le 2006; Le and Schmidt 2006; see also chapter 15). When valid estimates of the population range or variance are available, these adjustments provide reasonable estimates of the correlation that would be observed in a population with such variance. However, it is more problematic to adjust for range restriction in a manipulated variable (that is, a carefully planned intervention) for which population values may not be known.

22.2.3 Missing Effect Sizes in Primary Studies

Missing effect sizes occur when study reports fail to include findings for all the sampled groups, outcome measures, or times. This sometimes happens because space limitations in a journal prevent reporting results in full detail, although the increased availability of online supplemental materials makes this reason less plausible. Missing effect size may also come about when research reports focus on only a part of the overall study, or when authors decide to report only a subset of their findings (for example, short-term effects, treatment completers, or placebo-controls). Some effect sizes also go unreported because the findings were not statistically significant, thus inflating the average effect size from a meta-analysis. Except for the last example, the impact of unreported effect sizes will vary, depending on why an author decided to include some findings and exclude others.

To prevent this kind of bias, it is desirable to code the most complete version of a report (for example, dissertations and technical reports rather than published articles) and to contact study authors to obtain additional information, as Steven Premack and John Hunter (1988) and George Kelley, Kristi Kelley, and Zung Vu Tran (2004) did with some success. However, this last strategy is not feasible if authors cannot be located or they have discarded the

needed information. Meta-analysts must then code research reports for evidence of missing effect sizes and use sensitivity analyses to explore how the data known to be missing might have affected overall findings. This can involve examining whether studies with little or no missing data yield comparable findings to studies with more missing data or missing data that is patterned a particular way.

Another approach is to pursue the imputation strategies discussed in chapter 17 of this book, though it is clear that they hold only under certain model assumptions. When these assumptions cannot be convincingly justified, the impact of these procedures must then be questioned. An example of this comes from early meta-analyses of the effects of psychotherapy, where it was assumed that the most likely estimate for a missing effect size was zero, based on the assertion that researchers failed to report only those outcomes that did not differ from zero (that is, were not statistically significant). Although this assumption has never been comprehensively tested, David Shapiro and Diana Shapiro coded 540 unreported effect sizes as zero and then added them to 1,828 reported effects, reducing the average effect from 0.93 to 0.72 (1982). But effects sizes may be unreported for many other reasons—for example, when they are negative, or positive but associated with unreliable measures—raising concern about the assumption that they average zero (for a general introduction to multiple imputation methods, see Little and Rubin 2002; Rubin 1987; Sinharay, Stern, and Russell 2001; for applications in meta-analysis, see Sutton 2000; Kelley 2004; Shadish et al. 1998).

22.2.4 Unreliability of Codings in Meta-Analyses

Meta-analytic data are the product of a coding process susceptible to human error. Unreliability at the level of research synthesis (such as unreliable determination of means, standard deviations, or sample sizes) is not expected to bias average effect-size estimates. This is because in classical measurement theory, measurement error is independently distributed and uncorrelated with true scores. However, measurement error will increase the variance of the observed effect sizes, increasing estimates of standard error and reducing statistical power for hypothesis tests. Chapter 9 in this volume discusses several strategies for controlling and reducing error in coding. In our experience pilot testing the coding protocol, comprehensive coder training, engaging coders with expertise in the substantive area being reviewed, consulting external literature, contacting primary authors, using reliability estimates as controls, generating confidence ratings for individual codings, and conducting sensitivity analyses are all helpful strategies for reducing and controlling for error in data coding

22.2.5 Capitalizing on Chance in Meta-Analyses

Although research syntheses may combine findings from hundreds of studies and thousands of individual respondents, they are not immune to inflated type I error when many statistical tests are conducted without adequate control for error rate—a problem that is exacerbated in research syntheses with few studies. Typically, meta-analysts conduct many analyses as they probe the robustness of an effect size across various methodological and substantive characteristics that might moderate effect sizes, and as they otherwise explore the data. To reduce capitalizing on chance, researchers must adjust error rates, examine families of hypotheses in multivariate analyses, or stick to a small number of a priori hypotheses.

22.2.6 Biased Effect-Size Sampling

Research reports frequently present more than one estimate, especially when multiple outcome measures, multiple treatment and control groups, and multiple delayed assessment time points are involved. Some of these effect estimates may be irrelevant for a particular topic, and some of the relevant ones will be substantively more important than others. Meta-analysts must then decide which estimates will enter the meta-analysis. Bias occurs when estimates are selected that are as substantively relevant as those not selected but that have different average effect sizes. Georg Matt (1989) discovered this when three independent coders recoded a subsample of the studies in Mary Lee Smith, Gene Glass, and Thomas Miller's (1980) meta-analysis of the benefits of psychotherapy. Following what seemed to be the same rules as Smith and colleagues for selecting and calculating effect estimates, the recoders extracted almost three times as many effect estimates whose mean effect size was approximately 0.50 against the 0.90 from the original codings. Rules are required in each meta-analysis that clearly indicate which effect estimates to include. Best practice is to specify these rules before data collection begins. If several plausible rules are identified, it is then important to examine whether they lead to the same results. Specifying such rules is relatively easy, but implementing them validly (and reliably) depends on training coders to identify relevant effect sizes

within studies. Data analyses should also explore for possible coder differences in results.

22.2.7 Publication Bias

Some studies are conducted but never written up; of those written up, some are not submitted for publication; of those submitted, some are never published; and of those published, some are difficult to find and retrieve. Publication bias exists when the average effect estimate from published studies differs from that of the population of studies ever conducted on the topic. Anthony Greenwald (1975) and Robert Rosenthal (1979) both argue that published studies in the behavioral and social sciences are likely to be a biased sample of all the studies actually carried out, because studies with statistically significant findings supporting a study's hypotheses are more likely to be submitted for publication and ultimately published, given reviewer and editor biases against the null hypothesis. An extension of this bias may even operate among published studies if those that are easier to identify and retrieve have different effect sizes than other studies. This could arise because of the greater visibility of studies in "major" publications that are abstracted clearly by the major referencing services, because authors in major outlets are better connected in professional networks, or because the paper's title or abstract is more likely to contain keywords relevant for a particular meta-analysis. Unsuccessful replications by lesser known researchers are not likely to appear in major journals, however appropriate they might be for a conference or a more obscure publication.

Strenuous attempts should be made to find unpublished or difficult-to-retrieve studies, and separate effect-size estimates should be calculated for published and unpublished studies as well as for studies that varied in how difficult they were to locate. In a review of thirty-five meta-analyses of randomized trials, Ikhlaaq Ahmed, Alexander Sutton, and Richard Riley find that only ten discussed or investigated the potential for publication bias (2012). The chapters in this book on scientific communication (chapter 4), reference databases (chapter 5), grey literature (chapter 6), and publication bias (chapter 18) all provide additional suggestions for dealing with publication bias (see also Rothstein, Sutton, and Borenstein 2005).

Several developments are especially important. First, research registries are particularly promising for avoiding publication biases, especially in research domains under the influence of regulatory agencies (such as the Food and Drug Administration). Second, the Cochrane Collaboration has taken the idea of research registries one step further, developing a registry for prospective research syntheses, in which eligible studies for a meta-analysis are identified and evaluated before their findings a re e ven k nown. T hird, meta-analysts can now rely on a set of powerful exploratory data analysis methods to detect biases in published studies (Egger et al. 1997; Sterne and Egger 2001; Sutton, Duval, et al. 2000; Sutton, Song, et al. 2000; Peters et al. 2006). Finally, if the evidence indicates that effect estimates depend on publication history, the likely consequences of this bias should be assessed with a variety of methods, including Larry Hedges and Jack Vevea's selection method approach (Vevea and Hedges 1995; Hedges and Vevea 1996; Vevea, Clements, and Hedges 1993; Vevea and Woods 2005), and Susan Duval's trim and fill method (Duval and Tweedie 2000; Sutton, Duval, et al. 2000).

22.2.8 Bias in Computing Effect Sizes

Meta-analyses typically require transforming findings from primary studies into a common metric such as the correlation coefficient, a standardized mean difference, or an odds ratio (see chapter 11). This is necessary because studies differ in the type of quantitative information they originally provide, and bias results if some types of transformation lead to systematically different estimates of average effect size or standard error when compared to others.

The best understood case of transformation bias concerns the situation in which probability levels are aggregated across studies and some have been truncated (Rosenthal 1990; Wolf 1990), for instance, when a study reports a group difference statistically significant at $p < .05$ without specifying the exact probability level. A conservative estimate of the observed effect size can be obtained by assuming that $p = .05$ then finding the relevant critical value of the test statistic that would have been observed (given the appropriate degrees of freedom). Because it is known that the actual probability value was smaller than the one assumed, the transformed effect size is known to be conservatively biased.

Michael Rosenberg points out a potential bias in the conversion of χ^2 statistics to correlations if the expected values in treatment and control groups or not equal (2010). Another case of potential bias involves studies that used dichotomous outcome measures or continuous variables that have been dichotomized (for example, improved versus not improved, convicted versus not convicted). Julio Sánchez-Meca, Fulgencio Marín-Martínez, and Salvador Chacon-Moscoso compare seven approaches to convert-

ing effect sizes derived from continuous and dichotomous outcome variables into a common metric, concluding that the Cox and Probit–based effect-size indices showed the least bias across simulated population effect sizes, δ, ranging from 0.2 to 0.8 (2003). Whenever possible, meta-analysts are advised to calculate effect sizes directly from means, standard deviations, or the like rather than from approximations such as truncated *p*-levels or proportions improved. This will be possible for many studies, though not all. Simulation studies of the performance of different indicators like that of Sánchez-Meca and his colleagues can help make an informed choice, and additional tools for retrieving accurate effect size estimates from incomplete summary data in primary studies are needed (Olkin 2012). But when they are not available, meta-analysts should empirically examine whether estimates differ by the type of effect-size transformation used.

22.2.9 Lack of Statistical Independence Among Effect Sizes

Stochastic dependencies among effect sizes may influence average effect estimates and their precision. The effect-size estimates in a meta-analysis may lack statistical independence for at least four reasons: collecting data on multiple outcomes for the same respondents; comparing different interventions to a single control group, or different controls to a single intervention; calculating an effect estimate for each of several subsamples of person within the same study (for example, women and men); and the same research team conducts multiple studies on the same topic (Hedges 1990). These situations can be conceptualized hierarchically—for instance as multiple outcomes or repeated assessments of the same measure nested within each study. Ignoring or misspecifying the resulting covariance structure of effect sizes can lead to invalid estimates of mean effect sizes and their standard errors (see also chapter 13, this volume).

There are several approaches to dealing with such dependencies. The most simple involves analyzing for each study only one of the set of possible correlated effects, for example, the mean or median of all effects, a randomly selected effect estimate, or the most theoretically relevant estimate (Lipsey and Wilson 2001). Another involves Bonferroni and "ensemble adjusted" *p*-values. Although these strategies are relatively simple to apply, each is conservative and fails to take into account all of the available data. Larry Hedges and Ingram Olkin therefore developed a multivariate statistical framework in which dependencies can be directly modeled (1985; see also Raudenbush, Becker, and Kalaian 1988; Rosenthal and Rubin 1986). More recently, Larry Hedges, Elizabeth Tipton, and Matthew Johnson developed a robust estimator of the covariance matrix of meta-regression coefficients for within-study correlated effect-size estimates (2010), and Emily Tanner-Smith and Elizabeth Tipton (2014) develop and tested corresponding software macros. Bayesian, hierarchical linear (HLM), and structural equation (SEM) modeling approaches have also been successfully applied to deal with multiple effect sizes within studies (Raudenbush and Bryk 1985; van Houwelingen, Arends, and Stijnen 2002; Saleh et al. 2006; Eddy, Hasselblad, and Shachter 1990; Sutton and Abrams 2001; Scott et al. 2007; Nam, Mengersen, and Garthwaite 2003; Hox and Leeuw 2003; Prevost, Abrams, and Jones 2000; Cheung 2008, 2014; Cheung and Chan 2005; see also chapter 13, this volume). These multivariate models are more complex than the previously discussed univariate techniques, and some require estimates of the covariance structure among the correlated effect sizes that may be difficult to obtain, in part because of missing information in primary studies. Moreover, the gains in estimation due to using these multivariate techniques may be small.

22.2.10 Failure to Weight Effect Sizes Proportional to Their Precision

Everything else being equal, studies with larger sample sizes yield effect-size estimate that are more precise (that is, smaller standard errors) than studies with smaller ones. Simply averaging effect sizes of different precision may yield biased average effects and sampling errors even if each study's estimates are themselves unbiased. Therefore, when effect sizes are combined, it has become common practice to use weighted averages, allowing more precise estimates to have a stronger influence on the overall findings. Larry Hedges shows that the optimal weight is the inverse of the sampling variance of an effect size (1982; Hedges and Vevea 1998). Michael Brannick and his colleagues and Fulgencio Marín-Martínez and Julio Sánchez-Meca demonstrate this outcome in a Monte Carlo simulation for random-effects meta-analysis under scenarios common in organizational research (Brannick, Yang, and Cafri 2011; Marín-Martínez and Sánchez-Meca 2010). This validity threat was common in early meta-analyses of psychotherapy outcomes but has become increasingly rare in recent years in meta-analyses using common effect-size metrics, such as d, r, and odds ratios (Shapiro and Shapiro 1982; Smith, Glass, and Miller 1980). Although

this remains controversial, there appears to be continued interest in overall effect estimates based on effect-size weights that take into account differences between studies in their methodological quality (Doi et al. 2015; Doi and Thalib 2009, 2008).

22.2.11 Underjustified Use of Fixed- or Random-Effects Models

When analyzing effect sizes, Hedges and Olkin (1985) and Hedges and Vevea (1998) stress the importance of deciding on a model with fixed or random effects. Perhaps the most important difference between the two models concerns the inferences they allow. In its simplest form, a fixed-effect model assumes that all the studies in a meta-analysis involve independent samples of participants from the same population, and so it is legitimate to postulate a single but unknown effect-size parameter. If that is the case, each study in a meta-analysis provides an independent estimate of this unknown effect-size parameter, estimates of which differ only as a result of sampling variability. In a more complex fixed-effects model, effect sizes may differ between fixed groups or levels such that each study within a group or level provides an estimate of that unknown effect-size parameter. In contrast, random-effects models assume that each study has its own unique "true" effect and to be a sample realization from a universe of related, yet distinct effects. Thus, observed differences in treatment effects have two sources: different samples of participants (as in the fixed-effect model) and true differences in treatment effects between studies (relative to a single true effect size in the fixed model). This means that in the random-effects model, treatment effects are best represented as a distribution of true effects represented by their expected value and variance.

These two models have important implications for the analysis and interpretation of effect-size estimates (Hedges and Vevea 1998; see chapter 12, this volume). The fixed-effect model is analytically simpler, requiring estimates of the specific fixed effects of interest (for example, mean effect sizes for treatment A, B, and C) and their precision (that is, standard error). The random-effects model requires estimating the population variance associated with the universe of treatment effects given the estimates observed in a sample of treatments and their precision. As a consequence of the different underlying statistical models, fixed-effects analyses limit inferences to the specific fixed levels of a factor that were included in a meta-analysis (for example, treatments A, B, and C). In contrast, random-effects models strive for inferences about the population of levels of a factor (for example, population of treatments consisting of A, B, C, F, . . . , Z), given the samples of levels that were included in a meta-analysis (for example, sample of treatments consisting of C, G, M, and T).

The decision to assume a fixed- or random-effects model is primarily influenced by the substantive assumptions meta-analysts make about the processes generating an effect and about the desired inferences. That is, are different studies estimating a single common fixed effect or do the effects vary between studies? Are the treatments, settings, outcomes, and participants examined in different studies sufficiently standardized and similar (that is, fixed) that they should be considered equivalent for purposes of interpreting treatment effects? Are we only interested in drawing inferences about the specific instances of treatments, settings, outcomes, and participants that were studied (that is, fixed-effect model)? Or, are we interested in drawing more general inferences about the classes of treatments, settings, and outcomes to which the specific instances that were studied belong (that is, random-effect model)? Decisions about fixed- or random-effects models should not be influenced by the heterogeneity observed in the data and the results of a homogeneity test. This is because such tests cannot provide meaningful results if they are applied to a meta-analysis of studies estimating the same fixed effects (for instance, the speed of light). Heterogeneity tests should also be used cautiously in meta-analyses of a random effects (for instance, the effects of supplemental instruction in introductory college courses). This is because tests of homogeneity tend to have low statistical power in situations commonly found in research syntheses, for example, modest heterogeneity, unequal sample sizes; small within-study sample sizes (Hardy and Thompson 1998; Harwell 1997; Hedges and Pigott 2001; Jackson 2006). In summary, there is no simple indicator of which model is correct, and the choice of a fixed- or random-effects model should be based on our understanding of the nature of the effect and our understanding of how the included primary studies sample the effects.

22.2.12 Lack of Statistical Power for Detecting an Association

Although the focus of meta-analyses is on estimating the magnitude of effects and their precision rather than null hypothesis testing, meta-analysts often report findings from hypothesis tests about the existence of an association. Under most circumstances, the statistical power for

detecting an association in a meta-analysis is influenced by the number of studies, the sample sizes within studies, the type of assignment of units to experimental conditions in the primary studies, such as cluster randomization, and, for analyses employing random-effects assumptions, the between-studies variance component (Hedges and Pigott 2001; Donner, Piaggio, and Villar 2003; Cohn and Becker 2003). When compared with statistical analyses in primary studies, tests of mean effect sizes will typically be more powerful in meta-analyses, particularly in fixed-effect models estimating the average effect of a class of interventions based on many similar studies. There are some paradoxical exceptions to this rule in random-effect models, for instance, when small studies add more between-studies variance than they compensate for by adding information about the mean effect (Hedges and Pigott 2001). Similarly, research syntheses of cluster randomization trials must take design effects into account to obtain correct estimates of statistical power (Donner, Piaggio, and Villar 2001, 2003).

Some meta-analysts are limited to a few studies, each having small sample sizes. Other meta-analysts are interested in examining effect sizes for subclasses of treatments and outcomes, different types of settings, and different subpopulations. Careful power analyses can help clarify the type II error rate of a test failing to reject a null hypothesis, and Jeffrey Valentine, Therese Pigott, and Hannah Rothstein (2010) provide a useful primer on statistical power for meta-analysis for fixed and random models. The meta-analyst then has to decide which trade-off to make between the number and type of statistical tests and the statistical power of these tests.

22.3 THREATS TO INFERENCES ABOUT THE CAUSAL NATURE OF THE TREATMENT-OUTCOME ASSOCIATION

Threats to inferences about whether an association between treatment and outcome classes is causal or spurious arise mostly out of the designs of primary studies. It is at the level of the individual study that the temporal sequence of cause and effect, randomization of units to conditions, and other design features aimed at strengthening causal inferences are implemented.

The following threats (see table 22.2) are in addition to those presented earlier that refer to the likelihood of an association between treatment and outcome classes. The logic of causal inference is fairly straightforward at the primary study level, but is complicated at the research synthesis level because findings from partially flawed primary studies often need to be combined. Inferences about causation are not necessarily jeopardized by deficiencies in primary studies because—at least in theory—individual sources of bias in the primary studies may cancel each other out exactly when aggregated in a research synthesis. So, at the research synthesis level, a threat only arises if the deficiencies within each primary study combine across studies to create a predominant direction of bias.

Table 22.2 Threats to Inferences About the Causal Nature of the Treatment-Outcome Association

1. Absence of studies with successful random assignment
2. Primary study attrition

SOURCE: Authors' compilation.

22.3.1 Absence of Studies with Successful Random Assignment

In primary studies, unbiased causal inference requires establishing the direction of causality and ruling out third-variable alternative explanations. Inferring the direction of causality is easy in experimental and quasi-experimental studies where knowledge is usually available about the temporal sequence from manipulating the treatment to measuring its effects. In theory, third-variable alternative explanations are ruled out when participants in primary studies are randomly assigned to treatment conditions or if a regression-discontinuity study is done (Shadish, Cook, and Campbell 2001; West, Biesanz, and Pitts 2000; Imbens and Lemieux 2008).

In research syntheses of well-implemented randomized trials, the strong causal inferences at the primary study level are transferred to the causal inferences at the level of the research synthesis. This is why, for example, the Cochrane Collaboration originally restricted systematic reviews of medical research to randomized controlled trials (Chalmers 1993). Systematic reviews conducted through the Campbell Collaboration cover interventions from broader range of disciplines (including criminal justice, education, social welfare, and international development) and may include randomized experiments, high-quality quasi-experiments, and observational studies (Shemilt et al. 2008). If randomization was poorly implemented or absent, then causal inferences are ambiguous and the pervasive possibility of third-variable explanations arises (Campbell and Boruch 1975). This was illustrated in

a meta-analysis of quasi-experimental studies of neurocognitive functioning in post-traumatic stress disorder (PTSD), where PTSD diagnoses may be confounded with seeking treatment and comorbid conditions (Scott et al. 2015). If lack of random assignment in primary studies yields a predominant bias across these studies, causal inference at the level of the research synthesis is jeopardized.

One approach to this threat compares the average effect estimates from studies with random assignment to those studies on the same question with more systematic assignment. If effect estimates differ, causal primacy must then be given to findings from randomized designs, assuming that the randomized and non-randomized studies are equivalent on other characteristics. Though a set of randomized experiments and quasi-experiments sometimes result in similar estimates, there are some notable exceptions (Shadish, Luellen, and Clark 2006; Jacob and Ludwig 2005; Boruch 2005; Lipsey 1992; Smith, Glass, and Miller 1980; Wittmann and Matt 1986; Chalmers et al. 1983).

A second strategy involves the careful explication of possible biases caused by flawed randomization or associated with a particular quasi-experimental design in each primary study as Cobb Scott and colleagues attempted to do (Scott et al. 2015). Based on a sufficiently complete understanding of possible biases, adjustments for pretreatment differences between groups can sometimes be made to project causal effects, controlling for the identified biases. The problem here of course is justifying that all important biases have been identified and their operation has been correctly modeled. This latter approach is particularly useful when the first approach fails if too few primary studies with randomized designs exist (for further discussion of these issues and a checklist for meta-analysts planning to include non-randomized studies, see Norris et al. 2013; Wells et al. 2013; Reeves et al. 2013; Higgins et al. 2013).

22.3.2 Primary Study Attrition

Attrition of participants from treatment or measurement is common in even the most carefully designed randomized experiments and quasi-experiments. If attrition in these primary studies is differential across treatment groups, effect estimates may then be biased, inflating or deflating causal estimates. If the biases operating in each direction completely cancel out, then causal inference at the research synthesis level is not affected. However, if predominant bias persists across the primary studies, then causal inferences from the synthesis are called into question. For instance, Mark Lipsey's (1992) meta-analysis of juvenile delinquency interventions found that the more amenable juveniles might have dropped out of treatment groups and the more delinquent juveniles out of control groups. The net effect at the level of the research synthesis is a potential bias. Ying Yuan and Roderick Little (Yuan and Little 2009) examined the effect of missing data in primary studies on estimates of treatment effect and concluded that standard random-effects models lead to biased estimates if study-level attrition rates and effect sizes are correlated. The magnitude of the bias is positively associated with the strength of the association and the relative size of the within-study and between-study variance and may be corrected for using methods they developed.

To address this threat, it is crucial to code, for each primary study, information about the level and differential nature of attrition from the treatment groups. The latter information cannot be inferred from the sheer level of attrition, but requires the primary study to have reported the possible differential attrition of participants. If this information is available, the analyst can then examine the effect of attrition on effect estimates by disaggregating studies according to their level of total and differential attrition. Jeffrey Valentine and McHugh use this strategy in combination with sensitivity analyses to explore whether the reported levels of attrition presented plausible threats in randomized experiments in education (2007).

22.4 THREATS TO GENERALIZED INFERENCES

Although some of the threats already discussed are germane to generalization (for example, underjustified use of fixed and random-effects models), we now turn explicitly to generalization issues, first discussing threats that may lead to erroneous conclusions about target constructs and universes of people, treatments, outcomes, and settings (threats 1–7 in table 22.3). Some of these threats have their origins in deficiencies and constraints of the primary studies; others are independent of primary studies and linked to deficiencies of the coding and rating system a meta-analyst uses. Next, we examine the threats pertinent to analyses of potential moderator variables from which generalized inferences are drawn about empirical robustness and causal contingency (threats 8–13 in table 22.3). This is followed by an additional threat (14) relevant to generalizations that extrapolate findings beyond the universes included in a meta-analysis to unstudied people, treatments, outcomes, and settings.

Table 22.3 Threats to Generalized Inferences in Research Syntheses

Inferences to Target Constructs and Universes
1. Biased sampling of inference domains
2. Underrepresentation of prototypical attributes
3. Restricted heterogeneity of substantively irrelevant third variables
4. Mono-operation bias
5. Mono-method bias
6. Rater drift
7. Reactivity effects
Inferences About Robustness and Moderating Conditions
8. Restricted heterogeneity in inference domains
9. Moderator variable confounding
10. Failure to test for homogeneity of effect sizes
11. Lack of statistical power for homogeneity tests
12. Lack of statistical power for studying disaggregated groups
13. Misspecification of causal mediating relationships
Extrapolations to Novel Constructs and Universes
14. Misspecification of models for extrapolation

SOURCE: Authors' compilation.

22.4.1 Biased Sampling of Inference Domains

Statistical sampling theory is sometimes invoked as the justification for generalizing from the obtained instances (or samples) of people, treatments, outcomes, and settings to the universes or domains they are thought to represent, and about which researchers want to draw inferences. However, it is rare in individual primary studies for instances to be selected at random from their target universes, just as it is rare in research syntheses to randomly sample primary studies, treatments, outcomes or settings from the universe of research on a given topic. Some exceptions are noteworthy, as when Robert Orwin and David Cordray (1985) and Georg Matt (1989) randomly sampled studies from those included in the Smith, Glass, and Miller (1980) meta-analysis of the psychotherapy outcome—assuming it had a census of all studies or a random sample from the census. Perhaps the most feasible random selection in meta-analysis is when, as some methodologists have suggested (Lipsey and Wilson 2001), only one effect size is sampled per study so as to provide an unbiased estimate of the mean effect size per study.

More common study selection strategies are for meta-analysts to seek collecting the entire population of published and unpublished studies on a topic, or restricting the selection of studies to all those with specific person, treatment, outcome, setting or time characteristics of substantive importance (Mick et al. 2003; Moyer et al. 2002). In either of these circumstances, inferences from the samples to their target universes may be biased if the meta-analysts are unable to retrieve all studies from the target populations and the missing studies have different average effect sizes than those that were included. As a result, generalizations can easily be overstated, even if they are supported by data from hundreds of studies and thousands of research participants.

22.4.2 Underrepresentation of Prototypical Attributes

Research syntheses should start with the careful explication of the target constructs about which inferences are to be drawn, at a minimum identifying their prototypical attributes and any less central features at the boundaries with other constructs. Thus, it can be that the collection of primary studies in the research synthesis does not contain representations of all the prototypical elements. This was the case when William Shadish and his colleagues attempted to investigate the effects of psychotherapy in clinical practice, for example (2000). They observed that many studies included a subset of prototypical features of clinical practice, but no single study included all of its features as they defined them. As a result, generalization to real-world psychotherapy practice is problematic, despite thousands of studies on the effectiveness of psychotherapy practice. In such a situation, the operations implemented in the primary studies force us to revise the construct about which inferences are possible, reminding us that we cannot generalize to the practice of psychotherapy as it is commonly conducted in the United States. An important task of meta-analysis is to inform the research community about the underrepresentation of prototypical elements of core constructs in the literature on hand, turning attention to the need to incorporate them into the sampling designs of future studies.

22.4.3 Restricted Heterogeneity of Substantively Irrelevant Third Variables

Even if sampling from the universes about which generalized inferences are sought were random and important prototypical elements were represented in the reviewed

studies, a threat arises if a research synthesist cannot demonstrate that the causal association is robust and holds across substantively irrelevant characteristics. For instance, if the reviewed studies on the effectiveness of homework in middle school were conducted by just one research team, relied on voluntary participation by students, or depended on teachers being highly motivated, the threat would then arise that all conclusions about the general effectiveness of homework are confounded with substantively irrelevant aspects of the research context. To give an even more concrete example, if private schools were explicated to be those where school expenses come from student fees, donations, and the proceeds on endowments (rather than from taxes), it is irrelevant whether the schools are parochial or nonparochial, military schools or elite academic schools. To generalize to the universe of private schools requires being able to show that relationships are not limited to one or a few of these contexts—say parochial or military schools.

Limited heterogeneity of universes will also impede the transfer of findings to new universes (that is, extrapolation), because it hinders the ability to demonstrate the robustness of a causal relationship across substantive irrelevancies of design, implementation, or measurement method. The wider the range and the more substantively irrelevant aspects across which a finding is robust, and the better moderating influences are understood, the stronger the belief that the finding will also hold under the influence of not yet examined contextual irrelevancies.

22.4.4 Mono-Operation Bias

The coding and rating systems of research syntheses often rely exclusively on single items to measure such complex target constructs as setting, diagnosis, or treatment type. It is well known from the psychometric literature that single-item measures have poor measurement properties. They are notoriously unreliable, tend to underrepresent a construct, and are often confounded with irrelevant constructs. To address and improve on common measurement limitations of meta-analytic coding manuals, rigorous procedures for establishing inter-rater reliability and standard scale development procedures have to become common practice to allow valid inferences about target constructs of a meta-analysis.

22.4.5 Mono-Method Bias

This bias occurs if the measurement method in a meta-analysis relies on a single coder who reads and codes a research publication, following the operations delineated in a coding manual. To avoid this bias, coding procedures for meta-analyses have to incorporate multi-method coding approaches. This could include having multiple coders (perhaps with different substantive backgrounds related to the research topic) code all items of a coding manual, contacting the original authors to provide clarification and additional information, obtaining additional write-ups of a study, such as complete dissertations or reports to a funding agency, and relying on external, supplementary sources, such as to describe psychometric properties of an outcome measure or allegiance and experience of a researcher (Robinson, Berman, and Neimeyer 1990). In the absence of such improvements, conclusions about important target constructs relying on single coder ratings remain suspect.

22.4.6 Rater Drift

Reading, understanding, and coding publications of multifaceted primary studies involve many cognitively challenging tasks. Over time, coders learn through practice, develop heuristics to simplify complex tasks, fatigue and become distracted, and may change their cognitive schemas as a result of exposure to study reports. As a consequence, the same coder may unknowingly change over time such that earlier codings of the same evidence differ from later ones. To address this validity threat, rater drift needs to be monitored as part of a continuing coder training program as Bryce McLeod, John Weisz, and Jeffrey Wood do in their meta-analysis of the association between parenting and childhood depression (2007). Moreover, changes in coding manuals should be made publicly to reflect changes in the understanding of a code, which may necessitate recoding studies that were examined under the prior coding rules.

22.4.7 Reactivity Effects

A measure is said to be reactive if the measurement process itself influences the outcome (Webb et al. 1981). In the case of a research synthesis, the coding process itself may inadvertently influence the coding outcome. For instance, knowing that the author of a study is a well-known expert in the field rather than an unknown novice may predispose a coder to rate research design characteristics more favorably for the expert. Similarly, knowing that the treatment yielded no benefits over the control condition may bias a coder to rate the implementation of the treatment condition more critically. To minimize reactivity biases, it is desirable to mask raters to influences that could bias their codings,

including authorship and study results. Further, as much as possible, attempts should be made to avoid reactivity biases, including codings that require as little inference as possible on the part of the raters. For example, rather than asking raters to arrive at an overall code describing study quality, it would be better to ask them to code specific aspects of studies that pertain to quality, such as the method of allocating participants to groups and attrition, which are less likely to be reactive (see chapters 7 and 9, this volume).

22.4.8 Restricted Heterogeneity in Inference Domains

Inferences about the conditions moderating the magnitude and direction of an association are facilitated if the relationship can be studied for a large number of diverse domains or universes of people, treatments, outcomes, settings, and times. This is the single most important potential strength of research syntheses over individual studies. Although meta-analyses of standardized treatments and specifically designated outcomes, populations, and settings may increase the precision of some effect-size estimates, such restrictions hamper our ability to better understand the conditions under which such relationships can and cannot be observed. Rather than limiting a meta-analysis to a review of a single facet of the universe of interest or lumping together a heterogeneous set of facets, we encourage meta-analysts—sample sizes permitting—to explicitly represent and take advantage of such heterogeneities.

Such advice has implications for those observers who have suggested that causal inferences in general—and causal moderator inferences in particular—could be enhanced if meta-analysts relied only on studies with superior methodology, particularly randomized experiments with standardized treatments, manifestly valid outcomes and clearly designated settings and populations (Chalmers et al. 1989; Chalmers and Lau 1992; Sacks et al. 1987; Slavin 1986). Such a strategy appears to be useful in areas where research is fairly standardized. However, other limitations arise because this standardization limits the heterogeneity in research designs, treatment implementations, outcome measures, recruitment strategies, subject characteristics, and the like. Hence it is not possible to examine empirically how robust a particular effect is that has been obtained in a restricted, standardized context, leaving uncharted the realm of application of meta-analytic conclusions.

Meta-analysis has the potential to increase confidence in generalizations to new universes (that is, extrapolation) if findings are robust across a wide range and large number of different universes. The more robust the findings and the more heterogeneous the populations, settings, treatments, outcomes, and times in which they were observed, the greater the belief that similar findings will be observed beyond the populations studied. If the evidence for stubborn empirical robustness can be augmented by evidence for causal moderating conditions, the novel universes in which a causal relationship is expected to hold can be even better identified.

The logical weakness of this argument lies in its inductive basis. That psycho-educational interventions with adult surgical patients have consistently shown earlier release from the hospital across a broad range of major and minor surgeries and across diverse respondents and of treatment providers throughout the 1960s, 1970s, and 1980s cannot logically guarantee the same effect will hold for as-yet-unstudied surgeries, and in the future (Devine 1992). However, the robustness of the relationship does strengthen the belief that psycho-educational interventions will have beneficial effects with new groups of patients and with novel surgeries in the near future (that is, extrapolation). Homogeneous populations, treatments, outcomes, settings, and times limit the examination of causal contingencies and robustness, and consequently, impede inferences about the transfer of findings to novel contexts.

22.4.9 Moderator Variable Confounding

At the synthesis level, moderator variables describe characteristics of the classes of treatment, outcomes, settings, populations, or times across which the magnitude or direction of a causal effect differs—a generalizability question. Claims about moderator variables are involved when a research synthesist concludes that treatment type A is less effective in population C than D, stronger with outcome of type E than F, weaker in setting G than H or positive in past times but not recently. Moderator variables are even involved when the claim is made in a synthesis that treatment type A is superior to treatment type B, for this assumes that the studies of A are equivalent to those of B on everything correlated with the outcome other than A versus B. Any uncertainty about this pertains to the role that average study differences might have played in achieving the obtained difference between studies of A and B; thus it is an issue of moderator variables.

Threats to valid inference about the causal moderating role in research syntheses are pervasive (Lipsey 2003; Shadish and Sweeney 1991; Scott et al. 2007; Valentine

and Thompson 2013). This is because participants in primary studies are rarely assigned to moderator conditions at random, for example, to outcome type E versus F or to settings G versus H (see chapter 2, this volume). Claims about the causal role of moderators are often questionable in syntheses because any one moderator is often confounded with other study characteristics. For instance, studies of smoking cessation in medical primary care facilities are likely to attract older participants than similar studies on college campuses. Likewise, behavioral treatments tend to involve behavioral outcome measures of a narrow target behavior, whereas psycho-dynamic interventions are likely to rely on broader measures of adjustment. If the moderator variable (for example, primary care setting versus college campus) is confounded with characteristics of the design, setting, or population (for example, age), differences in the size or direction of a treatment effect that might be attributable to the moderator are instead confounded with other attributes of the set of studies in one setting and the set of studies in another.

To deal with this possibility, meta-analysts should examine within-study comparisons of the moderator effect because these are obviously not prone to between-study confounds. For instance, if the moderating role of treatment types A, B, and C is at stake, a meta-analysis can be conducted of all the studies with internal comparisons of treatment types A, B, and C (for example, Shapiro and Shapiro 1982). Or, if the moderating role of subject gender is of interest, inferences about gender differences might depend on within-study contrasts of males and females rather than on comparisons of studies with only males or females or that assess the percentage of males. An extension of direct head-to-head within-study comparisons has emerged out of networks of randomized trials (Salanti et al. 2008; Song et al. 2003; Salanti 2012). Known as mixed treatment comparison meta-analyses and network meta-analyses, this approach aims to improve meta-analyses comparing multiple treatments by providing a formal model for comparing treatments (for instance, ranking A, B, C, D) based on different multiple-treatments studies, none of which directly compared all treatments. Although such network meta-analyses are subject to potential moderator variable confounding (for instance, studies comparing A with B differ from those comparing B with C), they make transparent the network of available evidence and make explicit critical assumption regarding indirect comparisons based on multi-treatment studies, that is, consistency and transitivity. This allows the exploration of potential biases through sensitivity analyses and statistical model (Higgins et al. 2012; White et al. 2012). The latter is a common general approach to the exploration of a moderator variables (Shadish and Sweeney 1991; Donegan et al. 2015). The validity of the causal inferences based on such models depends on the ability of the meta-analyst to identify and reliably measure all confounding variables. Multivariate statistical adjustments can be informative here, but not definitive given the difficult task of conceptualizing all such confounds and then measuring them well across the studies under review.

22.4.10 Failure to Test for Homogeneity of Effect Sizes

Under a fixed-effect statistical model, a statistical test for homogeneity assesses whether the variability in effect estimates exceeds that expected from sampling error alone (Hedges 1982; Rosenthal and Rubin 1982). If the null hypothesis of homogeneous effect sizes is retained, a single population effect size provides a parsimonious model of the data, and the weighted mean effect size provides an adequate estimate. If the null hypothesis is rejected, the implication is that subclasses of studies may exist that differ in population effect size, triggering the search to identify the nature of such subclasses. Hence, heterogeneity tests play an important role in examining the robustness of a relationship and in initiating the search for potential moderating.

The failure to test for heterogeneity may result in lumping manifestly different subclasses of people, treatments, outcomes, settings, or times into one class. This problem has been referred to as the apples and oranges problem of meta-analysis. However, Gene Glass, Robert Rosenthal, and Georg Matt have argue that apples and oranges should indeed be mixed if the interest is in generalizing to such higher-order characteristics as fruit or whatever else inheres in an array of treatments, outcomes, settings, people, and times (Glass 1978; Smith, Glass, and Miller 1980; Rosenthal 1990; Matt 2003, 2005). We should indeed be willing to combine studies of manifestly different subclasses of people, treatments, outcomes, settings, or times if they yield equivalent results in reviews. In this context, the homogeneity test indicates when studies yield such different results that a single, common average effect size needs to be disaggregated through blocking on study characteristics that might explain the observed variance in effect sizes.

Coping with this threat is relatively simple, and homogeneity tests have become standard both as a way of

identifying the most parsimonious statistical model to describe the data and as a means of testing model specification (see chapter 12). However, the likelihood of design flaws in primary studies, of publication biases and the like makes the interpretation of homogeneity tests more complex (Hedges and Becker 1986; Schulze 2004). Evidence also indicates that the choice of the effect-size metric, such as *r* versus Fisher *z*-transform or odds ratio versus log odds, affects the outcome of heterogeneity tests (Engels et al. 2000). If all the studies being meta-analyzed have the same flaw, or if studies with zero and negative effects are less likely to be published, then a consistent bias results across studies and can make the effect sizes appear more homogeneous than they really are. Further, even if the collection of studies is not biased, a failure to reject the null hypothesis that observed effect sizes are a sample realization of the same population effect does not prove it. Conversely, if all the studies have different design flaws, effect sizes could be heterogeneous even though they actually have the same population effect. Obviously, the causes of heterogeneity that are of greatest interest are substantive rather than methodological. Consequently, it is useful to differentiate between homogeneity tests conducted before and after the assumption, that all study-level differences in methodological irrelevancies have been accounted for, has been defended.

22.4.11 Lack of Statistical Power for Homogeneity Tests

A homogeneity test examines whether the observed variability in effect sizes is more than would be expected from sampling error alone. It is thus the gatekeeper test for deciding to continue the search for variables that moderate the average effect size obtained in a review—a generalization task. But when these homogeneity tests have little statistical power, as is usually the case (Gavaghan, Moore, and McQuay 2000), their type II error rate will be large and lead to the erroneous conclusions that the search for moderator variables should be abandoned (Jackson 2006; Hedges and Pigott 2001; Harwell 1997). Moreover, Elena Kulinskaya, Michael Dollinger, and Kirsten Bjørkestøl developed a modification of the standard Q test that provides more accurate results especially for small and moderate study sizes (2011b). They show that the improved accuracy leads to a decrease in power for risk difference effect sizes and an increase in power for standardized mean difference effect sizes (2011a). When statistical power is suspect, statistically nonsignificant homogeneity tests provide inclusive evidence and should not be exclusively relied on to justify conclusions about the robustness of an association.

22.4.12 Lack of Statistical Power for Studying Disaggregated Groups

If there is reason to believe treatment effects may differ across types of treatments, outcomes, people, settings, or over time (that is, moderator effects), highly aggregated classes (for example, therapy or well-being) have to be disaggregated to examine the conditions under which an effect changes direction or magnitude. Such subgroup analyses rely on fewer studies than main effect analyses and involve additional statistical tests that may necessitate procedures for type I error control, lowering the statistical power for the subanalyses in question. Consequently, the chances are reduced to find statistically significant differences even if such differences exist in the population. The flaws of this erroneous inference are compounded if a meta-analyst then concludes that an effect generalizes across subgroups of the universe because no statistically significant differences were found. Large samples of studies mitigate this problem to some extent, although the power in research syntheses is more complex than in primary studies. This is because power depends not only on the effect size, type I error rate, and sample size of primary study participants, but also on the number of studies and the underlying statistical model (Cohn and Becker 2003; Hedges and Pigott 2004; Valentine, Pigott, and Rothstein 2010).

22.4.13 Misspecification of Causal Mediating Relationships

Mediating relationships are examined to shed light on the processes by which a causal agent, such as secondhand smoke, transmits its effects on an outcome, such as asthma exacerbation. They provide highly treasured explanations of otherwise merely descriptive causal associations. Instead of simply noting that a treatment affects an outcome, mediating relationships inform us why and how a treatment affects an outcome. Explanatory models rely on causal mediating relationships to justify extrapolations and to specify causal contingencies; both are important generalization tasks.

Few meta-analyses of mediating processes exist, and those that do utilize quite different strategies. Monica Harris and Robert Rosenthal's meta-analysis of the mediation of

interpersonal expectancy effects used the available experimental studies that had information on at least one of the four mediational links abstracted from relevant substantive theory, which were never tested together in any individual study (1985). Steven Premack and John Hunter's meta-analysis of mediation processes underlying individual decisions to form a trade union relied on combining correlation matrices from different studies on subsets of variables believed to mediate a cause-effect relationship (1988). The entire mediation model had not been probed in any individual study. Betsy Becker's meta-analysis relied on the collection of individual correlations or correlation matrices to generate a combined correlation matrix (1992). That is, individual studies may contribute evidence for one or more of the causal links postulated in the mediational model. This approach has been further developed in an effort to conduct meta-analyses of structural equation models based on a pooled correlation matrix (Cheung and Chan 2005). Recent advances on this topic and persistent challenges are discussed in a special issue of *Research Synthesis Methods* (Cheung and Cheung 2016; Cheung and Hafdahl 2016; Gnambs and Staufenbiel 2016; Hedges 2016; Oort and Jak 2016; Sheng et al. 2016; Wilson, Polanin, and Lipsey 2016; Yuan 2016). All these approaches examine mediational processes where the data about links between variables come from within-study comparisons. However, a fourth mediational approach infers causal links from between-study comparisons. William Shadish's research on the mediation of the effects of therapeutic orientation on treatment effectiveness is a salient example (Shadish 1992; Shadish and Sweeney 1991).

Although the four approaches differ considerably, they all struggle with the same practical problem—how to test mediational models involving multiple causal connections when only few (or no) studies are available in which all the connections have been examined within the same study. A major source of threats to the meta-analytic study of mediation processes arises from between-study heterogeneity (Hedges 2016) such that the different correlation matrices or different mediational links are supported by different quantities and qualities of evidence, perhaps because of excessive missing data on some mediational links, the operation of a predominant direction of bias in some studies, or publication or attrition biases involving yet other parts of the mediational model. How to quantify and model this heterogeneity remains an important issue to be addressed.

On the basis of the earlier meta-analyses of causal mediational processes as well as more recent meta-analyses of structural equation models, it is clear that missing data is a pervasive reason for misspecifying causal models in a meta-analysis (Cook et al. 1992). Missing data within studies may or may not be completely at random, and reported correlations may or may not provide consistent estimates. Missing correlations within studies may lead meta-analyst to omit important mediational links from the model that is actually tested and may also prevent consideration of alternative plausible models. Although in many substantive areas large numbers of studies are probing a particular descriptive causal relationship, in far fewer are large numbers of studies providing a broad range of information on mediational processes, and even fewer where two or more causal mediating models are directly compared. There are many reasons for these deficiencies. Process models are derived from substantive theory, and these theories change over time as they are improved or as new theories emerge. Obviously, the novel constructs in today's theories are not likely to have been measured well in past work. Moreover, today's researchers are reluctant to measure constructs from past theories that are now out of fashion or obsolete. Thus, the dynamism of theory development does not fit well the meta-analytic requirement for a stable set of mediating constructs or the simultaneous comparison of several competing substantive theories. Another relevant factor is that, even if a large number of mediational variables have been measured, they are often measured less well and documented less thoroughly than measures of cause and effect constructs are. Measures of mediating variables are given less attention, and are sometimes not analyzed unless a molar causal relationship has been established. Given the pressure to publish findings in timely fashion and to write succinct rather than comprehensive reports, the examination of mediating processes is often postponed, limited to selected findings, or based on inferior measures.

22.4.14 Misspecification of Models for Extrapolation

The most challenging type of generalization involves inferences to new populations, treatments, outcomes, and settings. Such extrapolations to novel conditions have to rely on the available empirical evidence and specifically on the strength of the generalized inferences this evidence allows to the target universes represented in these studies. Therefore, all threats discussed to this point for generalized inferences to target universes (1–7 in table 22.3) and robustness and moderators (8–13 in table 22.3) necessarily

apply and affect extrapolations to novel constructs and universes.

Based on the available evidence, extrapolations rely on explicit and implicit models or informal heuristics to project treatment effects under yet unstudied conditions, such as the effects of a new cancer drug in pediatric patients based on evidence in adult populations or the effects of psychotherapy in clinical practice based on studies conducted in research settings. At issue, then, is the validity of the model used to project the effects of an off-label use of a drug (based, for example, on weight, metabolism, development stage of organs) or the effects of psychotherapy in practice (based on caseload of therapists, patient mix, manualized or eclectic therapy). Under the best conditions, these models are informed by a comprehensive understanding of the causal mediating processes and empirical evidence about the robustness of effects and their moderating conditions across a broad range of substantively relevant and irrelevant variables. To the extent that this is not the case, the threat arises that the explicit or implicit extrapolation models may be misspecified, creating the risk for incorrect projections of effects under novel conditions.

Our discussion makes it clear that extrapolations to novel conditions carry higher uncertainty than generalized inferences to universes from which particulars have been studied. In this situation, the delicate goal of the meta-analyst is to communicate the existing evidence and the models from which an extrapolation is made, the assumptions built into such models, the uncertainty of the inference, and the conditions necessary to reduce the uncertainty. For the user of a meta-analysis, it may then be possible to weigh the risks and harm of an incorrect extrapolation against the likelihood and benefits of a correct extrapolation.

22.5 CONCLUSION

We argue that the major promise of research syntheses lies in strengthening empirical generalizations of associations between classes of treatments and outcomes. The history of meta-analytic practice, however, has demonstrated that this promise is threatened, because meta-analysts cannot rely on the two major scientific warrants for generalizability claims, random sampling from designated populations and strong causal explanations. This chapter offers an alternative approach, a threats-to-validity framework, to explore and make a case for generalized inferences in meta-analysis when the established models cannot be applied.

Following Donald Campbell and Lee Cronbach, we distinguish between three types of generalized inferences that are central to meta-analyses (Campbell and Stanley 1963; Cook and Campbell 1979; Cronbach 1980). The first deals with inferences about an association in target universes of people, treatments, outcomes, and settings based on particulars sampled from these universes. The second concerns the robustness and conditional nature of an association across heterogeneous, substantively relevant and irrelevant conditions. The third concerns extrapolating or projecting an association to novel, as-yet unstudied universes. The proposed threats-to-validity framework seeks to assist meta-analysts in exploring alternative explanations to the generalized inferences of interest. Because meta-analysts cannot realistically rely on the elegance and strength of sampling theory to warrant generalizability claims, the proposed framework offers a falsificationist approach to identify and rule out plausible alternative explanations to justify generalized inferences in meta-analyses.

Although the threats-to-validity framework makes no assumptions about random sampling or comprehensive causal explanations, it does require that we critically investigate all plausible concerns that can lead to spurious inferences and rule out each concern on grounds that it is implausible or based on evidence that it did not operate in a specific situation. Similar to the causal inferences based on Campbell's threats to validity for quasi-experimental designs, the generalized inferences based on the proposed threats to validity for meta-analyses are tentative and only as good as the critical evaluation of the plausible threats. The threats we present are, in large measure, a summary of what other scholars have identified since Gene Glass's pioneering work (Glass 1976; Smith, Glass, and Miller 1980). We expect the list to continue to change at the margin as new threats are identified, current ones are seen as less relevant than we now think, and as meta-analysis is put to new purposes.

To the fledgling meta-analyst, our list of validity threats may appear daunting and the proposed remedies overwhelming. More experienced practitioners will be less intimidated by the numerous threats because they operate from an implicit theory about the differential seriousness and prevalence of these threats and can recognize when the needed substantive and technical expertise and resources are on hand. Even so, all research syntheses have to make important trade-offs between partially conflicting goals. Thus, should methodologically less rigorous studies be excluded to strengthen causal inferences if this also limits the ability to explore potential moderating conditions and

the empirical robustness of effects? Should resources be allocated to code fewer study features with greater validity or to code more study features that might capture even more potentially important reasons why effect sizes differ? When should one stop searching for more relevant studies if they are increasingly fugitive, or stop trying to obtain missing data from primary study authors, or stop checking coder reliability and drift? To the experienced practitioner, the definitive research synthesis is perhaps even more of a fiction than the definitive primary study. The goal is research syntheses that answer important questions about the generalizability of an association while making explicit the limitations of the findings claimed, raising new and important questions about the boundaries of our understanding, and setting the agenda for the next generation of research.

In the earlier editions of this book, we called for the development of a viable new theory of generalization that can guide the design of research syntheses. We believe that the principles proposed by Cook (1990, 1993) and further elaborated by Shadish, Cook, and Campbell (2001) can serve as a starting point for such a theory. From this perspective, secure general knowledge can only emerge from studies that examine under different conditions heterogeneous variations of a common theme or template of a causal association. In combination, the proposed principles offer practical guidelines for exploring generalizability claims and ruling out threats to validity, for which meta-analyses can then provide empirical warrants. They are not, though, principles firmly ensconced in statistical theory as is the case with probability sampling. But the latter has a limited reach across the persons, settings, times and instances of the cause and effect that are necessarily involved in generalizing from sampling details to causal claims. For all its undisputed theoretical elegance, probability sampling is most relevant to generalizing to populations of units and, sometimes, to settings. Its practical relevance to the other entities involved in causal generalization is much less clear. Although the past forty years of practice have seen many improvements, significant further progress is still needed to achieve a more complete understanding of how research syntheses can achieve even better causal generalization. The necessary condition for this is the existence of one or more comprehensive and internally cogent theories of causal generalization. But this we do not yet have in any form that leads to novel practical actions when conducting meta-analyses. Until we have such cogent theories of generalization, the proposed principles and validity threats can guide the meta-analyst to identify the realm of application of a knowledge claim to sustain the evolution of scientific knowledge.

22.6 REFERENCES

Ahmed, Ikhlaaq, Alexander J. Sutton, and Richard D. Riley. 2012. "Assessment of Publication Bias, Selection Bias, and Unavailable Data in Meta-Analyses Using Individual Participant Data: A Database Survey." *British Medical Journal* 344: d7762. DOI: 10.1136/bmj.d7762.

Anderson, Samantha F., and Scott E. Maxwell. 2016. "There's More Than One Way to Conduct a Replication Study: Beyond Statistical Significance." *Psychological Methods* 21(1): 1–12. DOI: 10.1037/met0000051.

Arocas Casañ, Vicente, J. Mateo Carmona, O. Molina Garcia, Ma Angeles Fernandez de Palencia Espinosa, Blázquez Alvarez, Ma Amelia de la Rubia Nieto, and Jesús del Rio Garcia. 2016. "Off-Label Prescription of Drugs at Hospital." *Farmacia Hospitalaria* 40(2): 63–71. DOI: 10.7399/fh.2016.40.2.9268.

Becker, Betsy J. 1992. "Models of Science Achievement: Forces Affecting Male and Female Performance in School Science." In *Meta-Analysis for Explanation: A Casebook*, edited by Thomas D. Cook et al. New York: Russell Sage Foundation.

———. 2001. "Examining Theoretical Models Through Research Synthesis: The Benefits of Model-Driven Meta-Analysis." *Evaluation and the Health Professions* 24(2): 190–217.

Boruch, Robert F. 2005. "Comments on 'Can the Federal Government Improve Education Research?' by Brian Jacob and Jens Ludwig." *Brookings Papers on Education Policy* 1: 67–80.

Brannick, Michael T., Liu-Qin Yang, and Guy Cafri. 2011. "Comparison of Weights for Meta-Analysis of *r* and *d* Under Realistic Conditions." *Organizational Research Methods* 14(4): 587–607. DOI: 10.1177/1094428110368725.

Brauner, Julie V., Lily M. Johansen, Troels Roesbjerg, and Anne K. Pagsberg. 2016. "Off-Label Prescription of Psychopharmacological Drugs in Child and Adolescent Psychiatry." *Journal of Clinical Psychopharmacology* 36(5): 500–507. DOI: 10.1097/JCP.0000000000000559.

Campbell, Donald T. 1957. "Factors Relevant to the Validity of Experiments in Social Settings." *Psychological Bulletin* 54(4): 297–312.

———. 1986. "Relabeling Internal and External Validity for Applied Social Scientist." In *Advances in Quasi-Experimental Design and Analysis*, edited by William M. K. Trochim. San Francisco: Jossey-Bass.

Campbell, Donald T., and Robert F. Boruch. 1975. "Making the Case for Randomized Assignment to Treatments by Considering the Alternatives: Six Ways in Which Quasi-Experimental Evaluations Tend to Underestimate Effects." In *Evaluation and Experiment : Some Critical Issues in Assessing Social Programs*, edited by Carl A. Bennett and Arthur A. Lumsdaine. New York: Academic Press.

Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally.

Chalmers, Iain. 1993. "The Cochrane Collaboration: Preparing, Maintaining, and Disseminating Systematic Reviews of the Effects of Health Care." *Annals of the New York Academy of Science* 703(1): 156–65.

Chalmers, Thomas C., Paul Celano, Henry S. Sacks, and Harry Smith Jr. 1983. "Bias in Treatment Assignment in Controlled Clinical Trials." *New England Journal of Medicine* 309(22): 1358–61.

Chalmers, Thomas C., Peg Hewett, Dinah Reitman, and Henry S. Sacks. 1989. "Selection and Evaluation of Empirical Research in Technology Assessment." *International Journal of Technology Assessment in Health Care* 5(4): 521–36.

Chalmers, Thomas C., and Joseph Lau. 1992. "Meta-Analysis of Randomized Control Trials Applied to Cancer Therapy." In *Important Advances in Oncology, 1992*, edited by Vincent T. Devita, Samuel Hellman and Steven A. Rosenberg. Philadelphia, Pa.: Lippincott.

Cheung, Mike W. L. 2008. "A Model for Integrating Fixed-, Random-, and Mixed-Effects Meta-Analyses into Structural Equation Modeling." *Psychological Methods* 13(3): 182–202. DOI: 10.1037/a0013163.

———. 2014. "Modeling Dependent Effect Sizes with Three-Level Meta-Analyses: A Structural Equation Modeling Approach." *Psychological Methods* 19(2): 211–29. DOI: 10.1037/a0032968.

Cheung, Mike W. L., and Wai Chan. 2005. "Meta-Analytic Structural Equation Modeling: A Two-Stage Approach." *Psychological Methods* 10(1): 40–64. DOI: 10.1037/1082–989X.10.1.40.

Cheung, Mike W. L., and Shu Fai Cheung. 2016. "Random-Effects Models for Meta-Analytic Structural Equation Modeling: Review, Issues, and Illustrations." *Research Synthesis Methods* 7(2): 140–55. DOI: 10.1002/jrsm.1166.

Cheung, Mike W. L., and Adam R. Hafdahl. 2016. "Special Issue on Meta-Analytic Structural Equation Modeling: Introduction from the Guest Editors." *Research Synthesis Methods* 7(2): 112–20. DOI: 10.1002/jrsm.1212.

Cohn, Lawrence D., and Betsy J. Becker. 2003. "How Meta-Analysis Increases Statistical Power." *Psychological Methods* 8(3): 243–53. DOI: 10.1037/1082–989X.8.3.243.

Collingwood, Robin G. 1940. *An Essay on Metaphysics*. Oxford: Clarendon Press.

Cook, Thomas D. 1990. "The Generalization of Causal Connections." In *Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data*, edited by L. Sechrest, E. Perrin and J. Bunker. Washington: U.S. Department of Health and Human Services.

———. 1992. *Meta-Analysis for Explanation: A Casebook*. New York: Russell Sage Foundation.

———. 1993. "Understanding Causes and Generalizing About Them." In *New Directions in Program Evaluation*, edited by L. B. Sechrest and A. G. Scott. San Francisco: Jossey-Bass.

———. 2014. "Generalizing Causal Knowledge in the Policy Sciences: External Validity As A Task Of Both Multiattribute Representation and Multiattribute Extrapolation." *Journal of Policy Analysis and Management* 33(2): 527–36.

Cook, Thomas D., and Donald T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand McNally.

Cook, Thomas D., Harris M. Cooper, David S. Cordray, Heidi Hartmann, Larry V. Hedges, Richard J. Light, Thomas A. Louis, and Frederick Mosteller, eds. 1992. *Meta-Analysis for Explanation: A Casebook*. New York: Russell Sage Foundation.

Cronbach, Lee J. 1980. *Toward Reform of Program Evaluation*. San Francisco: Jossey-Bass Publishers.

———. 1982. *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey-Bass.

Devine, Elizabeth C. 1992. "Effects of Psychoeducational Care with Adult Surgical Patients: A Theory Probing Meta-Analysis of Intervention Studies." In *Meta-Analysis for Explanation: A Casebook*, edited by Thomas D. Cook et al. New York: Russell Sage Foundation.

Doi, Suhail A., Jan J. Barendregt, Shanjahan Khan, Lukman Thalib, and Gail M. Williams. 2015. "Simulation Comparison of the Quality Effects and Random Effects Methods of Meta-Analysis." *Epidemiology* 26(4): e42–4. DOI: 10.1097/EDE.0000000000000289.

Doi, Suhail A., and Lukman Thalib. 2008. "A Quality-Effects Model for Meta-Analysis." *Epidemiology* 19(1): 94–100. DOI: 10.1097/EDE.0b013e31815c24e7.

———. 2009. "an Alternative Quality Adjustor for the Quality Effects Model for Meta-Analysis." *Epidemiology* 20(2): 314. DOI: 10.1097/EDE.0b013e318196a8d0.

Donegan, Sarah, Lisa Williams, Sofia Dias, Catrin Tudur-Smith, and Nicky Welton. 2015. "Exploring Treatment by Covariate Interactions Using Subgroup Analysis and Meta-Regression in Cochrane Reviews: A Review of Recent Practice." *PLoS ONE* 10 (6):e0128804. DOI: 10.1371/journal.pone.0128804.

Donner, Allan, Gilda Piaggio, and José Villar. 2001. "Statistical Methods for the Meta-Analysis of Cluster Randomization Trials." *Statistical Methods in Medical Research* 10(5): 325–38.

———. 2003. "Meta-Analyses of Cluster Randomization Trials. Power Considerations." *Eval Health Prof* 26(3): 340–51.

Duval, Sue, and Richard Tweedie. 2000. "Trim and Fill: A Simple Funnel-Plot-Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis." *Biometrics* 56(2): 455–63.

Ebell, Mark H., Jay Siwek, Barry D. Weiss, Steven H. Woolf, Jeffrey Susman, Bernard Ewigman, and Marjorie Bowman. 2004. "Strength of Recommendation Taxonomy (SORT): A Patient-Centered Approach to Grading Evidence in the Medical Literature." *Journal of the American Board of Family Practice* 17(1): 59–67.

Eddy, David M., Vic Hasselblad, and Ross Shachter. 1990. "An Introduction to a Bayesian Method for Meta-Analysis: The Confidence Profile Method." *Medical Decision Making* 10(1): 15–23.

Egger, Matthias, George Davey Smith, Martin Schneider, and Christoph Minder. 1997. "Bias in Meta-Analysis Detected by a Simple, Graphical Test." *British Medical Journal* 315(7109): 629–34.

Engels, Eric A., Christopher H. Schmid, Norma Terrin, Ingram Olkin, and Joseph Lau. 2000. "Heterogeneity and Statistical Significance in Meta-Analysis: An Empirical Study of 125 Meta-Analyses." *Statistics in Medicine* 19(13): 1707–28.

Gasking, Douglas. 1955. "Causation and Recipes." *Mind* 64(256): 479–87.

Gavaghan, David J., R. Andrew Moore, and Henry J. McQuay. 2000. "An Evaluation of Homogeneity Tests in Meta-Analyses in Pain Using Simulations of Individual Patient Data." *Pain* 85(3): 415–24.

Glass, Gene V. 1976. "Primary, Secondary, and Meta-Analysis of Research." *Educational Researcher* 5(10): 3–8.

———. 1978. "Integrating Findings: The Meta-Analysis of Research." *Review of Research in Education* 5: 351–79.

Gnambs, Timo, and Thomas Staufenbiel. 2016. "Parameter Accuracy in Meta-Analyses of Factor Structures." *Research Synthesis Methods* 7(2): 168–86. DOI: 10.1002/jrsm.1190.

Greenwald, Anthony G. 1975. "Consequences of Prejudice Against the Null Hypothesis." *Psychological Bulletin* 82(1): 1–20.

Guyatt, Gordon H., Andrew D. Oxman, Regina Kunz, Roman Jaeschke, Mark Helfand, A. Liberati, Gunn Elisabeth Vist, Holger J. Schunemann, and GRADE Working Group. 2008. "GRADE: Incorporating Considerations of Resources Use into Grading Recommendations." *British Medical Journal* 336(7654): 1170–73. DOI: 10.1136/bmj.39504.506319.80.

Hardy, Rebecca J., and Simon G. Thompson. 1998. "Detecting and Describing Heterogeneity in Meta-Analysis." *Statistics in Medicine* 17(8): 841–56.

Harris, Monica J., and Robert Rosenthal. 1985. "Mediation of Interpersonal Expectancy Effects: 31 Meta-Analyses." *Psychological Bulletin* 97(3): 363–86.

Harwell, Michael. 1997. "An Empirical Study of Hedge's Homogeneity Test." *Psychological Methods* 2(2): 219–31.

Hedges, Larry V. 1982. "Estimation of Effect Sizes from a Series of Independent Experiments." *Psychological Bulletin* 92(2): 490–99.

———. 1990. "Directions for Future Methodology." In *The Future of Meta-Analysis*, edited by Kenneth W. Wachter and Miron L. Straf. New York: Russell Sage Foundation.

———. 2016. "Applying Meta-Analysis to Structural Equation Modeling." *Research Synthesis Methods* 7(2): 209–14. DOI: 10.1002/jrsm.1214.

Hedges, Larry V., and Betsy J. Becker. 1986. "Statistical Methods in the Meta-Analysis of Research on Gender Differences." In *The Psychology of Gender: Advances Through Meta-Analysis*, edited by Janet S. Hyde and Marcia C. Linn. Baltimore, Md.: Johns Hopkins University Press.

Hedges, Larry V., and Ingram Olkin. 1985. *Statistical Methods for Meta-Analysis*. Orlando, Fl.: Academic Press.

Hedges, Larry V., and Therese D. Pigott. 2001. "The Power of Statistical Tests in Meta-Analysis." *Psychological Methods* 6(3): 203–17. DOI: 10.1037/1082–989X.6.3.203.

———. 2004. "The Power of Statistical Tests for Moderators in Meta-Analysis." *Psychological Methods* 9(4): 426–45. DOI: 10.1037/1082–989X.9.4.426.

Hedges, Larry V., Elizabeth Tipton, and Matthew C. Johnson. 2010. "Robust Variance Estimation in Meta-Regression with Dependent Effect Size Estimates." *Research Synthesis Methods* 1(1): 39–65. DOI: 10.1002/jrsm.5.

Hedges, Larry V., and Jack L. Vevea. 1996. "Estimating Effect Size Under Publication Bias: Small Sample Properties and Robustness of a Random Effects Selection Model." *Journal of Educational and Behavioral Statistics* 21(4): 299–332.

———. 1998. "Fixed- and Random-Effects Models in Meta-Analysis." *Psychological Methods* 3(4): 486–504. DOI: 10.1037/1082–989x.3.4.486.

Higgins, Julian P. T., Dan Jackson, Jessica K. Barrett, Guobing Lu, Anthony E. Ades, and Ian R. White. 2012. "Consistency and Inconsistency in Network Meta-Analysis:

Concepts and Models for Multi-Arm Studies." *Research Synthesis Methods* 3(2): 98–110. DOI: 10.1002/jrsm.1044.

Higgins, Julian P. T., Craig Ramsay, Barnaby C. Reeves, Jonathan J. Deeks, Beverley Shea, Jeffrey C. Valentine, Peter Tugwell, and George Wells. 2013. "Issues Relating to Study Design and Risk of Bias When Including Non-Randomized Studies in Systematic Reviews on the Effects of Interventions." *Research Synthesis Methods* 4(1): 12–25. DOI: 10.1002/jrsm.1056.

Hox, Joop J., and Edith D. de Leeuw. 2003. "Multilevel Models for Meta-Analysis." In *Multilevel Modeling: Methodological Advances, Issues, and Applications*, edited by Steven P. Reise and Naihua Duan. Mahwah, N.J.: Lawrence Erlbaum.

Hunter, John E., and Frank L. Schmidt. 2004. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*, 2nd ed. Newbury Park, Calif.: Sage Publications.

Hunter, John E., Frank L. Schmidt, and Huy Le. 2006. "Implications of Direct and Indirect Range Restriction for Meta-Analysis Methods and Findings." *Journal of Applied Psychology* 91(3): 594–612. DOI: 10.1037/0021-9010.91.3.594.

Imbens, Guido, and Thomas Lemieux. 2008. "Special Issue Editors' Introduction: The Regression Discontinuity Design—Theory and Applications." *Journal of Econometrics* 142(2): 611–14. DOI: 10.1016/j.jeconom.2007.05.008.

Jackson, Dan. 2006. "The Power of the Standard Test for the Presence of Heterogeneity in Meta-Analysis." *Statistics in Medicine* 25(15): 2688–99.

Jacob, Brian, and Jens Ludwig. 2005. "Can the Federal Government Improve Education Research?" *Brookings Papers on Education Policy* 8: 47–87.

Kelley, George A., Kristi S. Kelley, and Zung Vu Tran. 2004. "Retrieval of Missing Data for Meta-Analysis: A Practical Example." *International Journal of Technology Assessment in Health Care* 20(3): 296–99.

Kulinskaya, Elena, Michael B. Dollinger, and Kirsten Bjørkestøl. 2011a. "Testing for Homogeneity in Meta-Analysis I: The One-Parameter Case: Standardized Mean Difference." *Biometrics* 67(1): 203–12. DOI: 10.1111/j.1541-0420.2010.01442.x.

———. 2011b. "On the Moments of Cochran's Q Statistic Under the Null Hypothesis, with Application to the Meta-Analysis of Risk Difference." *Research Synthesis Methods* 2(4): 254–70. DOI: 10.1002/jrsm.54.

Le, Huy, and Frank L. Schmidt. 2006. "Correcting for Indirect Range Restriction in Meta-Analysis: Testing a New Meta-Analytic Procedure." *Psychological Methods* 11(4): 416–438. DOI: 10.1037/1082–989X.11.4.416.

Lipsey, Mark W. 1992. "Juvenile Delinquency Treatment: A Meta-Analytic Inquiry into the Variability of Effects." In *Meta-Analysis for Explanation: A Casebook*, edited by Thomas D. Cook et al. New York: Russell Sage Foundation.

———. 2003. "Those Confounded Moderators in Meta-Analysis: Good, Bad, and Ugly." *Annals of the American Academy of Political and Social Science* 587(1): 69–81.

Lipsey, Mark W., and David B. Wilson. 2001. *Practical Meta-Analysis*. Applied Social Research Methods Series, vol. 49. Thousand Oaks, Calif.: Sage Publications.

Little, Roderick J. A., and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*, 2nd ed. New York: John Wiley & Sons.

Mackie, John L. 1974. *The Cement of the Universe: A Study of Causation*. Oxford: Oxford University Press.

Marín-Martínez, Fulgencio, and Julio Sánchez-Meca. 2010. "Weighting by Inverse Variance or by Sample Size in Random-Effects Meta-Analysis." *Educational and Psychological Measurement* 70(1): 56–73. DOI: 10.1177/0013164409344534.

Matt, Georg E. 1989. "Decision Rules for Selecting Effect Sizes in Meta-Analysis: A Review and Reanalysis of Psychotherapy Outcome Studies." *Psychological Bulletin* 105(1): 106–15.

———. 2003. "Will It Work in Münster? Meta-Analysis and the Empirical Generalization of Causal Relationships." In *Meta-Analysis*, edited by Heinz Holling, Dankmar Böhning, and Ralf Schulze. Berlin: Springer.

———. 2005. "Uncertainty in the Multivariate World: A Fuzzy Look Through Brunswik's Lens." In *Multivariate Research Strategies: A Festschrift in Honor of Werner W. Wittmann*, edited by André Beauducel et al. Maastricht: Shaker Publishing.

McLeod, Bryce D., John R. Weisz, and Jeffrey J. Wood. 2007. "Examining the Association Between Parenting and Childhood Depression: A Meta-Analysis." *Clinical Psychology Review* 27(8): 986–1003. DOI: 10.1016/j.cpr.2007.03.001.

Mick, Eric, Joseph Biederman, Gahan Pandina, and Stephen V. Faraone. 2003. "A Preliminary Meta-Analysis of the Child Behavior Checklist in Pediatric Bipolar Disorder." *Biological Psychiatry* 53(11): 1021–27.

Moyer, Anne, John W. Finney, Carolyn E. Swearingen, and Pamela Vergun. 2002. "Brief Interventions for Alcohol Problems: A Meta-Analytic Review of Controlled Investigations in Treatment-Seeking and Non-Treatment-Seeking Populations." *Addiction* 97(3): 279–92.

Nam, In-Sum, Kerrie Mengersen, and Paul Garthwaite. 2003. "Multivariate Meta-Analysis." *Statistics in Medicine* 22(14): 2309–33.

Norris, Susan L., David Moher, Barnaby C. Reeves, Beverley Shea, Yoon Loke, Sarah Garner, Laurie Anderson, Peter

Tugwell, and George Wells. 2013. "Issues Relating to Selective Reporting When Including Non-Randomized Studies in Systematic Reviews on the Effects of Healthcare Interventions." *Research Synthesis Methods* 4(1): 36–47. DOI: 10.1002/jrsm.1062.

Olkin, Ingram. 2012. "Retrieving Treatment and Control Proportions from Incomplete Summary Data in Meta-Analysis." *Research Synthesis Methods* 3(3): 250–54. DOI: 10.1002/jrsm.1043.

Oort, Frans J., and Suzanne Jak. 2016. "Maximum Likelihood Estimation in Meta-Analytic Structural Equation Modeling." *Research Synthesis Methods* 7(2): 156–67. DOI: 10.1002/jrsm.1203.

Orwin, Robert G., and David S. Cordray. 1985. "The Effects of Deficient Reporting on Meta-Analysis: A Conceptual Framework and Reanalysis." *Psychological Bulletin* 97(1): 134–47.

Peters, Jaime L., Alexander J. Sutton, David R. Jones, Keith R. Abrams, and Lesley Rushton. 2006. "Comparison of Two Methods to Detect Publication Bias in Meta-Analysis." *Journal of the American Medical Association* 295(6): 676–80. DOI: 10.1001/jama.295.6.676.

Premack, Steven L., and John E. Hunter. 1988. "Individual Unionization Decisions." *Psychological Bulletin* 103(2): 223–34.

Prevost, Teresa C., Keith R. Abrams, and David R. Jones. 2000. "Hierarchical Models in Generalized Synthesis of Evidence: An Example Based on Studies of Breast Cancer Screening." *Statistics in Medicine* 19(24): 3359–76.

Raudenbush, Stephen W., Betsy J. Becker, and Hripsime Kalaian. 1988. "Modeling Multivariate Effect Sizes." *Psychological Bulletin* 103(1): 111–20.

Raudenbush, Stephen W., and Anthony S. Bryk. 1985. "Empirical Bayes Meta-Analysis." *Journal of Educational Statistics* 10(1): 75–98.

Reeves, Barnaby C., Julian P. T. Higgins, Craig Ramsay, Beverley Shea, Peter Tugwell, and George A. Wells. 2013. "An Introduction to Methodological Issues When Including Non-Randomised Studies in Systematic Reviews on the Effects of Interventions." *Research Synthesis Methods* 4(1): 1–11. DOI: 10.1002/jrsm.1068.

Robinson, Leslie A., Jeffrey S. Berman, and Robert A. Neimeyer. 1990. "Psychotherapy for the Treatment of Depression: A Comprehensive Review of Controlled Outcome Research." *Psychological Bulletin* 108: 30–49.

Rosch, Eleanor. 1973. "Natural Categories." *Cognitive Psychology* 4(3): 328–50.

———. 1978. "Principles in Categorization." In *Cognition and Categorization*, edited by Eleanor Rosch and Barbara B. Lloyd. Hillsdale, N.J.: Lawrence Erlbaum.

Rosenberg, Michael S. 2010. "A Generalized Formula for Converting Chi-Square Tests to Effect Sizes for Meta-Analysis." *PLoS ONE* 5(4): e10059. DOI: 10.1371/journal.pone.0010059.

Rosenthal, Robert. 1979. "The File Drawer Problem and Tolerance for Null Results." *Psychological Bulletin* 86(3): 638–41.

———. 1990. "An Evaluation of Procedure and Results." In *the Future of Meta-Analysis*, edited by Kenneth W. Wachter and Miron L. Straf. New York: Russell Sage Foundation.

———. 1991. *Meta-Analytic Procedures for Social Research*, 2nd ed. Newbury Park: Sage Publications.

Rosenthal, Robert, and Donald B. Rubin. 1982. "Comparing Effect Sizes of Independent Studies." *Psychological Bulletin* 92(2): 500–504.

———. 1986. "Meta-Analytic Procedures for Combining Studies with Multiple Effect Sizes." *Psychological Bulletin* 99(3): 400–406.

Rothstein, Hannah R., Alexander J. Sutton, and Michael Borenstein, eds. 2005. *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. New York: John Wiley & Sons.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Sacks, Henry S., Jayne Berrier, Dinah Reitman, V. A. Ancona-Berk, and Thomas C. Chalmers. 1987. "Meta-Analyses of Randomized Controlled Trials." *New England Journal of Medicine* 316(8): 450–55.

Salanti, Georgia. 2012. "Indirect and Mixed-Treatment Comparison, Network, or Multiple-Treatments Meta-Analysis: Many Names, Many Benefits, Many Concerns for the Next Generation Evidence Synthesis Tool." *Research Synthesis Methods* 3(2): 80–97. DOI: 10.1002/jrsm.1037.

Salanti, Georgia, Julian P. T. Higgins, Anthony E. Ades, and John P. A. Ioannidis. 2008. "Evaluation of Networks of Randomized Trials." *Statistical Methods in Medical Research* 17(3): 279–301. DOI: 10.1177/0962280207080643.

Saleh, A. K. Ehsanes, Khatab M. Hassanein, Ruth S. Hassanein, and Hyo Mi Kim. 2006. "Quasi-Empirical Bayes Methodology for Improving Meta-Analysis." *Journal of Biopharmaceutical Statistics* 16(1): 77–90.

Salgado, Jesús F., Silvia Moscoso, and Neil Anderson. 2016. "Corrections for Criterion Reliability in Validity Generalization: The Consistency of Hermes, the Utility of Midas." *Revista de Psicología del Trabajo y de las Organizaciones* 32(1): 17–23. DOI: 10.1016/j.rpto.2015.12.001.

Sánchez-Meca, Julio, Fulgencio Marín-Martínez, and Salvador Chacon-Moscoso. 2003. "Effect-Size Indices for Dichotomized Outcomes in Meta-Analysis." *Psychological Methods* 8(4): 448–67.

Schulze, Ralf. 2004. *Meta-Analysis: A Comparison of Approaches*. New York: Hogrefe and Huber.

Scott, J. Cobb, Georg E. Matt, Kristen M. Wrocklage, Cassandra Crnich, Jessica Jordan, Steven M. Southwick, John H. Krystal, and Brian C. Schweinsburg. 2015. "A Quantitative Meta-Analysis of Neurocognitive Functioning in Posttraumatic Stress Disorder." *Psychological Bulletin* 141(1): 105–40. DOI: 10.1037/a0038039.

Scott, J. Cobb, Steven P. Woods, Georg E. Matt, Rachel A. Meyer, Robert K. Heaton, J. Hampton Atkinson, and Igor Grant. 2007. "Neurocognitive Effects of Methamphetamine: A Critical Review and Meta-Analysis." *Neuropsychology Review* 17(3): 275–97. DOI: 10.1007/s11065–007–9031–0.

Shadish, William R. 1992. "Do Family and Marital Therapies Change What People Do? A Meta-Analysis of Behavioral Outcomes." In *Meta-Analysis for Explanation: A Casebook*, edited by Thomas D. Cook et al. New York: Russell Sage Foundation.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2001. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, Mass.: Houghton Mifflin.

Shadish, William R., Xiangen Hu, Renita R. Glaser, Richard Kownacki, and Seok Wong. 1998. "A Method for Exploring the Effects of Attrition in Randomized Experiments with Dichotomous Outcomes." *Psychological Methods* 3(1): 3–22.

Shadish, William R., Jason K. Luellen, and M. H. Clark. 2006. "Propensity Scores and Quasi-Experiments: A Testimony to the Practical Side of Lee Sechrest." In *Strengthening Research Methodology: Psychological Measurement and Evaluation*, edited by Richard Bootzin and Patrick E. McKnight. Washington, D.C.: American Psychological Association.

Shadish, William R., Georg E. Matt, Ana Martinez Navarro, and George Phillips. 2000. "The Effects of Psychological Therapies under Clinically Representative Conditions: A Meta-Analysis." *Psychological Bulletin* 126(4): 512–29.

Shadish, William R., and Rebecca B. Sweeney. 1991. "Mediators and Moderators in Meta-Analysis: There's a Reason We Don't Let Dodo Birds Tell Us Which Psychotherapies Should Have Prizes." *Journal of Consulting and Clinical Psychology* 59(6): 883–93.

Shapiro, David A., and Diana Shapiro. 1982. "Meta-Analysis of Comparative Therapy Outcome Studies: A Replication and Refinement." *Psychological Bulletin* 92(3): 581–604.

Shemilt, Ian, Miranda Mugford, Sarah Byford, Mike Drummond, Eric Eisenstein, Martin Knapp, Jacqueline Mallender, Kevin Marsh, David McDaid, Luke Vale, and Damian Walker (Co-convenors of the Campbell and Cochrane Economics Methods Group). 2008. "The Campbell Collaboration: Economic Methods Policy Brief." *C2 Methods* policy brief. Oslo: The Campbell Collaboration. Accessed December 18, 2018. https://campbellcollaboration.org/images/pdf/plain-language/Economic_Methods_Policy_Brief.pdf.

Sheng, Zitong, Wenmo Kong, José M. Cortina, and Shuofei Hou. 2016. "Analyzing Matrices of Meta-Analytic Correlations: Current Practices and Recommendations." *Research Synthesis Methods* 7(2): 187–208. DOI: 10.1002/jrsm.1206.

Sinharay, Sandip, Hal S. Stern, and Daniel Russell. 2001. "The Use of Multiple Imputation for the Analysis of Missing Data." *Psychological Methods* 6: 317–29.

Slavin, Robert E. 1986. "Best-Evidence Synthesis: An Alternative to Meta-Analytic and Traditional Reviews." *Educational Researchers* 15(1): 5–11.

Smith, Edward E., and Douglas L. Medin. 1981. *Categories and Concepts*. Cambridge, Mass.: Harvard University Press.

Smith, Mary Lee, Gene V. Glass, and Thomas I. Miller. 1980. *The Benefits of Psychotherapy*. Baltimore, Md.: Johns Hopkins University Press.

Song, Fujian, Douglas G. Altman, Anne-Marie Glenny, and Jonathan J. Deeks. 2003. "Validity of Indirect Comparison for Estimating Efficacy of Competing Interventions: Empirical Evidence from Published Meta-Analyses." *British Medical Journal* 326(7387): 472–72.

Sterne, Jonathan A. C., and Matthias Egger. 2001. "Funnel Plots for Detecting Bias in Meta-Analysis: Guidelines on Choice of Axis." *J Clin Epidemiol* 54(10): 1046–55.

Sutton, Alexander J. 2000. *Methods for Meta-Analysis in Medical Research*, *Wiley Series in Probability and Mathematical Statistics.* New York: John Wiley & Sons.

Sutton, Alexander J., and Keith R. Abrams. 2001. "Bayesian Methods in Meta-Analysis and Evidence Synthesis." *Statistical Methods in Medical Research* 10(4): 277–303.

Sutton, Alexander J., Susan J. Duval, Richard L. Tweedie, Keith R. Abrams, and David R. Jones. 2000. "Empirical Assessment of Effect of Publication Bias on Meta-Analyses." *BMJ* 320(7249): 1574–7.

Sutton, Alexander J., Fujian Song, Simon M. Gilbody, and Keith R. Abrams. 2000. "Modelling Publication Bias in Meta-Analysis: A Review." *Stat Methods Med Res* 9(5): 421–45.

Tabarrok, Alexander T. 2000. "Assessing the FDA Via the Anomaly of off-Label Drug Prescribing." *Independent Review* V(1): 25–53.

Tanner-Smith, Emily E., and Elizabeth Tipton. 2014. "Robust Variance Estimation with Dependent Effect Sizes: Practical Considerations Including a Software Tutorial in Stata and

SPSS." *Research Synthesis Methods* 5(1): 13–30. DOI: 10.1002/jrsm.1091.

Twenge, Jean M., and W. Keith Campbell. 2001. "Age and Birth Cohort Differences in Self-Esteem: A Cross-Temporal Meta-Analysis." *Personality and Social Psychology Review* 5(4): 321–44.

Valentine, Jeffrey C., and Cathleen M. McHugh. 2007. "The Effects of Attrition on Baseline Comparability in Randomized Experiments in Education: A Meta-Analysis." *Psychological Methods* 12(3): 268–82. DOI: 10.1037/1082-989X.12.3.268.

Valentine, Jeffrey C., Therese D. Pigott, and Hannah R. Rothstein. 2010. "How Many Studies Do You Need? a Primer on Statistical Power for Meta-Analysis." *Journal of Educational and Behavioral Statistics* 35(3): 375–75. DOI: 10.3102/1076998610376621.

Valentine, Jeffrey C., and Simon G. Thompson. 2013. "Issues Relating to Confounding and Meta-Analysis When Including Non-Randomized Studies in Systematic Reviews on the Effects of Interventions." *Research Synthesis Methods* 4(1): 26–35. DOI: 10.1002/jrsm.1064.

van Houwelingen, Hans C., Lidia R. Arends, and Theo Stijnen. 2002. "Advanced Methods in Meta-Analysis: Multivariate Approach and Meta-Regression." *Stat Med* 21(4): 589–624.

Vevea, Jack L., Nancy C. Clements, and Larry V. Hedges. 1993. "Assessing the Effects of Selection Bias on Validity Data for the General Aptitude Test Battery." *Journal of Applied Psychology* 78(6): 981–987.

Vevea, Jack L., and Larry V. Hedges. 1995. "A General Linear Model for Estimating Effect Size in the Presence of Publication Bias." *Psychometrika* 60(3): 419–35.

Vevea, Jack L., and Carol M. Woods. 2005. "Publication Bias in Research Synthesis: Sensitivity Analysis Using A Priori Weight Functions." *Psychological Methods* 10(4): 428–43. DOI: 10.1037/1082-989X.10.4.428.

Viechtbauer, Wolfgang. 2007. "Confidence Intervals for the Amount of Heterogeneity in Meta-Analysis." *Statistics in Medicine* 26(1): 37–52.

Webb, Eugene J., Donald T. Campbell, Richard D. Schwartz, Lee Sechrest, and Janet B. Grove. 1981. *Nonreactive Measures in the Social Sciences*, 2nd ed. Chicago: Rand McNally.

Wells, George A., Beverley Shea, Julian P. T. Higgins, Jonathan Sterne, Peter Tugwell, and Barnaby C. Reeves. 2013. "Checklists of Methodological Issues for Review Authors to Consider When Including Non-Randomized Studies in Systematic Reviews." *Research Synthesis Methods* 4(1): 63–77. DOI: 10.1002/jrsm.1077.

West, Stephen G., Jeremy C. Biesanz, and Steven C. Pitts. 2000. "Causal Inference and Generalization in Field Settings: Experimental and Quasi-Experimental Designs." In *Handbook of Research Methods in Social and Personality Psychology*, edited by Harry T. Reid and Charles M. Judd. New York: Cambridge University Press.

Whitbeck, Caroline. 1977. "Causation in Medicine: The Disease Entity Model." *Philosophy of Science* 44(4): 619–37.

White, Ian R., Jessica K. Barrett, Dan Jackson, and Julian P. T. Higgins. 2012. "Consistency and Inconsistency in Network Meta-Analysis: Model Estimation Using Multivariate Meta-Regression." *Research Synthesis Methods* 3(2): 111–25. DOI: 10.1002/jrsm.1045.

Wilson, Sandra Jo, Joshua R. Polanin, and Mark W. Lipsey. 2016. "Fitting Meta-Analytic Structural Equation Models with Complex Datasets." *Research Synthesis Methods* 7(2): 121–39. DOI: 10.1002/jrsm.1199.

Wittmann, Werner W., and Georg E. Matt. 1986. "Meta-Analyse als Integration von Forschungsergebnissen am Beispiel deutschsprachiger Arbeiten zur Effektivität von Psychotherapie" (Integration of German-language psychotherapy outcome studies through meta-analysis). *Psychologische Rundschau* 37(1): 20–40.

Wolf, Frederic M. 1990. "Methodological Observations on Bias." In *The Future of Meta-Analysis*, edited by Kenneth W. Wachter and Miron L. Straf. New York: Russell Sage Foundation.

Yuan, Ke-Hai. 2016. "Meta Analytical Structural Equation Modeling: Comments on Issues with Current Methods and Viable Alternatives." *Research Synthesis Methods* 7(2): 215–31. DOI: 10.1002/jrsm.1213.

Yuan, Ying, and Roderick J. A. Little. 2009. "Meta-Analysis of Studies with Missing Data." *Biometrics* 65(2): 487–96. DOI: 10.1111/j.1541-0420.2008.01068.x.

23

POTENTIALS AND LIMITATIONS OF RESEARCH SYNTHESIS

HARRIS COOPER
Duke University

LARRY V. HEDGES
Northwestern University

JEFFREY C. VALENTINE
University of Louisville

CONTENTS

23.1 INTRODUCTION

We suspect that readers of this volume reacted to it in one of two ways: some may have been overwhelmed by the number and complexity of the issues that arise in a research synthesis. Alternatively, and perhaps simultaneously, some may have been delighted to have available a manual to help them through the synthesis process. We further suspect that the reaction experienced depended on whether the book was read while the reader was thinking about or while actually performing a synthesis. In the abstract, the concerns raised in this handbook may seem daunting. Concretely however, research syntheses have been carried out for decades and will be for decades to come, and the problems encountered in their conduct do not really go away when they are ignored. Knowledge builders need a construction manual to accompany their blueprints and bricks.

Further, there is no reason to believe that a sound research synthesis is any more complex than sound primary research. The analogies between research synthesis and survey research, content analysis, or even quasi-experimentation fit too well for this not to be the case. Certainly, each type of research has its unique characteristics. Still, when research synthesis procedures have become as familiar as primary research procedures, much of what now seems overwhelming will come to be viewed as difficult, but manageable, and obligatory.

Likewise, expectations have changed. As mentioned in chapter 1, we do not expect a primary study to provide unequivocal answers to research questions, nor do we expect it to be performed without flaw. Research syntheses are the same. The perfect research synthesis does not and never will exist. Synthesists who set standards that require them to implement all of the most rigorous techniques described in this book are bound to be disappointed. Time and other resources (not to mention the logic of discovery) will prevent such an accomplishment. The synthesist, like the primary researcher, must balance methodological rigor against practical feasibility.

In this chapter, we revisit some of the major issues that have emerged from the preceding pages. We state once more unique contributions that research synthesis can make to our understanding of scientific evidence. We suggest a few limitations and offer our hopes for how the usefulness of research syntheses can improve in the future. Finally, we briefly share some perspectives on what makes a knowledge synthesis of any kind, not just a research synthesis, most valuable to its readership.

23.2 UNIQUE CONTRIBUTIONS OF RESEARCH SYNTHESIS

23.2.1 Increased Precision and Reliability

Some comparisons were made in chapter 1 between primary research and research synthesis. It was suggested that the two forms of study had much in common. In contrast, when we compare research synthesis as it was practiced prior to the introduction of meta-analytic techniques it is the dissimilarities that capture attention. Perhaps most striking are the improvements in precision and reliability that new techniques have made possible.

Precision in literature searches has been improved dramatically by the procedures described herein. The revolution in information wrought by the internet has changed the way science is conducted and communicated. The development and maintenance of comprehensive research databases mean that a searcher for scientific literature can reach into thousands of journals. The proliferation of research registers means that literature searchers can know about research that is ongoing and can even reach into another researcher's data files. In addition, the near limitless information capacity of computers in the digital age means that the common excuse for incomplete reporting (that journal space is limited) is moot. Although the present process is far from perfect, it is an enormous improvement over past practice.

Related to discovering what literature is "out there" is access to the literature itself. The use of digital means to transmit copies of documents has simplified and sped immensely the acquisition of research reports. Gone are the days of spending hours in the library copying journal articles and dissertations (most of the latter were archived on microfiche, making copying particularly painful). Institutional subscriptions mean research intensive institutions can deliver electronic copies of scholarly work to searchers in seconds. And, open-access journals are proliferating, making much research available to anyone with a connection to the internet (see the Directory of Open Access Journals, https://doaj.org/).

Finally, the movement to promote data sharing (for example, National Science Foundation, http://www.nsf.gov/bfa/dias/policy/dmp.jsp) adds yet another level of potential precision and reliability to research syntheses. Now, synthesists can go directly to the data used in experiments to recalculate statistics of interest and even to generate new statistics that were not included in research reports.

The coding of studies and their evaluation for quality has also come a long way. The old process was rife with potential bias. In 1976, Gene Glass wrote,

> A common method for integrating several studies with inconsistent findings is to carp on the design or analysis deficiencies of all but a few studies—those remaining frequently being one's own work or that of one's students and friends—and then advancing the one or two "acceptable" studies as the truth of the matter. (4)

Today, research synthesists collect, code, and evaluate their studies with considerably more care to avoid bias. Explicit, well-defined coding frames are used by multiple coders. Judgments about quality of research are often eschewed entirely, allowing the data to determine if research design influences study outcomes.

Finally, the process of integrating and contrasting primary study results has taken a quantum leap forward. It is no longer acceptable to string together paragraph descriptions of studies, with details of significance tests, and then conclude that "the data are inconsistent, but seem to indicate . . ." Today's synthesist can provide (a) confidence and prediction intervals around effect-size estimates (b) for separate parts of a literature distinguished by both methodological and theoretical criteria (c) calculated several different ways using different assumptions about the adequacy of the literature search and coding scheme as well as different underlying statistical models.

These are only a few of the ways that the precision and reliability of current syntheses outstrip those of their predecessors. The increases in precision and reliability that a synthesist can achieve when a body of literature grows from two to twenty to two hundred studies are squandered with less rigorous methods.

23.2.2 Testing Generalizability of Claims

Current methods not only capitalize on accumulating evidence by making estimates more precise, but also permit the testing of hypotheses that may have never been tested in primary studies. This advantage is most evident when the objective of the synthesist is to assess the generality or specificity of results. The ability of a synthesist to test whether a finding appears to vary across settings, times, people, measurements, and researchers surpasses by far the ability of most primary studies. However, for yesterday's synthesist, the variation among studies was a nuisance; it clouded interpretation. The old methods of synthesis were handicapped severely when it came to judgments on whether a set of studies revealed consistent results and, if not, what might account for the inconsistency. Yesterday's synthesist often chose the wrong datum (statistical significance tests rather than effect sizes) and the wrong evaluation criteria (implicit cognitive algebra rather than formal statistical test) on which to base these judgments. For today's synthesist, variety is the spice of life. The methods described in this handbook make such analyses routine. When the outcomes of studies prove too discrepant to support a claim that one estimate of a relationship underlies them all, current techniques provide the synthesist with a way of systematically searching for moderating influences, using consistent and explicit rules of evidence.

As mentioned in the introductory chapter, the reproducibility of results in the social and medical sciences has become an issue of concern. We pointed out that attempts at direct replication of research findings seemingly often fail (Open Science Collaboration 2015). The mindset of research synthesists provides a different perspective on the replication issue. From this perspective, the results of the studies that go into the same meta-analysis are best thought of as conceptual replications, or tests of the same hypothesis, rather than direct replications. This may be true even for direct replication attempts, in part because researchers often do not have strong theories that inform them about why and how the relationships observed come about (it is hard to replicate an intervention when you do not know how the intervention works). For synthesists, then, variations in study methods are to be expected, as are variations in results.

Therefore, if studies in the social and medical sciences replicate less often than anticipated, this is perhaps not surprising. Furthermore, determining whether a study, or a series of studies, have "replicated" a given result is harder than it seems (Hedges and Schauer 2018). But perhaps more interesting from the standpoint of a synthesist is that even protocol-based attempts at replication often result in a rejection of the null hypothesis of homogeneous effect sizes. Half of the direct replications that Richard Klein and his colleagues reported rejected the null hypothesis of effect-size homogeneity, a situation that most scholars would not have predicted (2014). It suggests that it is difficult to reproduce the exact conditions under which a previous study has been conducted. We encourage synthesists to treat this as an opportunity to explore the causal mechanisms and the generality of findings across circumstances.

23.3 LIMITATIONS OF RESEARCH SYNTHESIS

23.3.1 Correlational Nature of Review Evidence

Research syntheses have limitations as well. Among the most critical is the correlational nature of synthesis-generated evidence. As detailed in chapter 2, study-generated evidence appears when individual studies contain results that directly test the relation under consideration. Synthesis-generated evidence appears when the results of studies using different procedures to test the same hypothesis are compared. As noted in chapter 2, only study-generated evidence can be used to support claims about causal relations (and even then, usually only when experimental designs are used to generate the evidence). Specific confounds can be controlled statistically at the level of synthesis-generated evidence, but the result can never lead to the same confidence in inferences produced by study-generated evidence from investigations employing random assignment. This is an inherent limitation of research syntheses. It also provides the synthesist with a fruitful source of suggestions concerning directions for future primary research.

23.3.2 Post Hoc Nature of Synthesis Tests

It is almost always the case that researchers interested in conducting syntheses set upon the process with some, if not considerable, knowledge of the empirical data base they are about to summarize. They begin with a good notion of what the literature says about which interventions work, which moderators operate, and which explanatory variables are helpful. This being the case, the synthesist cannot state a hypothesis so derived and then also conclude that the same evidence supports the hypothesis. As Kenneth Wachter and Miron Straf point out, once data have been used to develop a theory, they cannot be used to test it (1990). Synthesists who wish to test specific, a priori hypotheses in meta-analysis must go to extra lengths to convince their audience that the generation of the hypothesis and the data used to test it are truly independent.

23.3.3 Need for New Primary Research

Each of these limitations serves to underscore an obvious point: a research synthesis should never be considered a replacement for new primary research. Primary research and research synthesis are complementary parts of a common practice, not competing alternatives. The additional precision and reliability brought to syntheses by the procedures described in this text do not make research integrations substitutes for new data-gathering efforts. Instead, they help ensure that the next wave of primary research is sent off in the most illuminating direction. A synthesis that concludes a problem is solved, that no future research is needed, should be viewed with extreme skepticism. Even the best architects can see how the structures they have built can be improved.

23.4 EMERGING DEVELOPMENTS IN RESEARCH SYNTHESIS

23.4.1 Improving Data in Research Syntheses

Two problems that confront research synthesists recur throughout this text. They concern the comprehensiveness of literature retrieval processes, especially because searches are influenced by publication bias and missing data in research reports. Despite new techniques for locating research, the correspondence between the studies that can be accessed by the synthesist and the target population of studies is an issue confronted during data collection, analysis, and interpretation. Data that goes unreported in research descriptions is an issue in coding, analysis, and interpretation as well.

There is little need to reiterate the frustration that synthesists feel when they contemplate the possible biases in their conclusions caused by studies and results they cannot retrieve. Instead, we note that the earlier editions of this book made a call for improvements in research synthesis procedures related to ways to restructure the scientific information delivery systems so that more of the needed information becomes available to the research synthesist. Obviously, this has occurred, as has improvements in techniques for estimating and addressing problems associated with publication bias. That said, as Jack Vevea and Kathleen Coburn (chapter 18) suggest, there are no perfect publication bias tests. In fact, a strong argument can be made for the assertion that there are no good publication bias tests that can be applied universally. For this reason, synthesists need to be much more cautious than they typically are when making claims about publication bias, particularly its absence. There is no sense is which it is legitimate to read a reasonably symmetric funnel plot as proof that publication bias is not a problem. Furthermore, Vevea and Coburn argue for a triangulation approach to publication bias. Users can take

some comfort when multiple approaches yield approximately the same answer. It is also informative (though not very comforting!) when multiple approaches yield different conclusions.

With regard to missing data within studies, the past decade has witnessed a significant upgrading of standards for reporting of primary research results in professional journals. Numerous guidelines are now available to help ensure that research reports contain the information needed for full inclusion of studies in meta-analyses. Many of these can be found online. The EQUATOR Network (http://www.equator-network.org/) lists at least ten guidelines for health-related research. Psychology has such guidelines as well, the Journal Article Reporting Standards Working Group (Appelbaum et al. 2018). The American Educational Research Association also provides standards for reporting on empirical social science research (AERA 2006). These standards are meant to encompass the broadest range of research approaches, from experiments using random assignment to qualitative investigations.

The second development that has enormous potential for ameliorating the problem of missing data is the establishment of auxiliary web sites by journals on which authors archive information about their research that could not be included in journal reports because of space limitations. For example, the American Psychological Association provides its journals with websites for this purpose. These are referred to in the published articles and access is free to anyone who reads the journal.

Journal reporting standards and auxiliary data storage, however, are only measures for addressing the problem of missing data. They do not address the issue of reporting biases. As noted, a partial solution to reporting biases is the continued development of research registries. Because research registers attempt to catalog investigations both when they are initiated and when they are completed, they represent a unique opportunity for overcoming reporting biases. In addition, they may be able to provide research synthesists with more thorough reports of study results. Evan Mayo-Wilson and Sean Grant attest to the advance of research registers (chapter 21, this volume). We suspect that as science journals increasingly provide incentives for registering studies before data collection begins, such as by agreeing to accept pre-registered studies regardless of their results, the issue of publication bias will become more trackable by research synthesists.

Of course, registers are not a panacea. Registers are helpful only if the researchers who register their studies keep and make available the results of their studies to synthesists who request them. Thus, we renew our call in the earlier editions of this book for professional standards that promote data sharing of even unpublished research findings. Still, forward-looking researchers, publishers, and funding agencies would be well advised to look toward the adoption of reporting standards and the creation of auxiliary websites and research registers as some of the most effective means for fostering knowledge accumulation. Given the pervasiveness of the publication bias and missing data problems, these efforts seem to hold considerable promise for helping research synthesists find the missing science.

Finally, throughout this chapter we emphasize the extent to which technology has helped improve the quality of research syntheses, and the ease with which they are conducted. We anticipate that additional developments will soon occur that will fundamentally change how research syntheses are conducted. In particular, we are enthusiastic about the extent to which software can be developed and trained to semi- or even fully automate some of the tasks, like abstract screening and data coding, that currently represent large portions of the total person hours that are required for a research synthesis. For example, RobotReviewer automates several aspects of the review process, including trial identification, data extraction, and even risk of bias assessment (Marshall et al. 2018). Research syntheses require a significant investment of time and resources. If RobotReviewer and other text mining tools continue to be developed and refined, our hope is that the quality of research syntheses will increase, and the resources needed to complete them will decrease. Among many other benefits, using technology to keep existing reviews up-to-date seems particularly rewarding and promising.

23.4.2 Improving Usefulness of Research Syntheses

Although there is little doubt that research syntheses are increasingly perceived as critical to understanding what the evidence says about a particular relationship, we strongly believe that more needs to be done to make the results of syntheses more useful to consumers. Jeffrey Valentine and his colleagues Ariel Aloe and Sandra Jo Wilson note in chapter 19 that much work is needed to understand how research results are interpreted by users.

If this call is taken seriously by researchers and funders of research, then it should result in tools for effect-size translations (such as visual aids) that can be routinely presented with effect-size estimates.

In this vein, if research synthesists hope to influence public policy decisions, more efforts to help decision makers understand implementation costs are needed. Even small effects that are associated with minimal costs are likely to be worthwhile. At the same time, potentially large effects are not helpful if the intervention costs are such that policymakers cannot justify the expense.

Similarly, we noted earlier that moderator tests can help researchers understand the association between characteristics of the research design, setting, and sample and effect-size estimates. It is common for potential users of a research synthesis to ask "Should I expect to see a similar effect in my context?" This is a particularly vexing problem, and one might agree with Donald Campbell, who argues that looking for "proximal similarity" (that is, the similarity between the contexts in the research and the user's context) is in some cases about the best that can be done (1986). It is usually not enough to point users to a confidence interval or even to a prediction interval and state that these present the range of likely effects for their context. Researchers have started to think seriously about this problem (see, for example, O'Muircheartaigh and Hedges 2014; Stuart, Bradshaw, and Leaf 2015). Development of tools to implement the underlying statistical concepts can be beneficial to users of research syntheses. One example of such work is by Elizabeth Tipton and Kate Miller-Bains, who developed The Generalizer (https://www.thegeneralizer.org/). In our view, this work is promising and our hope is that researchers continue to refine and test both the underlying statistical methods and different tools for presenting these methods to users of research syntheses.

Finally, network meta-analysis has become popular in clinical research, and our sense is that wider exploitation of this method will be helpful to those who would like to use research syntheses to inform decision making. The primary benefit of network meta-analysis is that it provides a formal method for evaluating indirect comparisons. For example, research synthesists might be interested in school-based efforts to improve children's sense of belonging in school. There might be several types of interventions (such as teacher professional development on engaging children, after-school programs, and parent involvement programs). Most studies will likely examine the effect of one of the interventions relative to a business-as-usual control. A few might examine one intervention compared with another intervention. Network meta-analysis involves integrating the indirect comparisons (those implied by the fact that the interventions are assessed against a common comparator) and the direct comparisons (for example, teacher professional development versus an after-school program). Decision making often involves trying to determine which alternative course of action is "best," and one of the appealing aspects of network meta-analysis is that the methodology maps well onto this reality. Readers interested in learning more about network meta-analysis should find the primer by Dimitris Mavridis and his colleagues (2015) to be an excellent introduction.

23.5 Criteria for Judging the Quality of Research Syntheses

In the previous chapter, Georg Matt and Thomas Cook began the process of extracting wisdom from technique. They attempt to help synthesists take the broad view of their work, to hold it up to standards that begin answering the question, "How valid (trustworthy) is this synthesis?"

The criteria for quality syntheses can vary with the needs of the reader. As long ago as 1981, Susan Cozzens found that readers using literature reviews to follow developments within their area of expertise valued comprehensiveness and did not consider important the reputation of the author. However, if the synthesis was outside the reader's area, the expertise of the author was important, as was brevity. Kenneth Strike and George Posner also address the quality of knowledge synthesis with a wide-angle lens (1983). They suggest that a valuable knowledge synthesis must have both intellectual quality and practical utility. A synthesis should clarify and resolve issues in a literature rather than obscure them. It should result in a progressive paradigm shift, that is, bring to a theory greater explanatory power, to a practical program expanded scope of application, and to future primary research an increased capacity to pursue unsolved problems. Always, a knowledge synthesis should answer the question asked; readers should be provided a sense of closure (or at the least a better sense of how the future research should proceed).

In fact, readers have their own criteria for what makes a good synthesis. Cooper asked post–master's degree students to read meta-analyses and to evaluate five separate aspects of the papers (1986). The results revealed that readers' judgments were highly intercorrelated: syntheses perceived as superior on one dimension were also perceived as superior on other dimensions. Further, the best predictor of the readers' quality judgments was their confidence in their ability to interpret the results. Syntheses seen as more interpretable were given higher quality ratings. Readers were also asked to provide open-ended comments on papers, which were subjected to an informal content analysis. The seven dimensions that readers mentioned most frequently were organization, writing style, clarity of focus, use of citations, attention to variable definitions, attention to study methodology, and manuscript preparation.

More specifically, today numerous checklists and other forms of guidance are available for assessing the quality of research syntheses and meta-analyses. Among the most frequently used is Assessing the Methodological Quality of Systematic Reviews (AMSTAR 2016). Cooper also developed an evaluative checklist for consumers of syntheses of social science research (2007, 2017). Table 23.1 provides the twenty questions Cooper offers as most essential to evaluating social science research syntheses. The questions are written from the point of view of a synthesis consumer and each question is phrased so that an affirmative response means confidence could be placed in that aspect of the synthesis' methodology. The list is not exhaustive, but most of the critical issues discussed throughout this text find expression in the questions, as do the dimensions used in medical and health checklists that seem most essential to work in the social sciences.

It is important to make several additional points about the checklist. First, it does not use a scaling procedure that could calculate a numerical score for a synthesis, such that higher scores might indicate more trustworthy synthesis. This approach was rejected because of concerns raised about similar scales used to evaluate the "quality" of primary research (see chapter 7, this volume; Valentine and Cooper 2008), in particular, that single "quality" scores are known to generate wildly different conclusions depending on what specific dimensions of quality are included and on how the different items on the scale are weighted. Also, when summary numbers are generated, studies with very different profiles of strengths and limitations can receive similar scores, making syntheses with very different validity characteristics appear more similar than they actually are.

Second, some questions on the checklist lead to clearer prescriptions than do others of what constitutes *good practice* by synthesists. This occurs for two reasons. First, the definition of good practice will depend at least somewhat on the topic under consideration. For example, the identification of information channels to search in order to locate studies that are relevant to a synthesis topic and what terms to use when searching reference databases are clearly topic-dependent decisions. So, the checklist can only suggest that syntheses based on complementary sources and proper and exhaustive database search terms should be considered by consumers as more trustworthy; the checklist cannot specify what these sources and terms might be. As many of the chapters in this book suggest, a consensus has not yet emerged, even among expert synthesists, on some judgments about the adequacy of synthesis methods.

Finally, the checklist makes a distinction between questions that relate to the conduct of research synthesis in general and to meta-analysis in particular. This is because not all research areas will have the needed evidence base to conduct a meta-analysis that produces interpretable results. However, this does not mean that other aspects of sound synthesis methodology can be ignored (for example, clear definition of terms, appropriate literature searches).

In discussing several meta-analyses of the desegregation literature, Wachter and Straf point out that sophisticated literature searching procedures, data quality controls, and statistical techniques can bring a research synthesist only so far (1990). Eventually, they write, "there doesn't seem to be a big role in this kind of work for much intelligent statistics, as opposed to much wise thought" (182). These authors are correct in emphasizing the importance of wisdom in research integration. Wisdom is essential to any scientific enterprise, but wisdom starts with sound procedure. Synthesists must consider a wide range of technical issues as they piece together a research domain. This handbook is meant to help research synthesists build well-supported knowledge structures. But structural integrity is a minimum criterion for synthesis to lead to scientific progress. A more general kind of design wisdom is also needed to build knowledge structures that people want to view, visit, and, ultimately, live within.

Table 23.1 A Checklist of Questions for Evaluating Research Syntheses

Defining the problem

1. Are the variables of interest given clear conceptual definitions?
2. Do the operations that empirically define the variables of interest correspond to the variables' conceptual definitions?
3. Is the problem stated so that the research designs and evidence needed to address it can be specified clearly?
4. Is the problem placed in a meaningful theoretical, historical, or practical context?

Collecting the research evidence

5. Were complementary searching strategies used to find relevant studies?
6. Were proper and exhaustive terms used in searches and queries of reference databases and research registries?
7. Were procedures employed to assure the unbiased and reliable (a) application of criteria to determine the substantive relevance of studies, and (b) retrieval of information from study reports?

Evaluating the correspondence between the methods and implementation of individual studies and the desired inferences of the synthesis

8. Were studies categorized so that important distinctions could be made among them regarding their research design and implementation?
9. If studies were excluded from the synthesis because of design and implementation considerations, where these considerations (a) explicitly and operationally defined, and (b) consistently applied to all studies?

Summarizing and integrating the evidence from individual studies

10. Was an appropriate method used to combine and compare results across studies?
11. If effect sizes were calculated, was an appropriate effect size metric used?
12. If a meta-analysis was performed (a) were average effect sizes and confidence intervals reported, and (b) was an appropriate model used to estimate the independent effects and the error in effect sizes?
13. If a meta-analysis was performed, was the homogeneity of effect sizes tested?
14. Were (a) study design and implementation features (as suggested by question 8) along with (b) other critical features of studies, including historical, theoretical and practical variables (as suggested by question 4) tested as potential moderators of study outcomes?

Interpreting the cumulative evidence

15. Were analyses carried out that tested whether results were sensitive to statistical assumptions and, if so, where these analyses used to help interpret the evidence?
16. Did the research synthesists (a) discuss the extent of missing data in the evidence base and (b) examine its potential impact on the synthesis's findings?
17. Did the research synthesists discuss the generality and limitations of the synthesis findings?
18. Did the synthesists make the appropriate distinction between study-generated and review-generated evidence when interpreting the synthesis's results?
19. If a meta-analysis was performed, did the synthesists (a) contrast the magnitude of effects with other related effect sizes or (b) present a practical interpretation of the significance of the effects?

Presenting the research synthesis methods and results

20. Were the procedures and results of the research synthesis clearly and completely documented?

SOURCE: Author's tabulation.

23.6 REFERENCES

American Educational Research Association. 2006. "Standard for Reporting Empirical on Social Science Research in AERA Publications." *Educational Researcher* 35(6): 33–40. Accessed December 18, 2018. https://journals.sagepub.com/doi/abs/10.3102/0013189x035006033?journalCode=edra.

AMSTAR. 2016. "AMSTAR Checklist." Accessed December 18, 2018. http://amstar.ca/Amstar_Checklist.php.

Appelbaum, Mark, Harris Cooper, Rex B. Kline, Arthur M. Nezu, Stephen M. Rao, and Evan Mayo-Wilson. 2018. "Journal Article Reporting Standards for Quantitative Research: The APA Publications and Communications Board Task Force Report." *American Psychologist* 73(1): 3–25.

Campbell, Donald. 1986. "Relabeling Internal and External Validity for Applied Social Scientist." In *Advances in Quasi-Experimental Design and Analysis*, edited by William M. K. Trochim. San Francisco: Jossey-Bass.

Cooper, Harris M. 1986. "On the Social Psychology If Using Integrative Research Reviews: The Case of Desegregation and Black Achievement." In *The Social Psychology of Education*, edited by Robert Feldman. Cambridge: Cambridge University Press.

———. 2007. *Evaluating and Interpreting Research Syntheses in Adult Learning and Literacy*. Cambridge, Mass.: National Center for the Study of Adult Learning and Literacy.

———. 2017. *Research Synthesis and Meta-Analysis: A Step-by-Step Approach*, 5th ed. Thousand Oaks, Calif.: Sage Publications.

Cozzens, Susan E. 1981. *Users Requirements for Scientific Reviews.* Grant report no. IST-7821947. Washington, D.C.: National Science Foundation.

Glass, Gene V. 1976. "Primary, Secondary, and Meta-Analysis of Research." *Educational Researcher* 5(1): 3–8.

Hedges, Larry V., and Jacob M. Schauer. 2018. "Statistical Analyses for Studying Replication: Meta-Analytic Perspectives." *Psychological Methods*. Published online August 2, 2018. DOI:10.1037/met0000189.

Klein, Richard, Kate Ratliff, Michelangelo Vianello, Reginald Adams Jr., Stepán Bahník, Michael Bernstein, Konrad Bocian, et al. 2014. "Data from Investigating Variation in Replicability: A 'Many Labs' Replication Project." *Journal of Open Psychology Data* 2(1). Published online April 4, 2014. DOI: 10.5334/jopd.ad.

Mavridis, Dimitris, Myrsini Giannatsi, Andrea Cipriani, and Georgia Salanti. 2015. "A Primer on Network Meta-Analysis with Emphasis on Mental Health." *Evidence-Based Mental Health* 18(2): 40–46.

Marshall, Iain J., Anna Noel-Storr, Joël Kuiper, James Thomas, and Byron C. Wallace. 2018. "Machine Learning for Identifying Randomized Controlled Trials: An Evaluation and Practitioner's Guide." *Research Synthesis Methods*. DOI:10.1002/jrsm.1287.

National Science Foundation. 2016. "Dissemination and Sharing of Research Results." Accessed December 18, 2018. http://www.nsf.gov/bfa/dias/policy/dmp.jsp.

O'Muircheartaigh, Colm, and Larry V. Hedges. 2014. "Generalizing from Unrepresentative Experiments: A Stratified Propensity Score Approach." *Journal of the Royal Statistical Society Series C* 63(2): 195–210.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349(6251): aac4716.

Strike, Kenneth, and George Posner. 1983. "Types of Syntheses and Their Criteria." In *Knowledge Structure and Use: Implications for Synthesis and Interpretation*, edited by Spencer Ward and Linda J. Reed. Philadelphia, Pa.: Temple University Press.

Stuart, Elizabeth A, Catherine Bradshaw, and Philip Leaf. 2015. "Assessing the Generalizability of Randomized Trial Results to Target Populations." *Prevention Science* 16(3): 475–85.

Valentine, Jeffrey C., and Harris M. Cooper. 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(2): 130–49.

Wachter, Kenneth W., and Miron L. Straf, eds. 1990. *The Future of Meta-Analysis.* New York: Russell Sage Foundation.

GLOSSARY

A Posteriori Tests: Tests that are conducted after examination of study results. Used to explore patterns that seem to be emerging in the data. The same as *Post Hoc tests.*

A Priori Tests: Tests that are planned before the examination of the results of the studies under analysis. The same as *Planned tests.*

Aggregate Analysis: The integration of evidence across studies when a description of the quantitative frequency or level of an event is the focus of a research synthesis.

Agreement Rate: The most widely used index of interrater reliability in research synthesis. Defined as the number of observations agreed on divided by the total number of observations. Also called "percentage agreement."

Apples and Oranges Problem: A metaphor for studies that appear related, but are actually measuring different things. A label sometimes given as a criticism of meta-analysis because meta-analysis combines studies that may have differing methods and operational definitions of the variables involved in the calculation of effect sizes.

Artifact Correction: The modification of an estimate of effect size (usually by an artifact multiplier) to correct for the effects of an artifact. See also *Artifact Multiplier.*

Artifact Distribution: A distribution of values for a particular artifact multiplier derived from a particular research literature.

Artifact Multiplier: The factor by which a statistical or measurement artifact changes the expected observed value of a statistic.

Artifacts: Statistical and measurement imperfections that cause observed statistics to depart from the population (parameter) values the researcher intends to estimate.

Artificial Dichotomization: The arbitrary division of scores on a measure of a continuous variable into two categories.

Attenuation: The reduction or downward bias in the observed magnitude of an effect size produced by methodological limitations in a study such as measurement error or range restriction.

Available Case Analysis: Also called pairwise deletion, a method for missing data analysis that uses all available data to estimate parameters in a distribution so that, for example, all available pairs of values for two variables are used to estimate a correlation.

Bayes Posterior Coverage: A Bayesian analogue to the confidence interval.

Bayes's Theorem: A method of incorporating data into a prior distribution to obtain a posterior distribution.

Bayesian Analysis: An approach that incorporates a prior distribution to express uncertainty about population parameter values.

Best Evidence Synthesis: Research syntheses that rely on study results that are the best available, not an a priori or idealized standard of evidence.

Between-Studies Sample Size: The number of studies in a meta-analysis.

Between-Study Moderators: Third variables or sets of variables that affect the direction or magnitude of relations between other variables. Identified on a between-studies basis when some reviewed studies represent one level of the moderator and other studies represent other levels. See *Within-Study Moderators*.

Between-Study Predictors: Measured characteristics of studies hypothesized to affect true effect sizes.

Bibliographic Database: A collection of records describing publications. Typically each record provides the title, author, source information, and date. Many databases also include an abstract. Some databases may add subject index terms and other codes to facilitate retrieval. Also known as a reference database.

Bibliographic reference management software: Software to manage records that have been downloaded from bibliographic databases. Examples include EndNote, Reference Manager, RefWorks, and Mendeley.

Bibliographic Search: An exploration of literature to find reports relevant to the research topic. A search is typically conducted by consulting sources such as paper indexes, reference lists of relevant documents, contents of relevant journals/books, and electronic bibliographic databases.

Binomial Effect Size Display: A 2×2 table created by considering two groups of scores as a single distribution. The binomial effect size display describes the proportion of each group that falls above and below the median of that distribution.

Birge Ratio: The ratio of certain chi-square statistics (used in tests for heterogeneity of effects, model fit, or that variance components are zero) to their degrees of freedom. Used as an index of heterogeneity or lack of fit. Has an expected value of one under correct model specification or when the between-studies variance component is zero.

Buck's method: A method for missing data imputation that replaces missing observations with the predicted value from a regression of the missing variable on the completely observed variable; first suggested by Buck (1960).

Categorical (Grouping) Variable: A variable that can take on a finite number of values used to define groups of studies.

Certainty: The confidence with which the scientific community accepts research conclusions, reflecting the truth value accorded to conclusions.

Citation Search: A literature search in which documents are identified based on their being cited by other documents.

Coder: A person who reads and extracts information from research reports.

Coder Reliability: The equivalence with which coders extract information from research reports.

Coder Training: A process for describing and practicing the coding of studies. Modification of coding forms, coding conventions, and code book may occur during the training process.

Coding Conventions: Specific methods used to transform information in research reports into numerical form. A primary criterion for choosing such conventions is to retain as much of the original information as possible.

Coding Forms: The physical records of coded items extracted from a research report. The forms, and items on them, are organized to assist coders and data entry personnel.

Combining Results: Putting effect sizes on a common metric (for example, r or d) an calculating measures of location (for example, mean, median) or combining tests of significance.

Comparisons: Linear combinations of means (of effect sizes), often used in the investigation of patterns of differences among three or more means (of effect sizes). The same as *Contrasts.*

Complete Case Analysis: A method for analysis of data with missing observations, which uses only cases that observe all variables in the model under investigation.

Concept: A topic or theme that forms part of a research question, for example, a population of interest such as children with attention deficit disorder.

Conceptual (or Theoretical) Definition: A description of the qualities of the variable that are independent of time and space but that can be used to distinguish events that are and are not relevant to the concept.

Conditional Distribution: The sampling distribution of a statistic (such as an effect size estimate) when a parameter of interest (such as the effect size parameter) is held constant.

Conditional Exchangeability of Study Effect Size: A property of a set of studies that applies when the investigator has no a priori reason to expect the true effect size of any study to exceed the true effect size of any other study in the set, given the two studies share certain identifiable characteristics.

Conditional Mean Imputation: Another term for Buck's method or regression imputation, a method for missing data imputation that replaces missing observations with the predicted value from a regression of the missing variable on the completely observed variable; first suggested by Buck (1960).

Conditional Variance: Variability due to sampling error that associated with estimation of an effect size. See also *Within-Study Variance of an Effect Size*.

Confidence Interval (CI): The interval within which a population parameter is expected to lie.

Confidence Ratings: A method for coders to directly rate their confidence level in the accuracy, completeness, and so on, of the data being extracted from primary studies. Can establish a mechanism for discerning high-quality from lesser-quality information in the conduct of subsequent analyses.

Construct Validity: The extent to which generalizations can be made about higher-order constructs on the basis of research operations.

Controlled Vocabulary: The standard terminology defined by an indexing service for a reference database and made available through a thesaurus.

Correlation Coefficient: An index that gives the magnitude of linear relationship between two quantitative variables. It can range from –1 (perfect negative relationship) to +1 (perfect positive relationship), with 0 indicating no linear relationship.

Correlation Matrix: A matrix (rectangular listing arrangement) of the zero-order correlations among a set of p variables. The diagonal elements of the correlation matrix equal one because they represent the correlation of each variable with itself. Within a single study the off-diagonal elements are denoted r_{ij}, for $i, j = 1$ ti p, $i \neq$ j. The correlation r_{ij} is the correlation between variables i and j.

Covariates: Covariates are variables that are predictive of the outcome measure and are frequently incorporated into the analysis of categorical data. In randomized intervention studies, they are sometimes adjusted for in order to improve the precision with which key parameters are estimated and to increase the power of significance tests. In nonrandomized studies, they are adjusted for in order to eliminate the bias caused by confounding, that is, the presence of characteristics associated with both the exposure variable and the outcome variable.

Credibility Interval: Interval within which the true effect sizes vary across primary studies examining ostensibly the same relationship. It can also be interpreted as a possible range of true effect sizes.

Cumulative Distribution Function: The probability that a random continuous or discrete variable will have a value that is less than or equal to the function's argument.

Data Coding: The process of obtaining data. In meta-analysis it consists of reading studies and recording relevant information on data collection forms.

Data Coding Protocol: A manual of instruction for data coders that explains how to code data items, how to handle exceptions, and when to consult the project manager to resolve unexpected situations.

Data Entry: The process of transferring information from data collection forms into a database, requiring clerical skills.

Data Reduction: The process of combining data items in a database to produce a cases-by-variables file for statistical analysis.

Data Sharing: Allowing researchers not involved in the original collection of data access to it for purposes of reanalysis and /or integration with other data sets.

Density Function: A mathematical function that defines the probability of occurrence of values of a random variable.

Descriptive Analysis: Descriptive statistics that characterize the results and attributes of a collection of research studies in an synthesis.

Descriptors: In literature searching, refers to subject headings used to identify important concepts included in a document. In data coding, refers to variables that identify characteristics of studies included in a research synthesis.

Disattenuation: The process of correcting for the reduction or downward bias in an effect size produced by attenuation.

Double Coding: A method of having two or more independent coders code studies to enhance and assess reliability of the coding process.

Effect Size: A generic term that refers to the magnitude of an effect or more generally to the size of the relation between two variables. Special cases include standardized mean difference, correlation coefficient, odds ratio, and the raw mean difference.

Effect Size Items (Coding): Items related to an effect size. Can include the nature and score reliability of outcome and predictor measures, sample size(s), means and standard deviations, and indices of association.

Eligibility Criteria: Conditions that must be met by a primary study in order for it to be included in the research synthesis. Also called "inclusion criteria."

EM Algorithm: An iterative estimation procedure that cycles between estimation of the posterior expectation (E) of a random variable (or collection of variables) and the maximization (M) of its posterior likelihood. Useful for maximum likelihood estimation with missing data.

Error of Measurement: Random errors in the scores produced by a measuring instrument.

Estimation Variance, v_i: The variance of the effect size estimator given a fixed true effect size. The same as *Conditional Variance.*

Evidence-Based Practices: Programs, products, and policies that are believed to be effective based on high-quality research designs and methods.

Exclusion Criteria: See *Ineligibility Criteria*.

Experiment: An investigation of the effects of manipulating a variable.

Experimental Research: Research in which both the introduction of the event and who is exposed to it are controlled by the researcher. The researcher uses a random procedure to assign students to conditions, essentially leaving the assignment to chance.

Experimental Units: The smallest division of the experimental (or observational) material such that any two units may receive different treatments.

Explicit Truncation: A process by which all scores in a distribution in a certain range (for example, all scores below $z = -0.10$) are eliminated from the data used in the data analysis. See *Range Restriction.*

Exploding Subject Headings: Offered in many databases, "exploding" subject index terms will automatically include any more specific terms in the search. In the Ovid interface in PsycINFO, exploding "short term memory" will also retrieve records for "iconic memory."

Exploratory Data Analysis: A descriptive (as opposed to inferential) analysis of data that often supplements numerical summaries with visual displays.

External Validity: The value of a study or set of studies for generalizing to individuals, settings, or procedures. Studies with high external validity can be generalized to a larger number of individuals, settings, or procedures than can studies with lower external validity. Largely determined by the sampling of individuals, settings, or procedures employed in the investigations from which generalizations are to be drawn.

Extrinsic Factors: The characteristics of research other than the phenomenon under investigation or the methods used to study that phenomenon, for example, the date of publication and the gender of the author.

Falsificationist Approach: A framework that stresses how secure knowledge depends on identifying and ruling out plausible alternative interpretation.

File-Drawer Problem: The situation in which study results go unreported (and thus are left in file drawers) when tests of hypotheses do not show statistically significant results.

Fixed Effects: Effects (effect sizes or components of a model for effect size parameters) that are unknown constants. Compare with *Random Effects*.

Fixed Effects Model: A model for combining effect sizes that assumes all effect sizes are fixed effects (that is constants as opposed to random quantities). If all the effect sizes estimate a common population parameter, so that the observed effect sizes differ from that parameter only by virtue of sampling error, this is sometimes called the fixed effect (singular) model or the common effect model.

Flat-File: In data management, a file that constitutes a collection of records of the same type that do not contain repeating items. Can be represented by two-dimensional array of data items, that is, a case-by variables data matrix.

Footnote Chasing: A technique for discovering relevant documents by tracing authors' footnotes (more broadly, their references) to earlier documents. Also known as the "ancestry approach." See also *Forward Citation Searching.*

Forward Citation Searching: A way of discovering relevant documents by looking up a known document in a citation index and finding the later documents that have cited it. See also *Footnote Chasing.*

Fourfold Table: A table that reports data concerning the relationship between two dichotomous factors. See *Odds Ratio, Log Odds Ratio,* and *Rate Difference*.

Free-Text Terms: Words in an information source record other than the indexing terms. Free text terms are usually those in the title and abstract of a database record.

Fugitive Literature: Papers or articles produced in small quantities, not widely distributed, and usually not listed in commonly used abstracts and indexes. See also *Grey Literature.*

Full-Text Database: A machine-readable file that consists of the complete texts of documents as opposed to merely citations and abstracts.

Funnel Plot: A graphic display of sample size plotted against effect size. When many studies come from the same population, each estimating the same underlying parameter, the graph should resemble a funnel with a single spout.

Generalization: Important purpose and promise of a meta-analysis. Refers to empirical knowledge about a general association. It can involve identifying whether an association (1) holds with specific populations of persons, settings, times, and ways of varying the cause or measuring the effect; (2) holds across different populations of people, settings, times, and ways of operationalizing a cause and effect; and (3) can even be extrapolated to other populations of people, settings, times, causes, and effects than those that have been studied to date.

Generalized Least Squares (GLS) Estimation: An estimation procedure similar to the more familiar ordinary least squares method. A sum of squared deviations between parameters and data is minimized, but GLS allows the data points for which parameters are being estimated to have unequal population variances and nonzero covariances (that is, to be dependent).

Grey Literature: Literature that is produced on all levels of government, academia, business, and industry in print and electronic formats, but that is not controlled by commercial publishers. See also *Fugitive Literature.*

Hand Searching: Searching the contents of a journal by looking at each article in sequence and making an assessment of the relevance of the article to the synthesis question. Handsearching may also be undertaken for sections of databases or websites.

Heterogeneity: The extent to which observed effect sizes differ from one another. In meta-analysis, statistical tests allow for the assessment of whether the variability in observed effect sizes is greater than would be expected

given chance (that is, sampling error alone). If so, then the observed effects are said to be heterogeneous.

Heterogeneity/Homogeneity of Classes: The range/similarity of classes of persons, treatments, outcomes, settings, and times included in the studies that have been reviewed.

Hierarchical Data Structure: In data management, a nested set of separate data files that each contain different coded information. This data structure allows for hierarchical data structures, such as multiple effect sizes for individual studies. Each data file can have a different number of rows per individual study.

High Inference Codes: Codes that involve attempting to infer how a contextual aspect of the studies might have been interpreted by participants.

Homogeneity: A condition under which the variability in observed effect sizes is not greater than would be expected given sampling error.

Homogeneity Test: A test that a collection of effect size estimates exhibit greater variability than would be expected if their corresponding effect size parameters were identical.

Hypothesis: A research problem containing a prediction about a particular link between the variables—based on theory or previous observation.

Identification Items (Coding): The category of coded items that document the research reports that are synthesized. Author, coder, country, and year and source of publication are typical items.

Ignorable Response Mechanism: When data are either missing completely at random (MCAR) or are missing at random (MAR), the mechanism that causes missing data is called ignorable because the response mechanism does not have to be modeled when using maximum likelihood or multiple imputation for missing data.

Indexing: The addition of indexing terms to database records, to provide standardized search terms for concepts that might be expressed by authors in various ways.

Indexing Language: A controlled vocabulary used to index records in a database to enhance consistent retrieval of records.

Indexing Term: A word or phrase from an indexing language.

Inclusion Criteria: See *Eligibility Criteria.*

Indirect Relation: A relationship between two variables where a third (or more) variable intervenes. A variable can have both direct and indirect relations to an outcome; its indirect relations are achieved by way of one more intermediate (mediator) variables. See *Mediator Variable.*

Ineligibility Criteria: Conditions or characteristics that render a primary study ineligible for inclusion in the research synthesis. The same as *Exclusion Criteria.*

Information Source: A database, website, or library that provides access to research evidence and other documents.

Information Specialist: An information scientist or librarian who has extensive experience of searching for research evidence from a variety of information sources.

Intercoder Correlation: An index of interrater reliability for continuous variables, based on the Pearson correlation coefficient. Used in research synthesis to estimate interrater reliability, analogous to its use in education to estimate test reliabilities when parallel forms are available.

Interface: This is a set of options or facilities that are available for searching a database. Options may include ways to combine sets of search results, including Boolean operators and proximity operators, ways to search for word variants (truncation, stemming, and wildcards) and the ability to restrict searches to specific fields such as the title.

Interjudge Reliability: The degree of agreement between judges or observers who are rating or observing the same events. Agreement can be expressed in numerous different ways (for example, percentage of cases for which ratings agreed, kappa).

Internal Validity: The value of a study or set of studies for concluding that a causal relationship exists between variables, that is, that one variable affects another. Studies with high internal validity provide a strong basis for making a causal inference. Largely determined by the control of alternative variables that could explain the relationship found within a study.

Inter-rater Reliability: The extent to which different raters rating the same studies assign the same rating to the coded variables.

Intraclass Correlation: Computed as the ratio of the variance of interest over the sum of the variance of interest plus error, the intraclass correlation is a family of indicies of agreement (or consistency) for continuous variables. The intraclass correlation can be used to isolate different sources of variation in measurement and to estimate their magnitude using the analysis of variance. The intraclass correlation is also used to describe the amount of clustering in a population in clustered or multilevel samples where it is the ratio of between-cluster variance to total variance.

Invisible College: A geographically dispersed network of scientists or scholars who share information in a particular research area through personal communication.

Kappa: A versatile family of indices of interrater reliability for categorical data, defined as the proportion of the best

possible improvement over chance that is actually obtained by the raters. Generally superior to other indices designed to remove the effects of chance agreement, kappa is a "true" reliability statistic that in large samples is equivalent to the intraclass correlation coefficient.

Knapp-Hartung Adjustment: An empirical adjustment to the variance or standard error of means or meta-regression coefficients in random effects meta-analyses. The Knapp-Hartung correction leads to larger but more accurate variance estimates.

Large-Sample Approximation: The statistical theory that is valid when each study has a "large" (within-study) sample size. The exact number of cases required to quality as "large" depends on the effect size index used. In meta-analysis, most statistical theory is based on large-sample approximations.

Listwise Deletion: A method for analyzing data with missing observations by analyzing only those cases with complete data; also termed complete case analysis.

Log Odds: See *Logit*.

Log Odds Ratio: The sample statistic, l, population parameter γ. Computed as the natural logarithm of the odds ratio $\ln(o)$. See *Fourfold Table.*

Logistic Regression: A statistical model in which the logit of a probability is postulated to be a linear function of a set of independent variables.

Logit: The logarithm of the odds value associated with a probability. If Π is a probability, the logit is $\ln \Pi/(1 - \Pi)$, where ln denotes natural logarithm.

Logit Transformation: A transformation commonly used with proportions or other numbers that have values between zero and one. The logit is the logarithmic transformation of the ratio $[p/(1 - p)]$, that is, $\ln [p/(1 - p)]$.

Low Inference Codes: Codes of study characteristics that require the synthesists only to locate the needed information in the research report and transfer it to the synthesis database.

Mantel-Haenszel Statistics: A set of statistics for combining odds ratios and testing the statistical significance of the resulting average odds ratio.

Maximum Likelihood: A commonly used method for obtaining an estimate for an unknown parameter from a population distribution.

Maximum Likelihood Methods for Missing Data: Methods for analysis of data with missing observations that provide maximum likelihood estimates of model parameters.

Mean of Random Effects: The mean of the true effect sizes in the random effects model.

Measurement Error: Random departure of an observed score of an individual from his/her actual (true) score. Sources of measurement error include random response error, transient error, and specific factor error.

Mediating Variable: Third variables or sets of variables that provide a causal account of the mechanisms underlying the relationship between other variables. Transmits a cause-effect relationship in the sense that it is a consequence of a more distal cause and a cause of more distal measured effects. A fundamental property of a mediator is that it is related to both a precursor and an outcome, while the precursor may show only an indirect relation to the outcome via the mediator.

Meta-Analysis: The statistical analysis of a collection of analysis results from individual studies for the purpose of integrating the findings.

Meta-Regression: The use of regression models to assess the influence of variation in contextual, methodological, participant, and program attributes on effect size estimates.

Method of Maximum Likelihood: A method of statistical estimation in which one chooses as the point estimates of a set of parameters values that maximize the likelihood of observing the data in the sample.

Method of Moments: A method of statistical estimation in which the sample moments are equated to their expectations and the resulting set of equations solved for the parameters of interest.

Missing at Random: Observations that are missing for reasons related to completely observed variables that are included in the model and not to the value of the missing observation.

Missing Completely at Random: Observations that are missing for reasons unrelated to any variables in the data. Missing observations occur as if they were deleted at random.

Missing Data: Data representing either study outcomes or study characteristics that are unavailable to the synthesist.

Model-Based Meta-Analysis: A meta-analysis that analyzes complex chains of events such as the prediction of behaviors based on a set of variables, and models the intercorrelations among variables.

Moderator Variable: A variable or set of variables that affect the direction or magnitude of relations between other variables. Variables such as gender, type of outcome, and other study features are often identified as moderator variables.

Multinomial Distribution: A probability distribution associated with the independent classification of a sample of subjects into a series of mutually exclusive and exhaustive categories. When the number of categories is two, the distribution is a binomial distribution.

Multiple Imputation: A method for data analysis with missing observations that generates several possible substitutes for each missing data point based on models for the reasons for the missing data.

Nonignorable Response Mechanism: A term that describes observations that are missing for reasons related to unobserved variables. Occurs when observations are missing because of their unknown values or because of other unknown variables.

Nonparametric Statistical Procedures: Statistical procedures (tests or estimators) whose distributions do not require knowledge of the functional form of the probability distribution of the data. Also called "distribution free statistics."

Normal Deviate: The location on a standard normal curve given in Z-score form.

Normal Distribution: A bell-shaped curve that is completely described by its mean and standard deviation.

Not Missing at Random (NMAR): A type of missing data where the probability of observing a value depends on the value itself; occurs in censoring mechanisms when, for example, high values on a variable are more likely to be missing than moderate or low values.

Odds Ratio: The *odds* for the outcome event of interest being positive, conditional on the exposure (intervention) being positive, are equal to the probability of the outcome being observed (in the intervention group), divided by one minus the probability of the outcome being observed in the intervention group. The *odds* for the outcome event being positive, conditional on the exposure (intervention) being absent, can be defined similarly. The *odds ratio* is simply the ratio of these two odds values. The underlying assumption is that fixed numbers of exposed and of unexposed individuals are sampled, and the probability of the outcome is observed. A unique property of the odds ratio is that it can be calculated from a study in which pre-specified numbers of units positive on the outcome and negative on the outcome are selected for a determination of their status on the exposure variable. Because the two study designs correspond to prospective and retrospective sampling, it is clear that the odds ratio is estimable using data from either of these two designs (as well as from a cross-sectional study).

Odds Value: A probability divided by its complement. If Π is a probability, its associated odds value is $\Pi(1-\Pi)$.

Omnibus Tests of Significance: Significance tests that address unfocused questions, as in F tests with more than 1 degree of freedom (df) for the numerator or in chi-square tests with more than 1 df.

Operational Definition: A definition of a concept that relates the concept to observable events.

Operational Effect Size: The effect size actually estimated in the study as conducted; for example, the (attenuated) effect size computed using an unreliable outcome measure. Compare with *Theoretical Effect Size*.

***p*-Value:** The probability associated with a statistical hypothesis test. The *p* value is the probability of obtaining a sample result at least as large as the one obtained given a true null hypothesis (that is, given the hypothesis that sampling error is the only reason that the observed result deviates from the hypothesized parameter value).

Pairwise Deletion: A method for missing data analysis that uses all available data to estimate parameters in a distribution so that, for example, all available pairs of values for two variables are used to estimate a correlation; also called available case analysis.

Parameter: The population value of a statistic.

Partial Effect Sizes: Partial effect sizes are effect sizes that represent the association between two variables adjusting for the effect of one or more variables on the focal predictor and outcome, depending on which index is used.

Partial Relations: Relationships between two variables where a third (or more) variable has been statistically controlled, such as in a multiple regression equation.

Partial Truncation: A process by which the frequencies of scores in certain parts of the range are reduced without completely eliminating scores in these ranges. For example, if one limits the data to high school graduates and above, the frequency of people with IQs below 85 will be reduced relative to the entire population. See *Range Restriction.*

Pearson Correlation: See *Correlation Coefficient*.

Peer Review: The review of research by an investigator's peers, usually conducted at the request of a journal editor prior to publication of the research and for purposes of evaluation to ensure that the research meets generally accepted standards of the research community.

Phi Coefficient (φ): A measure of association between two binary variables that cross-classified against each other. The phi coefficient is calculated as the product moment correlation coefficient between a pair of binary random variables.

Placebo: A pseudo-treatment administered to subjects in a control group.

Planned Tests: Tests that are planned before the examination of the results of the data under analysis. The same as *A Priori Tests*.

Pooling Results: Using data or results from several primary studies to compute a combined significance test or estimate.

Post Hoc Matching: A procedure whereby participants in different conditions of a study are identified as having the

same or similar scores on critical baseline variables. Participants without matches are removed from the pertinent analysis.

Post Hoc Tests: Tests that are conducted after examination of the data to explore patterns that seem to be emerging in the data. The same as *A Posteriori Tests*.

Posterior Distribution: An updated prior distribution that incorporates observed data (also posterior mean, posterior variance, and so on)

Power: See *Statistical Power.*

Power Analysis: See *Statistical Power Analysis.*

Precision: A measure used in evaluating literature searches. The ratio of documents retrieved and deemed relevant to total documents retrieved. Also known as specificity.

Precision of Estimation: Computationally, the inverse of the variance of an estimator. Conceptually, larger samples estimate population parameters more precisely than smaller samples. Results from larger samples will have smaller confidence intervals compared to results from smaller samples, all else being equal.

Pretest-Posttest Design: A study in which participants are tested on the outcome variable, exposed to an intervention, and tested again on the outcome variable. Often this design leaves so many validity threats plausible that it is difficult to interpret the results.

Primary Study: A report of original research, usually published in a technical journal or appearing as a thesis or dissertation; refers to the original research reports collected for a research synthesis.

Prior Distribution: An expression of the uncertainty of parameter values as a statistical distribution before observing the data (also prior mean, prior variance, and so on)

Prior Odds: The relative prior likelihood of pairs of parameter values.

Progressive Paradigm Shift: A fundamental change in the framework of science that results in the explanation of more natural phenomena or the conduct of empirical studies that are more in line with the rules of science. The term paradigm shift in science is attributed to Thomas Kuhn.

Prospective Registration: The compilation and collation of intended research projects before the projects formally start. See *Retrospective Registration*.

Proximal Similarity: The similarity in manifest characteristics (for example, prototypical attributes) between samples of persons, treatments, outcomes, settings, and times and their corresponding universes.

Proximity Operators: Search operators that specify that a search term can be retrieved when it occurs within a certain distance from another search term. The distance can usually be varied.

Proxy Variables: Variables, such as year of study publication, that serve as surrogates for the true, causal variables, such as developments in research methods or changes in data-reporting practices.

Publication Bias: The tendency for studies with statistically significant results to have a greater chance of being published than studies with non-statistically significant results. Because of this, research syntheses that fail to include unpublished studies may overestimate the true effect of an intervention.

Quasi-Experiment: An experiment in which the method of allocation of observations to study conditions is not done by chance. See *Randomized Experiment.*

Random Assignment: The allocation of individuals, communities, groups, or other units of study to an experimental intervention, using a method that assures that assignment to any particular group is by chance.

Random Effects: Effects that are assumed to be sampled from a distribution of effects. Compare with *Fixed Effects*.

Random Effects in a Regression Model for Predicting Effect Size: The deviation of Study *i*'s true effect size from the value predicted on the basis of the regression model.

Random Effects Model: A model for combining effect sizes under which observed effect sizes may differ from each other both due to sampling error and due to true variability in population parameters.

Random Effects Variance Component: A variance of the true effect size viewed as either (a) the variance of true effect sizes in a population of studies from which the synthesized studies constitute a random sample or (b) the degree of the investigator's uncertainty about the proposition that the true effect size of a particular study is near a predicted value.

Random Sampling: The allocation of individuals, communities, groups, or other units in a population to a sample, using a method that assures that assignment to the sample is by chance.

Randomized Experiment: An experiment that employs random assignment of observations to study conditions. See *Quasi-Experiment*.

Range Restriction: A situation in a data set in which a pre-selection process causes certain scores or ranges of scores to be missing from the data. For example, all people with scores below the mean may have been excluded from the data. Range restriction is produced by both *Explicit Truncation* and *Partial Truncation*.

Rank Correlation Test: A nonparametric statistical test in which the correlation between two factors is assessed by comparing the ranking of the two samples.

Rapid Review Search: A search undertaken for a synthesis conducted under a very short timeframe. These searches

usually involve a few selected databases and therefore are less likely to minimize publication bias.

Rate Difference: See *Fourfold Table.*

Rate Ratio / Risk Ratio: The ratio of two probabilities, *RR*, is a popular measure of the effect of an intervention. The names given to this measure in the health sciences, the *risk* ratio or relative *risk*, reflect the fact that the measure is not symmetric, but is the ratio of the first group's probability of an undesirable outcome to the second's probability (recognizing that the ratio could also be expressed in terms of the favorable outcome, at least in theory). (In fact, in the health sciences, the term "rate" is often reserved for quantities involving the use of "person-time," that is, the product of the number of individuals and their length of follow-up.)

Raw Mean Difference: The difference between the means of the intervention and comparison groups.

Recall: A measure used in evaluating literature searches. The ratio of relevant documents retrieved to total documents deemed relevant in a collection (for example, the total number of relevant reports in existence). Also known as sensitivity.

Reduction of Random Effects Variance: The difference between the baseline and the residual random effects variance.

Reference Database: A repository of citations to documents, for example, PsycINFO. Also known as bibliographic database.

Regression Discontinuity Design: The allocation of individuals, communities, groups, or other units of study to intervention conditions, using a specific cutoff score on an assignment variable

Regression Imputation: Also called Buck's (1960) method or regression imputation. A method for missing data imputation that replaces missing observations with the predicted value from a regression of the missing variable on the completely observed variable; first suggested by Buck (1960).

Relative Risk: See *Rate Ratio.*

Relevance: The use of construct and external validity as entry criteria for selecting studies for a synthesis.

Relevant Document: A document that matches a researcher's broad inclusion criteria (for example, is topically relevant) and possibly meets additional criteria for inclusion.

Reliability: Generally, the repeatability of measurement. More specifically, the proportion of observed variance of scores on a measuring scale that is not due to the variance of random measurement errors; or the proportion of observed variance of scores that is due to variance of true scores. Higher values indicate less measurement error and vice versa.

Replication: The repetition of a previously conducted study. Replications are sometimes referred to as "direct" (or sometimes, "statistical") when the researchers attempt to employ the same methods that were used in the original study, and as "conceptual" when researchers use different operations to test the same conceptual variables and hypotheses. Direct replications are often conceived of as attempts to assess whether the replication will achieve results similar to what was reported in an original study, assuming that the study conditions are similar. Conceptual replications are often undertaken to test whether the replication will achieve results similar to what was reported in the original despite intentional variations in the operations used to test the hypotheses.

Research Problem: A statement of (a) what variables are to be related to one another and (b) an implication of how the relationship can be tested empirically.

Research Quality: Factors such as study design or implementation features that affect the validity of a study or set of studies.

Research Register: A database of research studies (planned, active, and/or completed), usually oriented around a common feature of the studies such as subject matter, funding source, or design.

Research Synthesis: A review of a clearly formulated question that uses systematic and explicit methods to identify, select, and critically appraise relevant research, and to collect and analyze data from the studies that are included in the synthesis. Statistical methods (meta-analysis) may or may not be used to analyze and summarize the results of the included studies. Systematic review is a synonym of research synthesis.

Residual Random Effects Variance: The residual variance of the true random effects after taking into account a set of between-studies predictors.

Response Mechanism: Term used to describe the hypothetical reasons for missing data

Retrieved Studies: Primary studies identified in the literature search phase of a research synthesis and whose research report is obtained in full by the research synthesis investigators.

Review Generated Evidence: In a meta-analysis, evidence that arises from studies that do not directly test the relation being considered. Compare with *Study Generated Evidence*.

Robust Methods: Techniques that are not overly sensitive to the choice of a prior or sampling distribution or to outliers in the data.

Robust Variance Estimate: A technique used to produce variances and standard errors of coefficients in meta-regression using empirical as opposed to purely theoretical methods. Robust variance estimates are particularly useful when there are multiple correlated effect sizes in some studies.

Robustness: Consistency in the direction and/or magnitude of effect size for studies involving different classes of persons, treatments, settings, outcomes, or times.

Samples: Constituents or members of a universe, class, population, category, type, or entity.

Sampling Error: The difference between an effect size statistic and the population parameter that it estimates that occurs due to employing a sample instead of the entire population.

Scoping Search: A search undertaken to understand the scale, scope, and coverage of the literature. A scoping search can be used to identify which databases should be included in a systematic search, to help synthesists estimate the total number of records that might be retrieved in a systematic search, and to identify key search terms. Synthesists undertaking scoping searches may or may not intend these searches to be exhaustive.

Search Filters: A collection of search terms that identifies records about a specific population, study design or other issue; ideally derived by research.

Search Strategy: A collection of search terms that is used to interrogate a database to identify records relevant to a research synthesis question.

Search Terms: Words used to identify the key aspects of a subject for use in searching for relevant documents.

Second-Order Meta-Analysis: A meta-analysis conducted on multiple statistically independent and methodologically comparable mean true effect sizes in relevant primary meta-analyses (that is, meta-analyses conducted on primary studies) examining ostensibly the same relationship. Second-order meta-analysis addresses the problem of second-order sampling error still remaining in input the primary meta-analyses. Also known as umbrella review, meta-meta-analysis, and meta-analysis of meta-analyses.

Second-Order Sampling Error: The sampling error of primary meta-analytic estimates of mean and standard deviation of true effect sizes. This error still remains in primary meta-analyses because the number of studies in each primary meta-analysis is not infinite.

Selection Bias: An error produced by systematic differences between individuals, entities, or studies selected for analysis and those not selected.

Sensitivity: A measure used in evaluating literature searches. The ratio of relevant documents retrieved to total documents deemed relevant in a collection. Also known as recall. See *Precision*.

Sensitivity Analysis: An analysis used to determine whether and how sensitive the conclusions of the analysis are violations of assumptions or decision rules used by the synthesist. For example, a sensitivity analysis may involve eliminating dependent effect sizes to see if the average effect size changes.

Significance Level: See *p-Value*.

Significance Test: See *Statistical Significance Test*.

Single-Case Research: Research that studies how individual units change over time, and what causes this change.

Single-Value Imputation: A method for missing data analysis that replaces missing observations with one value such as the complete case mean or the predicted value from a regression of variables with missing data on variables completely observed.

Standardized Mean Difference: An effect size that expresses the difference (in standard deviation units) between the means of two groups. The d-index is one standardized mean difference effect size.

Statistical Power: The probability that a statistical test will correctly reject a false null hypothesis.

Statistical Power Analysis: A statistical analysis that estimates the likelihood of obtaining a statistically significant hypothesis test, given assumptions about the sample size employed, the desired Type I error rate, and the true magnitude of the relationship.

Stem-and-Leaf Display: A histogram in which the data values themselves are used to characterize the distribution of a variable. Useful for describing the center, spread, and shape of a distribution.

Study Descriptors: Coded variables that describe the characteristics of research studies other than their results or outcomes. For example, the nature of the research design and procedures used, attributes of the subject sample, and features of the setting are all study descriptors.

Study Protocol: A written document that defines the concepts used in the research and that describes study procedures in as much detail as possible before the research begins.

Study Generated Evidence: In a meta-analysis, evidence that arises from studies that directly test the relation being considered. For example, several studies might compare the relative effects of attending summer school (compared to control students) on boys versus girls. A meta-analysis of these comparisons is considered study generated evidence. Compare with *Synthesis Generated Evidence*.

Subject Subsamples: Groupings of a subject sample by some characteristic—for example, gender or research site.

Substantive Factors: Characteristics of research studies that represent aspects of the phenomenon under investigation, for example, the nature of the treatment conditions applied and the personal characteristics of the subjects treated.

Substantively Irrelevant Study Characteristic: Characteristics of studies that are irrelevant to the major hypotheses under study, but are inevitably incorporated into research operations to make them feasible. The analyst's goal is to show that tests of the hypothesis are not confounded with such substantive irrelevancies.

Sufficient Statistics: Term used for the statistics needed to estimate a parameter; the sum of the values of a variable is the sufficient statistic for computing the mean of a distribution.

Synthesis-Generated Evidence: In a meta-analysis, evidence that arises from studies that do not directly test the relation being considered. For example, some studies might examine the effect of attending summer school for low-SES students, while others might examine this relation for middle-SES students. A meta-analysis of the differential effectiveness of summer school for low versus middle-SES students is considered synthesis-generated evidence because the studies do not directly test this relation. Compare with *Study Generated Evidence*.

Synthetic Correlation Matrix: A matrix of correlations, each synthesized from one or more studies.

Systematic Artifacts: Statistical and measurement imperfections that bias observed statistics (statistical estimates) in a systematic direction (either upward or downward). See *Artifacts*.

Test of Significance: See *Statistical Significance Test.*

Text Mining: The use of software to analyze unstructured text and identify patterns and derive information about a body of literature. In the searching context, text mining can be used to identify terms, phrases, and collocated terms that might be used in search strategies. Text mining also has applications in record selection.

Theoretical Effect Size: The effect size that would be obtained in a study that differs in some theoretically defined way from one that is actually conducted. Typically, theoretical effect sizes are estimated from operational effect sizes with the aid of auxiliary information about study designs or measurement methods. For example, the (disattenuated) correlation coefficient corrected for the effects of measurement unreliability is a theoretical effect size. Compare to *Operational Effect Size*.

Theory Testing: The empirical validation of hypothesized relationships, which may involve the validation of a single relationship between two variables or a network of relations between several variables. Often validation depends on observing the consequences implied by, but far removed from, the basic elements of the theory.

Thesaurus: A list of subject terms in the controlled vocabulary authorized by subject experts to index the subject content of documents in a database.

Threat to Validity: A likely source of bias, alternative explanation, or "plausible rival hypothesis" for the research findings.

Time Series Design: Research design in which units are tested at different times, typically equal intervals, during the course of a study.

Total Variance of an Effect Size Estimate: The sum of the random effects variance and the estimation (or fixed effects) variance. The same as *Unconditional Variance*.

Transformation: The application of some arithmetic principle to a set of observations to convert the scale of measurement to one with more desirable characteristics, for example, normality, homogeneity, or linearity.

Treatment-by-Studies Interaction: The varying effect of a treatment across different studies; equivalently, the effect of moderators on the true effect size in a set of studies of treatment efficacy.

Trial Register: A database of research studies, planned, active or completed, usually oriented around a common feature of the studies such as the subject matter, funding source, or design.

Trim and Fill Method: A funnel-plot-based method that attempts to test and adjust for publication bias in meta-analysis.

True Effect Size: The effect magnitude that an investigator seeks to estimate when planning a study.

True Score: Conceptually, the score an individual would obtain in the complete absence of measurement error. A true score is defined as the average score that a particular individual would obtain if he/she could be independently measured with an infinite number of equivalent measures, with random measurement error being eliminated by the averaging process.

Truncation: Used to specify different word endings to a word root. The specific methods of truncation vary across databases. In the Ovid interface, "random$" will retrieve all terms with the stem "random" (examples include randomized, randomised, randomization, and randomness).

Unconditional Mean Imputation: Also called mean imputation, a method for missing data analysis that replaces missing observations with the complete case mean for that variable.

Unconditional Variance: An estimate of the total variability of a set of observed effect sizes, presumed to be due to both within-study and between-study variance.

Unit of Analysis: The study unit that contributes to the calculation of an effect size and its standard error within a study. This unit may be a single individual or a collection of individuals, like a classroom or dyad.

Universe: The population (or hyperpopulation) of studies to which one wishes to generalize.

Unpublished Literature: Papers or articles not made available to the public through the action of a publisher.

Unsystematic Artifacts: Statistical imperfections, such as sampling error, which cause statistical estimates to depart randomly from the population parameters that a researcher intends to estimate. See *Artifacts*.

Validity threat: See *Threat to Validity*.

Variance Components: The sources into which the variance of effect sizes is partitioned. Also, the random variables whose variances are the summands for the total variance.

Variance-Covariance Matrix: A matrix (rectangular listing arrangement) that contains the variances and covariances for a set of p variables. Within a single study the diagonal elements of the covariance matrix are the variances S_{ii}^2 for variables 1 to p. The covariances S_{ij}^2 for $i, j = 1$ to p, $i \neq j$, are the off-diagonal elements of the matrix.

Vote-Counting Procedure: A procedure in which one simply tabulates the number of studies with significant positive results, the number with significant negative results, and the number with nonsignificant results. The category with the most votes presumably provides the best guess about the direction of the population effect size. More modern vote-counting procedures provide an effect size estimate and confidence interval.

Weighted Least Squares Regression in Research Synthesis: A technique for estimating the parameters of a multiple regression model wherein each study's contribution to the sum of products of the measured variables is weighted by the precision of that study's effect size estimator.

Weighting: The process of allowing observations with certain desirable characteristics to contribute more to the estimation of a parameter. In meta-analysis, each study's contribution is often weighted by the precision of that study's effect size estimator.

Wildcards: In literature searching, wildcards are used to account for internal spelling variations. The specific methods for identifying wildcards vary across databases. In the Ovid interface, "randomi?ed" will identify records containing the word "randomized" and the word "randomised."

Within-Study Moderators: Moderators are identified on a within-study basis when two or more levels of the moderator are present in individual reviewed studies. See *Between-Studies Moderators*.

Within-Study Sample Size: The number of units within a primary study.

Z-Transformed Correlation Coefficient: The sample statistic z, population parameter ζ,. The Fisher Z transformation of a correlation r is $.5\ \ln[(1 + r)/(1 - r)]$.

INDEX

Boldface numbers refer to figures, tables and photos.