RESEARCH SYNTHESIS AND META-ANALYSIS

A Step-by-Step Approach

HARRIS COOPER   FIFTH EDITION

Research Synthesis and Meta-Analysis

Fifth Edition

To Elizabeth

APPLIED SOCIAL RESEARCH
METHODS SERIES

Research Synthesis and Meta-Analysis

A Step-by-Step Approach

Fifth Edition

APPLIED SOCIAL RESEARCH METHODS SERIES

Harris Coope
Duke University

\$SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC

Brief Contents

Detailed Contents

Preface to the Fifth Edition

Every scientific investigation begins with the researcher examining reports of previous studies related to the topic of interest. Without this step, researchers cannot expect their efforts to contribute to an integrated, comprehensive picture of the world. They cannot achieve the progress that comes from building on the efforts of others. Also, investigators working in isolation are doomed to repeat the mistakes made by their predecessors.

Similar to primary data collection, researchers need guidance about how to conduct a research synthesis—how to find research already conducted on a particular topic, gather information from research reports, evaluate the quality of research, integrate results, interpret the cumulative findings, and present a comprehensive and coherent report of the synthesis' findings. This book presents the basic steps in carrying out a research synthesis. It is intended for use by social and behavioral scientists who are unfamiliar with research synthesis and meta-analysis but who possess an introductory background in basic research methods and statistics.

Instead of a subjective, narrative approach to research synthesis, this book presents an objective, systematic approach. Herein, you will learn how to carry out an integration of research according to the principles of good science. The intended result is a research synthesis that can be replicated by others, can create consensus among scholars, and can lead to constructive debate on unresolved issues. Equally important, users of this approach should complete their research synthesis feeling knowledgeable and confident that their future primary research can make a contribution to the field.

The scientific approach to research synthesis has rapidly gained acceptance. In the years between its first and fifth editions, the procedures outlined in this book have changed from being controversial practices to being accepted ones. Indeed, in many fields the approach outlined herein is now obligatory. The years have also brought improvements in synthesis techniques. The technology surrounding literature searching has changed dramatically. The statistical underpinnings of meta-analysis—the quantitative combination of study results—have been developed and the application of these procedures has become widely accessible. Many techniques have been devised to help research synthesists present their results in a fashion that will be meaningful to their audience. Methodologists have proposed ways to make syntheses more resistant to criticism.

This fifth edition incorporates these changes. Most notably, Chapter 4 on conducting a literature search has been updated to include many of the recent developments wrought by the expanded use of the Internet for scientific communication. Many new developments have also occurred in the techniques for meta-analysis; these are covered in Chapter 6. They include new statistics for describing meta-analytic results and new techniques for combining complex data structures. The latter are touched on only briefly because they require more-advanced statistical training, unlike the other techniques I cover. Also, the references have been updated globally through the text.

Harris Cooper

Durham, North Carolina

Acknowledgments

About the Author

Harris Cooper
is Hugo L. Blomquist Professor in the Department of Psychology and Neuroscience at Duke University. He earned his doctorate degree in social psychology from the University of Connecticut. His research interests include research synthesis and applications of social and developmental psychology to educational policy issues including homework, school calendars, afterschool programs, and grading practices.

1 **Introduction** Literature Reviews, Research Syntheses, and Meta-Analyses

> ### This chapter describes
>
> - A justification for why attention to research synthesis methods is important
> - The goals of this book
> - A definition of the terms *research synthesis* and *meta-analysis*
> - A comparison of traditional narrative methods of research synthesis and methods based on scientific principles
> - A brief history of the development of the methods presented in this book
> - A seven-step model for the research synthesis process
> - An introduction to four research syntheses that will serve as practical examples in the chapters that follow

Much like a jigsaw puzzle you might do with family or friends, science is a cooperative, interdependent enterprise—only a puzzle in science is huge, and the puzzlers can span the globe and place pieces over decades. The hours you spend conducting a study contribute just one piece to a much larger puzzle. The value of your study will be determined as much by what direction it provides for future research (how it contributes to identifying the next needed puzzle piece) as from its own findings. Although it is true that some studies receive more attention than others, this is typically because the piece of the puzzle they solve (or the new puzzle they introduce) is important, not because they are puzzle solutions in and of themselves.

The Need for Attention to Research Synthesis

Science, then, is a cooperative and cumulative enterprise. As such, trustworthy accounts that describe past research are a necessary step in the orderly development of scientific knowledge. Untrustworthy accounts are similar to a puzzler forcing pieces to fit and putting pieces of the ocean in the sky. In order to make a contribution to our understanding of social and behavioral phenomena, researchers first need to know what is already known, with what certainty, and what remains unexplained. Yet, until four decades ago, social scientists paid little attention to how they conducted literature reviews that covered empirical findings, and how they located, evaluated, summarized, and interpreted past research. This omission from our research methodology became glaringly obvious when the explosion in the number of social researchers that occurred in the 1960s and 1970s resulted in a huge increase in the amount of social science research. It put in bold relief the lack of systematic procedures for conducting literature reviews that synthesized research. As the amount of research grew, so did the need for credible ways to integrate research findings, ways to ensure that fish did not fly and birds did not swim in our scientific puzzles.

Access to social science scholarship also has changed dramatically. In particular, the ability to find other people's research has been made easier by online reference databases and the Internet. Developing a list of research articles on a topic that interests you used to involve the lengthy and tedious scrutiny of printed compendia. Today, such lists can be generated, scrutinized, and revised with a few keystrokes. The number of reference databases you can search is hardly constrained by the time you have to devote to conducting your search. A half century ago, if you found an abstract of an article that interested you it could take weeks to communicate with its authors. Now, with electronic mail and file transfer, conversations and documents can be shared in seconds with the press of a button.

The need for trustworthy accounts of past research has also been heightened by growing specialization within the social sciences. Today, time constraints make it impossible for most social scientists to keep up with primary research except within a few topic areas of special interest to them. In 1971, Garvey and Griffith (1971) wrote,

> The individual scientist is being overloaded with scientific information. Perhaps the alarm over an "information crisis" arose because sometime in the last information doubling period, the *individual* psychologist became overburdened and could no longer keep up with and assimilate all the information being produced that was related to his primary specialty. (p. 350, emphasis in original)

What was true in 1971 is far truer today.

And finally, the call for use of evidence-based decision making has placed a new emphasis on the importance of understanding how a study was conducted, what it found, and what the cumulative evidence suggests is best practice (American Psychological Association's Presidential Task Force on Evidence-Based Practice, 2006). For example, in medicine there exists an international consortium of researchers, the Cochrane Collaboration (2015), producing hundreds of reports examining the cumulative

evidence on everything from public health initiatives to surgical procedures. In public policy, a similar consortium exists (Campbell Collaboration, 2015), as do organizations meant to promote government policy making based on rigorous evidence of program effectiveness (e.g., Coalition for Evidence-Based Policy, 2015). Each of these efforts, and many others, relies on trustworthy research syntheses to assist practitioners and policy makers in making critical decisions meant to improve human welfare.

Goals and Premises of the Book

This book is meant to serve as an introductory text on how to conduct a literature review of research and a meta-analysis in the social and behavioral sciences. The approach I will take applies the basic tenets of sound data gathering, analysis, and interpretation to the task of producing a comprehensive integration of past research on a topic. I will assume that you agree with me that the rules of rigorous, systematic social science inquiry are the same whether the inquirer is conducting a new data collection (a primary study) or a research synthesis. However, the two types of inquiry require techniques specific to their purpose.

There is one critical premise underlying the methods described in this text. It is that *integrating separate research projects into a coherent picture involves inferences as central to the validity of knowledge as the inferences involved in drawing conclusions from primary data analysis.* When you read a research synthesis, you cannot take for granted the validity of its conclusions, and that the author did a good job because you trust him or her; its validity must be evaluated against scientific standards. Social scientists performing a research synthesis make numerous decisions that affect the outcomes of their work. Each choice may enhance or undermine the trustworthiness of those outcomes. Therefore, if social science knowledge contained in research syntheses is to be worth believing, *research synthesists must meet the same rigorous methodological standards that are required of primary researchers.*

Judging the validity of primary research in the social sciences gained its modern foothold with the publication of Campbell and Stanley's (1963) monograph *Experimental and Quasi-Experimental Designs for Research*. A lineage of subsequent work refined this approach (e.g., Bracht & Glass, 1968; Campbell, 1969; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002). However, it was not until 15 years after Campbell and Stanley's pioneering work that social scientists realized they also needed a way to think about research syntheses that provided guidelines for evaluating the validity of syntheses that accumulated primary research outcomes.

This book describes (a) an organizing scheme for judging the validity of research syntheses, and (b) the techniques you can use to maximize the validity of conclusions drawn in syntheses you might conduct yourself.

Definitions of Literature Reviews

There are many terms that are used interchangeably to label the activities described in this book. These terms include *literature review, research review, systematic review, research synthesis,* and *meta-analysis*. In fact, some of these terms should be viewed as interchangeable, whereas some have broader or narrower meanings than others.

The term that encompasses all the rest is *literature review.* You would provide a brief literature review in the introduction to a report of new data. The scope of a literature review that introduces a new primary study typically is quite narrow: it will be restricted to those theoretical works and empirical studies pertinent to the specific issue addressed by the new study. The kind of literature review we are interested in here appears as a detailed independent work of scholarship. A literature review can serve many different purposes. It can have numerous different focuses and goals, take different perspectives in looking at the literature, cover more or less of the literature, and be written with different organizing principles for different audiences.

Based on interviews and a survey of authors, I presented a scheme for categorizing literature reviews (Cooper, 1988). This taxonomy is presented in Table 1.1. Most of the categories are easily understood. For instance, literature reviews can focus on the outcomes of research, research methods, theories, and/or applications of research to real-world problems. Literature reviews can have one or more goals: (a) to integrate (compare and contrast) what others have done and said, (b) to criticize previous scholarly works, (c) to build bridges between related topic areas, and/or (d) to identify the central issues in a field.

Table 1.1 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 of all studies
	Exhaustive with selective citation
	Representative citations
	Central or pivotal citations
Organization	Historical
	Conceptual
	Methodological
Audience	Specialized scholars
	General scholars
	Practitioners or policy makers
	General public

Petticrew and Roberts (2006) might add to my taxonomy a classification related to the time available to do the review. They use the term *rapid reviews* to describe reviews with a limited time for completion. Also, they use the term *scoping review* for a review meant to assess the types of relevant work currently in the literature and where they can be found. This type of review has the goal of helping the reviewers refine their research question (e.g., in terms of its conceptual breadth or years of coverage) and gauge the feasibility (in terms of time and resources) of conducting a full review. A scoping review is akin to a pilot study in primary research.

Literature reviews that combine two specific focuses and goals appear most frequently in the scientific literature. This type of literature review, and the focus of this book, has been alternately called a *research synthesis*, a *research review,* or a *systematic review. Research syntheses focus on empirical research findings and have the goal of integrating past research by drawing overall conclusions (generalizations) from many separate investigations that address identical or related hypotheses. The research synthesist's goal is to present the state of knowledge concerning the relation(s) of interest and to highlight important issues that research has left unresolved.* From the reader's viewpoint, a research synthesis is intended to "replace those earlier papers that have been lost from sight behind the research front" (Price, 1965, p. 513) and to direct future research so that it yields a maximum amount of new information.

A second kind of literature review that you will frequently encounter is a theoretical review. Here, the reviewer hopes to present the theories offered to explain a particular phenomenon and to compare them. The comparisons will examine the theories' breadth, internal consistency, and the nature of their

predictions. Typically, theoretical reviews contain descriptions of critical experiments already conducted and assessments of which theory is (a) most consistent with well-established research findings and (b) broadest in its ability to encompass the phenomena of interest. Sometimes theoretical reviews will also contain reformulations and integrations of notions drawn from different theories.

Often, a comprehensive literature review will address several sets of issues. Research syntheses are most common, however, and theoretical reviews will typically contain some synthesis of research. It is also not unusual for research syntheses to address multiple, related topics. For example, a synthesis might examine the relation between several different independent or predictor variables and a single dependent or criterion variable. For example, Scott and colleagues (2015) meta-analyzed the research on cognitive deficits that are associated with posttraumatic stress disorder. Nine cognitive domains were included in the meta-analysis: (1) attention/working memory, (2) executive functions, (3) verbal learning, (4) verbal memory, (5) visual learning, (6) visual memory, (7) language, (8) speed of processing, and (9) visuospatial abilities. The meta-analysis revealed that all of the cognitive deficits appeared more often in people classified as currently suffering from posttraumatic stress disorder but the strongest relationship (verbal learning) was about twice as large as the weakest relationship (visual memory).

As another example, a research synthesis might try to summarize research related to a series of temporally linked hypotheses. Harris and Rosenthal (1985) studied the mediation of interpersonal expectancy effects by first synthesizing research on how expectancies affect the behavior of the person who holds the expectation and then synthesizing research on how these behaviors influenced the behavior of the target.

This book is about research synthesis. Not only is this the most frequent kind of literature review in the social sciences, but it also contains many, if not most, of the decision points present in other types of reviews—and some unique ones as well. I have chosen to favor the label *research synthesis* over other labels for this type of literature review because the labels *research review* and *systematic review* occasionally cause confusion. They can also be applied to the process of peer review—that is, the critical evaluation of a manuscript that has been submitted for publication in a scientific journal. Thus, a journal editor may ask a scholar to provide a research review or a systematic review of a manuscript. The term *research synthesis* avoids this confusion and puts the synthesis activity front and center. Also, this label is used by *The Handbook of Research Synthesis and Meta-Analysis* (Cooper, Hedges, & Valentine, 2009), a text that describes approaches consistent with those presented here but in a more advanced manner.

The term *meta-analysis* is often used as a synonym for research synthesis, research review, or systematic review. In this book, meta-analysis will be employed solely to denote the quantitative procedures used to statistically combine the results of studies (these procedures are described in Chapter 5).

Why We Need Research Syntheses Based on Scientific Principles

Before the methods described in this book were available, most social scientists developed summaries of empirical research using a process in which multiple studies investigating the same topics were collected and described in a narrative fashion. These synthesists would describe one study after another, often arranged temporally, and then would draw a conclusion about the research findings based on their interpretation of what was found in the literature as a whole.

Research syntheses conducted in the traditional narrative manner have been much criticized. Opponents of the traditional research synthesis have suggested that this method—and its resulting conclusions—is imprecise in both process and outcome. In particular, traditional narrative research syntheses lack explicit standards of proof. Readers and users of these syntheses do not know what standard of evidence was used to decide whether a set of studies supported its conclusion (Johnson & Eagly, 2000). The combining rules used by traditional synthesists are rarely known to anyone but the synthesists themselves, if even they are consciously aware of what is guiding their inferences.

Four other disadvantages to traditional research syntheses, at least as they were carried out in the past, have often been leveled against this approach. First, traditional research syntheses rarely involve systematic procedures to ensure that all the relevant research was located and included in the synthesis. Traditional literature searches often stopped after the synthesists had gathered the studies they were already aware of or that they could locate through a search of a single reference database. Second, there was no way to check the accuracy with the information gathered from each study. Traditional research syntheses rarely, if ever, contained measures that assessed the reliability of the descriptions of the included research. Third, traditional narrative syntheses were prone to use post hoc criteria to decide whether individual studies met an acceptable threshold for methodological quality. This lack of explicit use of a priori quality standards led Glass (1976) to write,

> A common method of 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 or friends—and then advance the one or two "acceptable" studies as the truth of the matter. (p. 4)

Finally, traditional narrative syntheses, by their very nature, failed to result in statements regarding the overall magnitude of the relationship under investigation. They cannot answer the questions, "What was the size of the relationship between the variables of interest?" or "How much change was caused by the intervention?" or "Was this relationship or the effect of this intervention larger or smaller than that between other variables of interest or other interventions?"

Concern about the potential for error and imprecision in traditional narrative syntheses encouraged social science methodologists to develop the more rigorous and transparent alternatives described in this book. Today, state-of-the-art research syntheses use a collection of methodological and statistical techniques meant to reduce bias in accounts of the research, and to standardize and make explicit the procedures used to collect, catalog, and combine primary research. For example, today literature-searching strategies are designed to minimize differences between the results of retrieved studies and studies that were conducted but could not be uncovered by the literature search. Before the literature search begins, the criteria for deciding whether a study was conducted well enough to be included in the synthesis are explicitly stated. Then, these criteria are consistently applied to all studies, regardless of whether the results support or refute the hypotheses under investigation. Data from the research report are recorded using prespecified coding categories by coders trained to maximize interjudge agreement. Meta-analytic statistical methods are applied to summarize the data and provide a quantitative description of the cumulative research findings. Thus, research synthesis and the statistical integration of study results are conducted with the same rigorous procedures and are reported with the same transparency as is data analysis in primary scientific studies.

One example of how using state-of-the-art research synthesis methods can change cumulative findings was provided by a study conducted by Robert Rosenthal and me (Cooper & Rosenthal, 1980). In this study, graduate students and university faculty members were asked to evaluate a research literature on a simple research question: Are there sex differences in task persistence? All the participants in our study synthesized the same set of persistence studies but half of them used quantitative procedures and half used whatever criteria appealed to them—in other words, their own unstated inference test. We found statistical synthesists thought there was more support for the sex-difference hypothesis and a larger relationship between variables than did the other synthesists. Our finding revealed that synthesists we asked to use statistical techniques also tended to view future replications as less necessary than did other synthesists, although the difference between statistical and other synthesists did not reach statistical significance.

Principal Outcomes of a Research Synthesis

In addition to using a rigorous and transparent approach to cumulating the research, a state-of-the-art research synthesis is expected to provide information on several types of findings relating to the cumulative results of the research it covers. First, if a theoretical proposition is under scrutiny, readers of research syntheses will expect you to give them an overall estimate of the support for the hypothesis, both in terms of whether the null hypothesis can be rejected and the hypothesis' explanatory power—that is, the size of the relationship. Or if an intervention or public policy is under scrutiny, readers will expect you to estimate the effectiveness of the intervention or impact of the policy on the people it is meant to influence.

But you cannot stop there. Your audience also will expect to see tests of whether the relationship or estimate of effectiveness is influenced by variations in context. These may be suggested by characteristics of the theoretical hypothesis or intervention itself; how, when, and where the study was carried out; and who the participants were. Readers expect to be told whether the results of studies in your synthesis varied systematically according to characteristics of the manipulations or interventions, the settings and times at which the studies were conducted, differences between participants, characteristics of the measuring instruments, and so on.

A Brief History of Research Synthesis and Meta-Analysis

Above, I pointed out that the increase in social science research coupled with the new information technologies and the desire for trustworthy research syntheses in policy domains gave impetus to development of the methods described in this book. Here, I provide a brief history of people and events that have contributed to these techniques (see Cooper, Patall, & Lindsay, 2009, for a similar history).

Karl Pearson (1904) is credited with publishing what is believed to be the first meta-analysis (Shadish & Haddock, 2009). Pearson gathered data from 11 studies testing the effectiveness of a vaccine against typhoid and calculated for each a statistic he had recently developed, called the *correlation coefficient.* Based on the average correlations, Pearson concluded that other vaccines were more effective than the

new one.

In 1932 Ronald Fisher, in his classic text *Statistical Methods for Research Workers,* wrote, "It sometimes happens that although few or [no statistical tests] can be claimed individually as significant, yet the aggregate gives an impression that the probabilities are lower than would have been obtained by chance" (p. 99). Fisher was noting that statistical tests often fail to reject the null hypothesis because they lack statistical power. However, if the underpowered tests were combined, their cumulative power would be greater. For example, if you conduct a null hypothesis significance test and get a probability level of $p = .10$, the test is not statistically significant. But, what are the chances of getting a second independent test revealing $p = .10$ if the null hypothesis is true? Fisher presented a technique for combining the p-values that came from statistically independent tests of the same hypothesis.

Fisher's work would be followed by more than a dozen methodological papers published prior to 1960 (see Olkin, 1990), but the techniques were rarely put to use in research syntheses. Gene Glass (1976) introduced the term *meta-analysis* to mean the statistical analysis of results from individual studies "for purposes of integrating the findings" (p. 3). Glass (1977) wrote, "The accumulated findings of . . . studies should be regarded as complex data points, no more comprehensible without statistical analysis than hundreds of data points in a single study" (p. 352).

By the mid-1970s several high-profile applications of quantitative synthesis techniques focused the spotlight squarely on meta-analysis. Each of three research teams concluded that the traditional research synthesis simply would not suffice. Largely independently, they rediscovered and reinvented Pearson's and Fisher's solutions to their problem. In clinical psychology, Smith and Glass (1977) assessed the effectiveness of psychotherapy by combining 833 tests of the effectiveness of different treatment. In social psychology, Rosenthal and Rubin (1978) presented a research synthesis of 345 studies on the effects of interpersonal expectations on behavior. In education, Glass and Smith (1979) conducted a synthesis of the relation between class size and academic achievement. It included 725 estimates of the relation based on data from nearly 900,000 students. In personnel psychology, Hunter, Schmidt, and Hunter (1979) uncovered 866 comparisons of the differential validity of employment tests for black and white workers.

Independent of the meta-analysis movement but at about the same time, several attempts were made to draw research synthesis into a broad scientific context. In 1971, Feldman published an article titled "Using the Work of Others: Some Observations on Reviewing and Integrating," in which he wrote, "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" (p. 86). In the same year, Light and Smith (1971) argued that if treated properly, the variation in outcomes among related studies could be a valuable source of information, rather than a source of consternation, as it appeared to be when treated with traditional synthesis methods. Taveggia (1974) described six common activities in literature syntheses: selecting research; retrieving, indexing, and coding studies; analyzing the comparability of findings; accumulating comparable findings; analyzing the resulting distributions; and reporting the results.

Two articles that appeared in the *Review of Educational Research* in the early 1980s brought the meta-analytic and synthesis-as-research perspectives together. First, Jackson (1980) proposed six synthesis tasks "analogous to those performed during primary research" (p. 441). In 1982 I took the analogy between research synthesis and primary research to its logical conclusion and presented a five-stage model with accompanying threats to validity. This article was the precursor of the first edition of this book (Cooper, 1982).

Also in the 1980s, five books appeared that were devoted primarily to meta-analytic methods. Glass, McGaw, and Smith (1981) presented meta-analysis as a new application of analysis of variance and multiple regression procedures, with effect sizes treated as the dependent variable. Hunter, Schmidt, and Jackson (1982) introduced meta-analytic procedures that focused on (a) comparing the observed variation in study outcomes to that expected by chance and (b) correcting observed correlations and their variance for known sources of bias (e.g., sampling errors, range restrictions, unreliability of measurements). Rosenthal (1984) 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 effect size estimates were not based on traditional inferential statistics, but on a new set of techniques involving assumptions tailored specifically for the analysis of study outcomes. Light and Pillemer (1984) presented an approach that placed special emphasis on the importance of meshing both numbers and narrative for the effective interpretation and communication of synthesis results. Finally, in 1985, with the publication of *Statistical Methods for Meta-Analysis,* Hedges and Olkin helped elevate the quantitative synthesis of research to an independent specialty within the statistical sciences. Their book summarized and expanded nearly a decade of programmatic developments by the authors and established the procedures' legitimacy by presenting rigorous statistical proofs.

Since the mid-1980s, a large and growing number of books have appeared on research synthesis and meta-analysis. Some of these treat the topic generally (e.g., this text; Card, 2012; Lipsey & Wilson, 2001;

Petticrew & Roberts, 2006; Schmidt & Hunter, 2015), some treat it from the perspective of particular research designs (e.g., Bohning, Kuhnert, & Rattanasiri, 2008; Eddy, Hassleblad, & Schachter, 1992), and some are tied to particular software packages (e.g., Arthur, Bennett, & Huffcutt, 2001; Chen & Peace, 2013; Comprehensive Meta-Analysis, 2015). In 1994, the first edition of *The Handbook of Research Synthesis* was published, and the second edition appeared in 2009 (Cooper et al., 2009).[1]

The Stages of Research Synthesis

Textbooks on social research methodology present research projects as a sequenced set of activities. Although methodologists differ somewhat in their definitions of research stages, the most important distinctions in stages can be identified with a high degree of consensus.

As noted previously, I argued in 1982 that, similar to primary research, a research synthesis involved five distinct stages (Cooper, 1982). The stages encompass the principal tasks that need to be undertaken so that the synthesists produce an unbiased description of the cumulative state of evidence on a research problem or hypothesis. For each stage I codified the research question asked, its primary function in the synthesis, and the procedural differences that might cause variation in conclusions. For example, in both primary research and research synthesis, the problem formulation stage involves defining the variables of interest and the data collection stage involves gathering the evidence. Similar to primary data collectors, you can make different choices about how to carry out your inquiry; differences in your choices can create differences in your conclusions.

Most importantly, each methodological decision at each stage of a synthesis may enhance or undermine the trustworthiness of its conclusion or, in common social science parlance, can create threats to the validity of its conclusions. (A formal definition of the word *validity* appears in Chapter 4.) In my 1982 article and earlier editions of this book, I applied the notion of threats to inferential validity to research synthesis. I identified 10 threats to validity that might undermine the trustworthiness of the finding contained in a research synthesis. I focused primarily on validity threats that arise from the procedures used to cumulate studies—for example, conducting a literature search that missed relevant studies with a particular conclusion. This threats-to-validity approach was subsequently applied to research synthesis by Matt and Cook (1994, revised in 2009), who identified over 20 threats, and Shadish et al. (2002), who expanded this list to nearly 30 threats. In each case, the authors described threats related to potential biases caused by the process of research synthesis itself as well as to deficiencies in the primary research that made up the evidence base of the synthesis—for example, the lack of representation of important participant populations in the primary studies.

Table 1.2 summarizes a modification of the model that appeared in early editions of this book (Cooper, 2007, presented a six-step model). In my newest model, the process of research synthesis is divided into seven steps:

- Step 1: Formulating the problem
- Step 2: Searching the literature
- Step 3: Gathering information from studies
- Step 4: Evaluating the quality of studies
- Step 5: Analyzing and integrating the outcomes of studies
- Step 6: Interpreting the evidence
- Step 7: Presenting the results

These seven steps will provide the framework for the remainder of this book. Different from my earlier conceptualization, the new model separates two of the stages into four separate stages. First, the (a) literature search and (b) the process of extracting information from research reports are now treated as two separate stages. Second, the processes of (a) summarizing and integrating the evidence from individual studies and (b) interpreting the cumulative findings that arise from these analyses are treated separately. These revisions are based on much recent work that suggests these activities are best thought of as independent. They require separate decisions on the part of the synthesists and make use of distinct methodological tools. For example, you can thoroughly or cursorily search a literature. Then you can code much or little information from each report, in a reliable or unreliable manner. Similarly, you can correctly or incorrectly summarize and integrate the evidence from the individual studies and then, even if correctly summarized, interpret what these cumulative findings mean either accurately or inaccurately.

Table 1.2 Research Synthesis Conceptualized as a Research Project

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 (a) variables and (b) 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 (a) deemed relevant and/or (b) tested as moderating influences.
Searching the Literature	What procedures should be used to find relevant research?	Identify (a) sources (e.g., reference databases, journals) and (b) 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 (a) in information gathered might lead to differences in what is tested as an influence on cumulative results, (b) in coder training might lead to differences in entries on coding sheets, and/or

Table 1.2 (Continued)

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
			(c) 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 (a) the suitability of the methods for studying the synthesis question, and/or (b) 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 summarize and combine the research results?	Identify and apply procedures for (a) combining results across studies and (b) testing for differences in results between studies.	Variation in procedures used to summarize and compare results of included studies (e.g., narrative, vote count, averaged effect sizes) can lead to differences in cumulative results.

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
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 (a) criteria for labeling results as important and (b) 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 (a) lead readers to place more or less trust in synthesis outcomes and (b) influence others' ability to replicate results.

Also, I should note that the process of conducting a rigorous research synthesis, indeed any rigorous research, is never as linear as described in textbooks. You will find that "problems" you encounter at later stages in your synthesis will require you to backtrack and change decisions you made at an earlier stage. For example, your literature search might uncover studies that suggest you redefine the topic you are considering. Or, a dearth of studies with the desired design—for example studies with experimental manipulations—suggests you include other types of designs, such as studies that only correlated the variables of interest. For this reason, it is good to start with a plan for your synthesis in its entirety but remain open to the possibility of altering it as the project progresses.

Step 1: Formulating the Problem

The first step in any research endeavor is to formulate the problem. During problem formulation, the variables involved in the inquiry are given both abstract and operational definitions. At this stage you ask, "What are the concepts or interventions I want to study?" and "What operations are measureable expressions of these concepts and the outcomes that interest me?" In answering these questions, you determine what research evidence will be relevant (and irrelevant) to the problem or hypothesis of interest. Also, during problem formulation, you decide whether you are interested in simply describing the variable(s) of interest or in investigating a relationship between two or more variables, and whether this relationship is associational or causal in nature.

In Chapter 2 I examine the decision points you will encounter during the problem formulation stage. These decision points relate first and foremost to the breadth of the concepts involved in the relations of interest and how these correspond to the operations used to study them. They also relate to the types of research designs used in the primary research and how these correspond to the inferences you wish to make.

Step 2: Searching the Literature

The data collection stage of research involves making a choice about the population of elements that will be the target of the study. In primary social science research, the target will typically include human individuals or groups. In research synthesis, identifying target populations is complicated by the fact that you want to make inferences about two targets. First, you want the cumulative result to reflect the results of all previous research on the problem. Second, you hope that the included studies will allow generalizations to the individuals or groups that are the focus of the topic area.

In Chapter 3 I present a discussion of methods for locating studies. The discussion includes a listing of the sources of studies available to social scientists, how to access and use the most important sources,

and what biases may be present in the information contained in each source.

Step 3: Gathering Information From Studies

The study coding stage requires that researchers consider what information they want to gather from each unit of research. In primary research, the data-gathering instruments might include questionnaires, behavior observations, and/or physiological measures. In research synthesis, this involves the information about each study that you have decided is relevant to the problem of interest. This information will include not only characteristics of the studies that are relevant to the theoretical or practical questions—that is, about the nature of the independent and dependent variables—but also about how the study was conducted, its research design, implementation, and statistical results. Beyond deciding what information to collect and giving this clear definition, this stage requires that you develop a procedure for training the people who will gather the information and ensuring that they do so in a reliable and interpretable manner.

Chapter 4 will present some concrete recommendations about what information you should collect from empirical studies that have been judged relevant to your problem. It also introduces the steps that need to be taken to properly train the people who will act as study coders. Also, Chapter 4 contains some recommendations concerning what you can do when research reports are unavailable or when obtained reports do not have the information you need in them.

Step 4: Evaluating the Quality of Studies

After data are collected, the researcher makes critical judgments about the "quality" of data, or its correspondence to the question that is motivating the research. Each data point is examined in light of surrounding evidence to determine whether it is too contaminated by factors irrelevant to the problem under consideration to be of value in the research. If it is, the bad data must be discarded or given little credibility. For example, primary researchers examine how closely the research protocol was followed when each participant took part in the study. Research synthesists evaluate the methodology of studies to determine if the manner in which the data were collected might make it inappropriate for addressing the problem at hand.

In Chapter 2, I discuss how research designs (e.g., associational or causal) correspond to different research problems and in Chapter 5 I discuss how to evaluate the quality of research. I also look at biases in quality judgments and make some suggestions concerning the assessment of interjudge reliability.

Step 5: Analyzing and Integrating the Outcomes of Studies

During data analysis, the separate data points collected by the researcher are summarized and integrated into a unified picture. Data analysis demands that the researcher distinguish systematic data patterns from "noise" (or chance fluctuation). In both primary research and research synthesis, this process typically involves the application of statistical procedures.

In Chapter 6 I explain some methods for combining the results of separate studies, or methods of meta-analysis. Also, I show how to estimate the magnitude of a relationship or the impact of an intervention. Finally, I illustrate some techniques for analyzing why different studies find different relationship strengths.

Step 6: Interpreting the Evidence

Next, the researcher interprets the cumulative evidence and determines what conclusions are warranted by the data. These conclusions can relate to the evidence with regard to whether the relation(s) of interest are supported by the data and, if so, with what certainty. They can also relate to the generality (or specificity) of the findings over different types of units, treatments, outcomes, and situations.

In Chapter 7 I examine some of the decision rules you should apply as you make assertions about what your research synthesis says. This includes some ideas about interpreting the strength and generality of conclusions as well as the magnitude of relationships or intervention effects.

Step 7: Presenting the Results

Creating a public document that describes the investigation is the task that completes a research endeavor. In Chapter 8 I offer some concrete guidelines for what information needs to be reported regarding how the other six stages of the research synthesis were carried out.

Twenty Questions About Research Syntheses

I will frame the discussion of the stages of research synthesis by referring to 20 questions producers and consumers of research syntheses might ask that relate to the validity of conclusions. In my teaching, I have found this approach is easy to follow and helps students keep the big picture in mind as they move through the process. Each question is phrased so that an affirmative response would mean confidence could be placed in the conclusions of the synthesis. The relevant questions will be presented at the beginning of the discussion of each stage and will be followed by the related procedural variations that might enhance or compromise the trustworthiness of conclusions—in other words, what needs to be done to answer the question "yes." Although the 20 questions are not an exhaustive list of those that might be asked, most of the threats to validity identified in early editions of this work find expression in the questions. A list of the questions appears in Table 1.3. I will return to a discussion of the threats to validity of a research synthesis in Chapter 9.

Table 1.3 A Checklist of Questions Concerning the Validity of Research Synthesis Conclusions

Step 1: Formulating the problem

1. Are the variables of interest given clear conceptual definitions?

2. Do the operations that empirically define each variable of interest correspond to the variable's conceptual definition?

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, and/or practical context?

Step 2: Searching the literature

5. Were proper and exhaustive terms used in searches and queries of reference databases and research registries?

6. Were complementary searching strategies used to find relevant studies?

Step 3: Gathering information from studies

7. Were procedures employed to ensure the unbiased and reliable (a) application of criteria to determine the substantive relevance of studies and (b) retrieval of information from study reports?

Step 4: Evaluating the quality of studies

8. If studies were excluded from the synthesis because of design and implementation considerations, were these considerations (a) explicitly and operationally defined and (b) consistently applied to all studies?

9. Were studies categorized so that important distinctions could be made among them regarding their research design and implementation?

Step 5: Analyzing and integrating the outcomes of studies
10. Was an appropriate method used to combine and compare results across studies?
11. If a meta-analysis was performed, 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 above) along with (b) other critical features of studies, including historical, theoretical, and practical variables (as suggested by Question 4 above) tested as potential moderators of study outcomes?
Step 6: Interpreting the evidence
15. Were analyses carried out that tested whether results were sensitive to statistical assumptions and, if so, were 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' 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 synthesis-generated evidence when interpreting the synthesis' results?
19. If a meta-analysis was performed, did the synthesists (a) contrast the magnitude of effects with other related effect sizes and/or (b) present a practical interpretation of the significance of the effects?
Step 7: Presenting the results
20. Were the procedures and results of the research synthesis clearly and completely documented?

SOURCE: Adapted from Cooper, H. (2007). *Evaluating and Interpreting Research Syntheses in Adult Learning and Literacy.* Boston: National Center for the Study of Adult Learning and Literacy, World Education, Inc., p. 52.

Four Examples of Research Synthesis

I have chosen four research syntheses to illustrate the practical aspects of conducting rigorous summaries of research. The topics of the four syntheses represent a broad spectrum of social and behavioral science research, encompassing research from basic and applied social psychology, developmental psychology, curriculum and instruction in education, and the health-related professions. They involve diverse conceptual and operational variables. Some are also interdisciplinary in nature. More and more, research involves scholars drawn from different disciplines. Research syntheses are no different. One example I use—on aerobic exercise—involved researchers from a department of psychiatry and behavioral medicine in a School of Medicine and others from a department of psychology and neuroscience in a College of Arts and Sciences. In these circumstances, the different team members bring different perspectives on the problem. This can help with the identification of what variations in the conceptual and operational definition of variables will be important as well as where to look for relevant studies. It is not unusual for these teams to include a member who has advanced knowledge of the statistical techniques needed to perform a quantitative integration of the results of studies.

Even though the topics are very different, they are also general enough that readers in any discipline should find all four topics instructive and easy to follow without a large amount of background in the separate research areas. Most importantly, they cover research syntheses involving research designs that

have relevance to any topic area. You should be able to find among them a research paradigm that fits your particular topic of interest. A brief introduction to each topic will be helpful.

The Effects of Choice on Intrinsic Motivation (Patall, Cooper, & Robinson, 2008)

The ability to make personal choices—be they between courses of action, products, or candidates for political office, to name just a few—is central to Western culture. Not surprisingly, then, many psychological theories posit that providing individuals with choices between tasks will improve their motivation to engage in the chosen activity. In this research synthesis, we examined the role of choice in motivation and behavior. First, we examined the overall effect of choice on intrinsic motivation and related outcomes. We also examined whether the effect of choice was enhanced or diminished by a number of theoretically derived moderators including the type of choice, the number of options in the choice, and the total number of choices made. In this synthesis, the studies primarily used experimental designs and were conducted in social psychology laboratories.

The study was published in a journal serving a broad audience. It draws its topic from literatures in both social and developmental psychology. All the research designs it covers involved experimental manipulations with random assignment of subjects to conditions.

The Effect of Homework on Academic Achievement (Cooper, Robinson, & Patall, 2006)

Requiring students to carry out academic tasks during nonschool hours is a practice as old as formal schooling itself. However, the effectiveness of homework is still a source of controversy. Public opinion about homework fluctuated throughout the 20th century, and the controversy continues today. This synthesis focused on answering a simple question reflected in the article's title: "Does homework improve academic achievement?" We also looked at moderators of homework's effects, including the student's grade level and the subject matter.

This research synthesis focuses on a topic drawn from the education literature on instruction. It involved summarizing results from a few experimental studies using random and nonrandom (whole classroom) assignment. These studies were conducted in actual classrooms. Several studies that applied statistical models (multiple regressions, path analyses, structural equation models) to large databases were also included, as were many studies that simply correlated the time a student spent on homework with a measure of academic achievement.

Individual Differences in Attitudes Toward Rape (Anderson, Cooper, & Okamura, 1997)

Rape is a serious social problem. Every day, many women are forced by men to have sex without the woman's consent. This research synthesis examined the demographic, cognitive, experiential, affective, and personality correlates of attitudes toward rape. We found research that looked at the attitudes of both men and women. Demographic correlates of attitudes toward rape included age, ethnicity, and socioeconomic status (SES). Experiential correlates included involvement in previous rapes, knowing others who had been in a rape, and use of violent pornography. Personality correlates included the need for power and self-esteem. What value is there in summarizing research on rape attitudes? We hoped our synthesis would be used to improve programs meant to prevent rape by helping identify people who would benefit most from rape prevention interventions.

These studies were drawn from applied social psychology and were all correlational in nature. It cumulated studies associating a measure of an attitude or belief (about rape) with an individual differences measure.

Aerobic Exercise and Neurocognitive Performance (Smith et al., 2010)

Does physical exercise improve our ability to focus on and remember things? If so, exercise interventions could be used to counteract losses in attention, executive functioning (the ability to manage or regulate cognitive tasks), and memory. This might provide physicians with a way to forestall cognitive impairment due to age and dementia and even to lengthen life. While many studies have been conducted on whether exercise improves neurocognitive performance, we found that past reviews of this literature could not come to consensus on the magnitude of the effect. Nor did past reviews carefully examine possible influences on the results of different studies. Therefore, we conducted a meta-analysis examining (a) the effects of aerobic exercise interventions on cognitive abilities such as attention, processing speed and executive functioning, working memory, and memory; (b) how features of the exercise intervention (e.g.,

its components, duration, and intensity) influenced its outcomes; and (c) how individual differences between participants (e.g., age, initial level of cognitive functioning) might influence exercise effects. We included only studies that used experimental manipulations of exercise and randomly assigned participants to conditions.

This synthesis was based on studies of health interventions. They were all experimental in nature and used random assignment of subjects in field settings.

> **Exercise**
>
> The best way to benefit from reading this book is to plan and conduct a research synthesis in an area of interest to you. The synthesis should attempt to apply the guidelines outlined in the chapters that follow. If such an ambitious undertaking is not possible, you should try to conduct the more discrete exercises that appear at the end of each chapter. Often, these exercises can be further simplified by dividing the work among members of your class.

Note

1. Chalmers, Hedges and Cooper, (2002) also present a brief history of meta-analysis. Hunt (1997) wrote a popular book describing the early history of meta-analysis that contains interviews with the principal contributors. A special issue of the journal *Research Synthesis Methodology* (2015) provides first person accounts by developers of the early research synthesis and meta-analytic methods.

2 Step 1 Formulating the Problem

What research evidence will be relevant to the problem or hypothesis of interest in the synthesis?

Primary Function Served in the Synthesis

To define the (a) variables and (b) relationships of interest so that relevant and irrelevant studies can be distinguished from one another

Procedural Variation That Might Produce Differences in Conclusions

Variation in the conceptual breadth and distinctions within definitions might lead to differences in the research operations (a) deemed relevant and/or (b) tested as moderating influences.

Questions to Ask When Evaluating the Formulation for a Problem in a Research Synthesis

1. Are the variables of interest given clear conceptual definitions?
2. Do the operations that empirically define each variable of interest correspond to the variable's conceptual definition?
3. Is the problem stated so the research designs and evidence needed to address it can be specified clearly?
4. Is the problem placed in a meaningful theoretical, historical, and/or practical context?

This chapter describes

- The relationship between concepts and operations in research synthesis
- How to judge the relevance of primary research to a research synthesis problem
- The correspondence between research designs and research synthesis problems
- The distinction between study-generated and synthesis-generated evidence
- The treatment of main effects and interactions in research synthesis
- Approaches to establishing the value of a new research synthesis
- The role of previous syntheses in new synthesis efforts

All empirical research begins with a careful consideration of the problem that will be the focus of study. In its most basic form, the research problem includes the definition of two variables and the rationale for studying their association. One rationale can be that a theory predicts a particular association between the variables, be it a causal relationship or a simple association, positive or negative. For example, self-determination theory (Deci & Ryan, 2013) predicts that providing people with choices in what task to perform or how to perform it will have a positive causal effect on people's intrinsic motivation to do the task and persist at it. So manipulating choice, then measuring intrinsic motivation, will provide evidence on the veracity of the theory. Or a different rationale can be that some practical consideration suggests that any discovered relation might be important. For example, discovering the individual differences that correlate with attitudes about rape, even if there is little theory to guide us about what relationships to expect, might suggest ways to improve programs meant to prevent rape by helping identify people who would benefit most from different types of prevention interventions. Either rationale can be used for undertaking primary research or research syntheses.

The choice of a problem to study in primary research is influenced by your interests and the social conditions that surround you. This holds true as well for your choice of topics in research synthesis, with one important difference. When you do primary research, you are limited in your topic choice only by your imagination. When you conduct a research synthesis, you must study topics that already appear in the literature. In fact, a topic is probably not suitable for research synthesis unless it already has created sufficient interest within a discipline or disciplines to inspire enough research to merit an effort at bringing it all together.

The fact that syntheses are tied to only those problems that have generated previous research does not mean research synthesis is less creative than primary data collection. Rather, your creativity will be used in different ways in research synthesis. Creativity enters a research synthesis when you must propose overarching schemes that help make sense of many related, but not identical, studies. The variation in methods across studies is always much greater than the variation in procedures used in any single study. For example, studies of choice and intrinsic motivation vary in the types of choices they allow, some involving choices among tasks (e.g., anagrams versus number games) and others involving choices among the circumstances under which the task will be performed (e.g., the color of the stimuli, whether to use a pen or pencil), to name just two types of variations.

As a synthesist, you may find little guidance about how these variations should be meaningfully grouped to determine if they affect the relationship between choice and motivation. (Will grouping the choice

manipulations depending on whether they are task relevant versus task irrelevant lead to an important discovery?) Or theories may suggest meaningful groupings, but it will be up to you to discover what these theoretical predictions are. (What does self-determination theory say the effect of task relevance should be on how the ability to choose affects motivation?) Defining meaningful groupings of studies and justifying their use will be up to you. Your capacity for uncovering variables that explain why results differ in different studies and your ability to generate explanations for these relationships are creative and challenging aspects of the research synthesis process.

Definition of Variables in Social Science Research

Similarities in Concepts and Operations in Primary Research and Research Synthesis

The variables involved in any social science study must be defined in two ways. First, the variables must be given conceptual definitions. The term *conceptual definitions* describes qualities of the variable that are independent of time and space but can be used to distinguish observable events that are and are not relevant to the concept. For instance, a conceptual definition of the word *achievement* might be "a person's level of knowledge in academic domains." The term *neurocognitive functioning* might be conceptually defined as "mental processes associated with particular areas of the brain." The term *homework* might be conceptually defined as "tasks assigned by teachers meant to be carried out during nonschool hours."

Conceptual definitions can differ in their breadth—that is, in the number of events to which they refer. Thus, if achievement is defined as "something gained through effort or exertion," the concept is broader than it is if you use the definition in the paragraph above, relating solely to academics. The second definition would consider as achievement the effort exerted in social, physical, and political spheres, as well as academic ones. When concepts are broader, we also can say they are more abstract.

Both primary researchers and research synthesists must choose a conceptual definition and a degree of breadth for their problem variables. Both must decide how likely it is that an event represents an instance of the variable of interest. Although it is sometimes not obvious, even very concrete variables, such as homework, require conceptual definitions. So, the first question to ask yourself about how you have formulated the problem for your research synthesis is,

Are the variables of interest given clear conceptual definitions?

In order to relate concepts to observable events, a variable must also be operationally defined. An *operational definition* is a description of observable characteristics that determine if the event represents an occurrence of the conceptual variable. Put differently, a concept is operationally defined when the procedures used to make it observable and measurable are openly and distinctly stated. For example, an operational definition of the concept *intrinsic motivation* might include "the amount of time a person spends on a task during a free-time period." Again, both primary researchers and research synthesists must specify the operations included in their conceptual definitions.

Differences in Concepts and Operations in Primary Research and Research Synthesis

Differences in how variables are defined can also be found between the two types of research. Primary researchers have little choice but to define their concepts operationally before they begin their studies. They cannot start collecting data until the variables in the study have been given an empirical reality. Primary researchers studying choice must define how choice will be manipulated or measured before they can run their first participant.

On the other hand, research synthesists need not be quite so operationally precise, at least not initially. For them, the literature search can begin with only a conceptual definition and a few known operations relevant to it. Then, the associated operations can be filled out as the synthesists become more familiar with the research literature. For example, you might know that you are interested in interventions meant to increase physical activity among adults. Once you begin the literature search, you might also find types of interventions you were unaware existed. You might have thought of exercises classes but then find in the literature interventions involving self-monitoring (keeping a diary of physical activity), social modeling (watching others exercise), and providing a health-risk appraisal. Each of these might encourage exercise without directly manipulating it. You might also find interventions that involve lifting weights and other exercises that increase strength but do not involve aerobic activities (e.g., walking, jogging, biking). As a research synthesist, you have the comparative luxury of being able to evaluate the conceptual relevance of different operations as you find them in the literature. You can even change your conceptual definition

depending on the potentially relevant operational definitions your concept might cover that had not occurred to you when you began. Is weight training an intervention you are interested in if you are studying neurocognitive functioning, or is your conceptual definition better cast as "aerobic exercise interventions," thus excluding weight training? Primary researchers do not have this luxury, at least not without considerable retooling of their study after it has begun.

Of course, some a priori specification of operations is necessary, and you need to begin your synthesis with a conceptual definition and at least a few empirical realizations in mind. However, during a literature search, it is not unusual to come across operations that you did not know existed but are relevant to the construct you are studying. In sum, primary researchers must know exactly what operational definitions are of interest (i.e., those that will be measured or manipulated in their study) before they begin collecting data. Research synthesists may discover unanticipated operations that fit into the relevant domain along the way.

Another distinction between the two types of inquiry is that primary studies typically involve only one or a few operational definitions of the same construct. A particular exercise regimen or measure of academic achievement must be in hand before data collection begins. In contrast, research syntheses usually involve many empirical realizations for each variable of interest. Although no two participants[1] are treated exactly alike in any single study, this variation will ordinarily be small compared to variation introduced by the differences in the way participants are treated and outcomes are measured in separate studies. For example, a single study of choice and motivation might involve giving participants a choice to do either anagrams or sudokus. However, the synthesists looking at all the choice studies that have been conducted might find manipulations using anagrams, crosswords, sudokus, word finds, cryptograms, video games, and so on. Add to this the fact that research synthesists will also find much greater variation in the location in which studies were conducted (different geographical regions, labs, classrooms, or places of work) and in sampled populations (college students, children, or employees). The multiple operations contained in research syntheses introduce a set of unique issues that need to be examined carefully.

Multiple Operations in Research Synthesis

Research synthesists must be aware of two potential incongruities that can arise because of the variety of operations they encounter in the literature. First, you might begin a literature search with broad conceptual definitions in mind. However, you may discover that the operations used in previous relevant research have been narrower than your concepts imply. For instance, a synthesis of research on rape attitudes might begin with a broad definition of rape, including any instance of unwanted sexual relations, even women forcing sex on men. However, the literature search might reveal that past research dealt only with men as the perpetrators of rape. When such a circumstance arises, you must narrow the conceptual underpinnings of the synthesis to be more congruent with existing operations. Otherwise, its conclusions might appear more general than warranted by the data.

The opposite problem, starting with narrow concepts but then encountering operational definitions that suggest the concepts of interest should be broadened, can also confront a synthesist. Our example regarding the definition of "achievement" illustrates this problem. You might begin a search for studies on homework and achievement expecting to define achievement as relating solely to academic material. However, in perusing the literature, you might encounter studies of homework in classes on music and industrial arts, for example. These studies fit the definition of "homework" (i.e., tasks assigned by teachers meant to be carried out during nonschool hours), but the outcome variables might not fit the definition of achievement because they are not measures of verbal or quantitative ability. Should these studies be included? It would be fine to do so but you would have to make it clear that your conceptual definition of achievement now has broadened to include performance in nonacademic domains.

When conducting a research synthesis, *as your literature search proceeds, it is very important that you take care to reevaluate the correspondence between the breadth of the definitions of the concepts of interest and the variation in operations that primary researchers have used to define them.* Thus, the next question to ask yourself as you evaluate how well you have specified the problem for your research synthesis is,

> Do the operations that empirically define each variable of interest correspond to the variable's conceptual definition?

Make certain that your decisions to include certain studies have not broadened your definitions or that operations missing in the literature do not suggest that the conceptual definitions need to be narrowed. In primary research, this redefinition of a problem as a study proceeds is frowned upon. In research synthesis, it appears that some flexibility may be necessary, indeed beneficial.

Multiple Operationism and Concept-to-Operation Correspondence

Webb, Campbell, Schwartz, Sechrest, and Grove (2000) presented strong arguments for the value of having multiple operations to define the same underlying construct. They define the term *multiple operationism* as the use of many measures that share a conceptual definition "but have different patterns of irrelevant components" (p. 3). Having multiple operations of a construct 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., pp. 3–4)

While Webb and colleagues hold out the potential for strengthened inferences when a variety of operations exists, as happens in a research synthesis, their parting qualification also must not be ignored. Multiple operations can enhance concept-to-operation correspondence if the operations encompassed in your research synthesis are individually at least minimally related to the construct (Eid & Diener, 2006). This reasoning is akin to the reasoning applied in classical measurement theory. Small correlations between individual items on a multi-item test, say the items on an achievement test, and a participant's "true" achievement score can add up to a reliable indicator of achievement *if* a sufficient number of minimally valid items are present. Likewise, the conclusions of a research synthesis will not be valid if the operations in the covered studies bear no correspondence to the underlying concept or if the operations share a different concept to a greater degree than they share the intended one. This is true regardless of how many operations are included.

For example, it is easy to see the value of multiple operations when thinking about outcome variables. We are confident that homework affects the broad conceptual variable "achievement" when we have measures of achievement that include teacher-constructed unit tests, class grades, and standardized achievement tests, and the relationship between homework and achievement is in the same direction regardless of the achievement measure. We are less confident that the relationship exists if only class grades are used as outcomes. If only class grades are used, it may be that teachers include grades on homework assignments in the class grade and this explains the relationship, whereas homework might have no effect if unit tests or standardized tests serve as measures. These tests do not share the same source of error. But unit tests are highly aligned with the content of assignments, whereas standardized achievement tests typically are not.

Thus, when multiple operations provide similar results, they suggest the operations converge on the same construct, and our confidence grows in the conclusions. If the different operations do not lead to similar results, differences between the operations can give us clues about limitations to our conclusions. For example, if we find homework influences unit tests but not standardized tests, we might speculate that homework influences achievement only when the content of assignments and measures of achievement are highly aligned.

The value of multiple operations of independent variables (those manipulated in experiments meant to test theories) or intervention variables (treatments in applied settings) also can increase our confidence in conclusions. For example, if experimental studies of exercise interventions were all conducted using the same duration and intensity of exercises, we would not know whether more or less exercise might have different effects. Is there a threshold below which exercise has no effect? Can too much exercise cause fatigue that actually interferes with cognitive functioning?

In sum, the existence of a variety of operations in research literatures presents the potential benefit of allowing stronger inferences if the results allow you to rule out irrelevant sources of influence. If results are inconsistent across operations, it allows you to speculate on what the important differences between operations might be.

The use of operations not originally related to the concept.

Literature searches can sometimes uncover research that has been cast in a conceptual framework different from the one you want to study but that includes operational measures or manipulations relevant to the concepts of interest to you. For instance, there are several concepts similar to "job burnout" that appear in the research literature, such as "occupational stress" and "job fatigue." It is important to consider whether the operations associated with these different constructs are relevant to your synthesis, even if they have been labeled differently. When relevant operations associated with different abstract constructs are identified, they most certainly should be considered for inclusion in your synthesis. In fact, different concepts and theories behind similar operations can often be used to demonstrate the robustness of results. There probably is no better way to ensure that operations contain

different patterns of error than to have different researchers with different theoretical backgrounds perform related investigations.

Substituting new concepts for old.

Sometimes you will find that social and behavioral scientists introduce new concepts (and theories) to explain old findings. For example, in a classic social psychology experiment, the notion of "cognitive dissonance" was used to explain why an individual who is paid $1 to voice a counterattitudinal argument subsequently experiences greater attitude change than another person paid $25 to perform the same activity (Festinger & Carlsmith, 1959). Dissonance theory suggests that because a small amount of money is not sufficient to justify the espousal of the counterattitudinal argument, the person feels discomfort that can be reduced only through a shift in attitude. However, Bem (1967) recast the results of this experiment by proposing a self-perception theory. Briefly, he speculated that participants who observed themselves espousing counterattitudinal arguments inferred their opinions the same way as an observer: if participants see themselves making an argument for $1, they assume that because they are performing the behavior with little justification, they must feel positive toward the attitude in question (just like an observer would infer).

No matter how many replications of the $1/$25 experiment you uncovered, you could not use the results to evaluate the correctness of either of the two theories. You must take care to differentiate concepts and theories that predict similar and different results for the same set of operations. If predictions are different, the cumulative evidence can be used to evaluate the correctness of one theory or another, or the different circumstances in which each theory is correct. However, if the theories make identical predictions, no comparative judgment based on research outcomes is possible.

The effects of multiple operations on synthesis outcomes.

Multiple operations do more than introduce the potential for more-nuanced inferences about conceptual variables. They are also the most 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 two ways:

1. *Differences in the included operational definitions.* The operational definitions used in two research syntheses on the same topic can be different from one another. Two synthesists using an identical label for an abstract concept can search for and include different operational definitions. Each definition may contain some operations excluded by the other, or one definition may completely contain the other.
2. *Differences in operational detail.* Multiple operations also affect outcomes by leading to variation in the attention that synthesists pay to methodological distinctions in the literature. This effect is attributable to differences in the way study operations are treated *after* the literature has been searched. At this point, research synthesists become detectives who search for "distinctive clues about why two variables are related differently under different conditions" (Cook et al., 1992, p. 22). They use the observed data patterns as clues for generating explanations that specify the conditions under which a positive, null, or negative relationship will be found between two variables.

Synthesists differ in how much detective work they undertake. Some pay careful attention to study operations. They decide to identify meticulously the operational distinctions among retrieved studies. Other synthesists believe that method- or participant-dependent relations are unlikely, or they may simply use less care in identifying these relations.

Defining the Relationship of Interest

Whether you are doing primary research or research synthesis, in addition to defining the concepts you must also decide what type of relationship between the variables is of interest to you. While your conceptual definition of the variables will determine the relevance of different operations, it is the type of relationship that will determine the relevance of different research designs. In order to be able to determine the appropriateness of different research designs, there are three questions that need to be asked about the problem that motivates your research synthesis (see Cooper, 2006, for a more complete discussion of these issues):

1. Should the results of the research be expressed in numbers or narrative?
2. Is the problem you are studying a description of an event, an association between events, or a causal explanation of an event?
3. Does the problem or hypothesis seek to understand (a) how a process unfolds *within* an individual participant over time, or (b) what is associated with or explains variation *between* participants or groups of participants?

Quantitative or Qualitative Research?

With regard to the question, "Should the results of the research be expressed in numbers or narrative?" it should be clear that for the type of research synthesis I am focusing on here, the answer is "numbers." However, this does not mean that narrative or qualitative research will play no role in quantitative research syntheses. For example, in our synthesis of homework research, qualitative studies were used to help compile a list of possible effects of homework, both good and bad. In fact, even opinion pieces were used, such as complaints about homework ("It creates too much stress for children") that appeared in newspaper articles.

Qualitative research also was used to help identify possible moderators and mediators of homework's effects. For example, the homework literature search uncovered a survey and interview study (Younger & Warrington, 1996) that suggested girls generally hold more positive attitudes than boys toward homework and expend greater effort on doing homework. This study suggested that this individual difference among students might moderate relationships between homework and achievement. A case study of six families by Xu and Corno (1998) involved both interviews and home videotaping to examine how parents structure the homework environment and help children cope with distractions so they can pay attention to the homework assignment. This study clearly argued for the importance of parents as mediators in the homework process.

Of course, the results of qualitative research can also be the central focus of a research synthesis, not just an aid to quantitative synthesis. Discussions of how to carry out such reviews have occupied the thoughts of scholars much better versed in qualitative research than me. If you are interested in this type of research synthesis, you might examine Sandelowski and Barroso (2007) and/or Pope, Mays, and Popay (2007) for detailed examinations of approaches to synthesizing qualitative research.

Description, Association, or Causal Relationship?

Descriptive research.

The second question, "Is the problem you are studying a description of an event, an association between events, or a causal explanation of an event?" divides research problems into three groups. First, a research problem might be largely descriptive and take the general form, "What is happening?" Here, you might be interested in obtaining an accurate portrayal of some event or other phenomenon. In primary research, this might lead you to conduct a survey (Fowler, 2014). For example, older adults might be asked questions about the frequency of their physical activity. Your conclusion might be that "X% of adults over the age of Y routinely engage in physical activity." In research synthesis, you would collect all the surveys that asked a particular question and, perhaps, average the estimates of frequency in order to get a more precise estimate. Or you might examine moderators and mediators of survey results. For example, you could use the average age of participants in the surveys to test the hypothesis that physical activity deceases with age: "Studies with an average participant age of Y revealed more-frequent activity than studies with an average participant age of Z."

It is rare to see this kind of descriptive research synthesis in the scholarly social and behavioral science literature. However, a similar procedure does appear on the nightly news during the weeks leading up to an election, when a news anchor will report the cumulative findings of numerous polls of voters asking about support for candidates or ballot issues.

Part of the problem with synthesizing descriptive statistics across the types of studies that appear in social science journals is that the studies often use different scales to operationalize the same variable. For example, it would be difficult to synthesize the levels of activity found in intervention studies because some studies might measure activity by giving participants a pedometer and counting their miles walked. Other studies might measure activity by gauging lung capacity. Measuring time spent on homework would produce less difficulty. Metrics for measuring time should be consistent across studies or easily convertible from one to another (e.g., hours to minutes). Measures of achievement would likely be difficult because sometimes it will be measured as unit tests, sometimes as end-of-year grades, and sometimes as scores on standardized achievement tests.[2]

Another problem with aggregating descriptive statistics is that it is rarely clear what population the resulting averages refer to. Unlike the polls that precede elections, social scientists writing for scholarly outlets often use convenience samples. While we might be able to identify the population (often very narrow) from which the participants of each study have been drawn, it is rarely possible to say what population an amalgamation of such convenience samples is drawn from.

Associational research.

A second type of descriptive research problem might be, "What events or phenomena happen together?" Here, researchers take their descriptions a step farther and ask whether variables co-occur, or correlate, with one another. Several instances of interest in co-occurrence appear in our synthesis examples. Our synthesis of correlates of attitudes toward rape focused exclusively on simple correlations between attitudes and other characteristics of respondents. The synthesis about homework also looked at simple

correlations between the amount of time spent on homework reported by students, and their achievement.

Causal research.

The third research problem seeks an explanation for the event: "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). What constitutes good evidence of causal production is a complex question that I will return to in Chapter 5. In practice, three types of research designs are used most often to help make causal inferences. I will call the first *modeling research.* It takes correlational research a step farther by examining co-occurrence in a multivariate framework (Kline, 2011). For example, the synthesis on homework looked at studies that built complex models using multiple regression, path analysis, or structural equation modeling to describe the co-occurrence of many variables, one being homework, and academic achievement.

The second approach to discovering causes is called *quasi-experimental research*. Here, unlike the modeling approach, the researcher (or some other external agent) controls the introduction of an intervention or event but does not control precisely who may be exposed to it (May, 2012). For example, in the synthesis on homework, some studies looked at groups of children whose teachers chose whether to give homework rather than having the experimenter randomly assign classes to conditions. Then the researchers might try to match children in different classes on preexisting differences.

A unique type of quasi-experiment (often called *preexperimental*) involves a pretest–posttest design in which participants serve as their own control by being compared on the outcome variable before and after the intervention is introduced. If these appear frequently in a research literature, it is important to remember that whereas such designs equate groups on lots of differences (after all, they are the same people), these studies' results are open to all sorts of alternative interpretations. These interpretations are related to the passage of time, including changes in participants that would have occurred regardless of the introduction of the intervention (would you expect children to get better at reading over the course of a year even without homework?), as well as other interventions or general historical events that happened during the time between the pretest and the posttest.

Finally, in experimental research, both the introduction of the event *and* who is exposed to it are controlled by the researchers (or other external agents), who then leave treatment assignment to chance (Christensen, 2012). This approach minimizes average preexisting differences between the assigned participants in each group so that we can be most confident that any differences between participants are caused by the variable that was manipulated. Of course, there are numerous other aspects of the design that we must attend to for a strong inference about causality to be made, but for our current purposes, this unique feature of experimental research will suffice, until Chapter 5.

In the synthesis example about choice and motivation, all the included studies involved an experimental manipulation of choice and the random assignment of participants to choice and no-choice conditions. Also, the research synthesis on aerobic exercise was purposely constrained to include only experimental studies.

Within-Participant or Between-Participant Processes?

Finally, the third question you must ask about the posited relationship is, "Does the problem or hypothesis seek to understand (a) how a process unfolds within an individual participant over time or (b) what is associated with, or explains variation between, participants or groups of participants?" All the designs I have introduced relate to the latter, the differences between participants on a characteristic of interest. The former problem—the problem of change within a participant—would best be studied using the various forms of single-case or time series designs, research designs in which single participants are tested at multiple times, typically at equal time intervals, during the course of a study. As with between-participants differences, within-participant processes can be studied using designs that are purely descriptive (simple time series), that reveal associations between two processes over time (concomitant time series), or that assess the causal impact of an intervention in the process (interrupted time series).

Syntheses of time series research are still rare and the methodology is still quite new, so the remainder of this book focuses on syntheses of between-participants research. All our synthesis examples involve research that attempted to discover relations involving variation between participants. Still, this makes it no less important to ask whether the research question concerns processes within participants or differences between participants and to understand that the answer will dictate what research designs and synthesis methods will be appropriate for answering the question. If you are interested in within-participant processes, you can consult Shadish and Rindskopf (2007) for a discussion of synthesis of single-case research.

Simple and Complex Relationships

The problems that motivate most research synthesists begin by posing questions about a simple two-variable relationship: Does choice affect motivation? Does homework cause improvements in achievement? The explanation for this is simple: Bivariate relationships have typically been tested more often than more-complex relationships. That said, it is rare, if not unheard of, for a synthesis to have only one operation of each of the two variables. For example, in the choice synthesis, four different outcome variables were collected that related to the participants' motivation to engage in the task (i.e., tasks engaged in during free time, enjoyment or liking of the task, interest in the task, willingness to engage in the task again) and were tested for whether the different measures revealed different results. In the aerobic exercise research synthesis dozens of outcome variables were measured and then classified into four larger domains of neurocognitive functioning for purposes of analysis: attention, executive functioning, working memory, and memory.

In fact, all the example syntheses examined potential influences on the bivariate relationships, as do almost all syntheses, including not just third variables created because of how variables were defined but also variations created by differences in how the study was carried out. These will include design variations (e.g., experiments compared to quasi-experiments) and implementation variations (e.g., setting, time).

Although some specific hypotheses about three-variable relationships—that is, interactions—in the social sciences have generated enough interest to suggest that a research synthesis would be informative, for the vast majority of topics the initial problem formulation will involve a two-variable question. Again, however, your initial undertaking of the synthesis to establish the existence of a bivariate relationship should in no way diminish the attention you pay to discovering interacting or moderating influences. Indeed, discovering that a two-variable relationship exists would quite often be viewed as a trivial contribution by the research community. However, if bivariate relationships are found to be moderated by third variables, these findings are viewed as a step forward and are given inferential priority. Even when an interaction is the primary focus of a synthesis, the search for higher-order interactions should continue. I will say more on the relationships between variables in Chapter 6, when I discuss how main effects and interactions are interpreted in research synthesis.

Summary

In sum then, in addition to asking whether your research synthesis has (a) provided clear conceptual definitions of the variables of interest and (b) included operations that are truly correspondent to those conceptual definitions, you must also ask,

> Is the problem stated so that the research designs and evidence needed to address it can be specified clearly?

Figure 2.1 summarizes the differences that can arise between research syntheses due to variations in how concepts are defined, operationalized, and related to one another. In the top portion of the figure we see that two synthesists might use conceptual definitions of different breadth. The definitions will affect how many operations will be deemed relevant to the concepts. So, a synthesist who defines homework as "academic work done outside school" will include more operations—for example, tutoring would fit this definition—than a synthesist who defines homework as "tasks assigned by teachers meant to be carried out during nonschool hours." Furthermore, it is also possible that regardless of the breadth of the concepts, the synthesists might differ concerning their decisions about whether certain operations are relevant. For example, one synthesist might include music and industrial arts grades as measures of achievement whereas another might not.

Figure 2.1 Differences Between Research Syntheses Due to Differences in Conceptual Definitions, Relevant Operations, and Variable Relationships
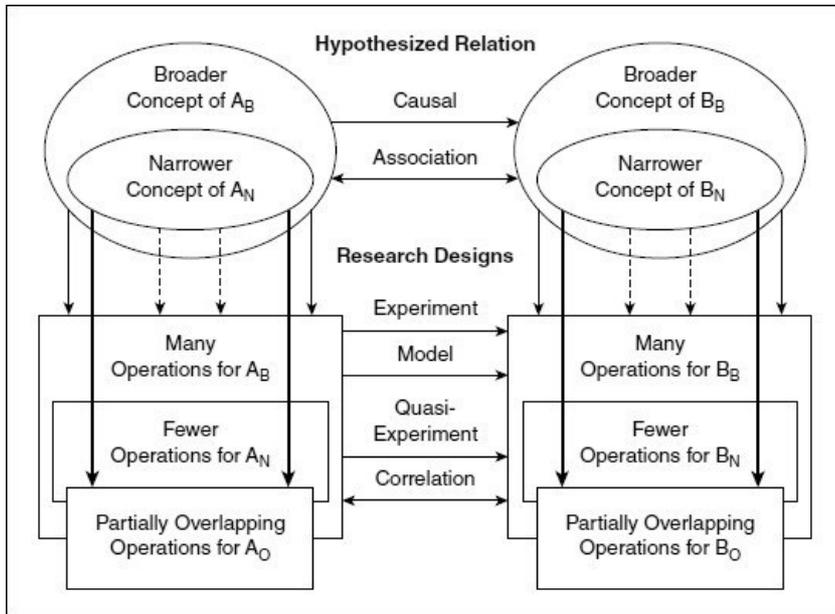
Also, the synthesists might differ in whether they are interested in research that studies an association or research that studies a causal link between the variables. This will influence the type of research designs that are deemed relevant and/or how the results of research using different designs are interpreted with regard to their ability to shed light on the relation of interest. So, synthesists who ask the question, "Is doing homework related to achievement?" would include both correlational and experimental research, while synthesists who ask the question, "Does homework cause improved achievement?" might restrict their synthesis to only experiments using random assignment to conditions and perhaps quasi-experiments. Or, if correlational research is included, it would need to be carefully interpreted as less than optimal for answering this question (a concern we will return to in Chapter 5).

And finally, it is important to remember that some variables in a synthesis can be relatively narrowly defined, whereas others are broadly defined. For example, in our synthesis concerning attitudes toward rape, the term *rape* was defined relatively narrowly as sexual intercourse between a man and a woman without the woman's consent. Still, our literature search uncovered 17 different measures of rape attitudes, but only 5 were used with much frequency (e.g., Attitude Toward Rape Scale, Rape Myth Acceptance). On the other hand, the concept used to define predictors of rape attitudes, "individual differences," was extremely broad. We identified 74 distinct individual difference variables that could be clustered into broader groupings (but narrower than "individual differences") consisting of demographic, cognitive, experiential, affective, and personality measures. As noted previously, much of the creative challenge and reward in doing research synthesis lies in identifying groupings like these and making sense of their different relationships to other variables.

Judging the Conceptual Relevance of Studies

It can always be the case that researchers disagree about the conceptual definition of a variable or about the operations relevant to it. In fact, many disputes surrounding research syntheses revolve around differences in what studies were included and excluded based on their relevance. Readers who are knowledgeable about the research area will say, "Hey, how come this study wasn't included?" or "How come this study was?" For example, many homework scholars would have objected if our research synthesis included studies that involved students receiving tutoring at the recommendation of the teacher even though including them might have met a broad conceptual definition of homework. Likely, had tutoring studies been included, these scholars would have suggested that the definition of homework, as most people understand it, involves assignments given to the entire class of students. They would have argued that the definition of homework needed to be more precise.

Beyond the breadth or narrowness of the conceptual definition, some research has examined other contextual factors that might affect whether a study is deemed relevant to a research problem. For example, judgments about the relevance of studies to a literature search appear to be related to the searcher's open-mindedness and expertise in the area (Davidson, 1977), whether the decision is based on titles or abstracts (Cooper & Ribble, 1989), and even the amount of time the searcher has for making relevance decisions (Cuadra & Katter, 1967). Thus, while the conceptual definition and level of abstractness that synthesists choose for a problem are certainly two influences on which studies are deemed relevant, a multitude of other factors also affect this screening of studies.

You should begin your literature search with the broadest conceptual definition in mind. In determining the acceptability of operations for inclusion within the broad concept, you should remain as open in your

interpretation as possible. At later stages of the synthesis, notably during data evaluation, it is possible to exclude particular operations due to their lack of relevance. However, *in the problem formulation and literature search stages, decisions about the relevance of studies should err on the side of being overly inclusive,* just as primary researchers collect some data that might not later be used in analysis. It is very distressing to find out *after* studies have been retrieved and catalogued that available pieces of the puzzle were passed over and a new search must be conducted. An initially broad conceptual search also will help you think more carefully about the boundaries of your concepts, leading to a more precise definition once the search is completed. So, if studies of tutoring are retrieved because an expansive interpretation of the concept of homework is used ("academic tasks carried out during nonschool hours"), and it is later decided that these ought not be included, it could lead to a refinement of the definition ("tasks assigned by teachers to all students").

It is also good practice to *have the initial decision about the potential relevance of a study, sometimes called initial or prescreening, made by more than one person.* Here, you give the screeners the conceptual definition of variables and examples of relevant operations and have them examine the documents retrieved by your literature search. The purpose of having multiple screeners is not only to see if the conceptual definitions lead to agreement among screeners, but also to flag for further screening any study that is deemed potentially relevant by any one screener. Often, the initial decision about relevance will be made on limited information about the study, such as the study's abstract. When this is the case, it is even more important to have at least two screeners judge each study and take a second look at studies even if just one screener thought it might be relevant to do so.

Table 2.1 provides an example of a screening sheet for coders to use to report their initial decision about whether a document is relevant to a search. The most critical code is the seventh which places each document into one of four categories depending on what the screener thinks it contains. Note that in addition to categories that identify documents as possibly containing data relevant to the search, the initial screening question includes a category for documents that might not include data for a meta-analysis but that might provide other important information or insights about the topic, perhaps for use in the introduction or discussion of the synthesis results. For example an article that does not contain empirical evidence but does include suggestions about possible influences on the impact of the intervention on adult activity would by classified as a background article.

Table 2.1   Initial Screening Coding Guide

Initial Screening for Relevance	
1. What is the report ID number?	___ ___ ___
2. What is the screener's name?	_____
3. What is the date of this screening?	__/__/____
4. What is the first author's last name?	_____
5. In what year did the document appear?	_____

Initial Screening for Relevance	

6. What type of document is this?
   a. Journal article
   b. Book or book chapter
   c. Dissertation
   d. MA thesis
   e. Private report
   f. Government report
   g. Conference paper
   h. Other (specify) _____
   i. Cannot tell
	_____

7. What type of information is contained in this document?
   a. Background
   b. Empirical evidence
   c. Both
   d. This document is irrelevant
	_____

8. If empirical, what type of empirical evidence does this document contain?
   1 = Descriptive
   2 = Association or experimental
   3 = Both
   4 = Other (specify) _____
	_____

9. If background, what type of background information does this document contain?
(Place a 1 for each item that applies, 0 for each item that does not apply)
   a. Descriptions of program variations   _____
   b. Issues in program implementation   _____
   c. Arguments for and/or against   _____
   d. Review of previous research   _____
   e. Other(specify)_____ _____	

The rest of the information on the sheet relates to characteristics of the document and its producers. This information is typically found in the document records contained in most computerized reference databases, so it typically is not necessary for the screener to examine the full document to find it. Some of these codes might be used to make decisions about whether to include a study. For example, the year of the report might be used if the decision is made to limit the synthesis only to studies appearing after a certain date. When a literature search requires the screening of large numbers of documents (a search of the ERIC database for the mention of the term *homework* reveals more than 2,700 documents since 1996) the initial screening will occur at this level. And, of course, the questions in an initial screening might be altered depending on their relevance to a particular search.

Study-Generated and Synthesis-Generated Evidence

I have pointed out that most research syntheses focus on main-effect questions but then also test for moderators by grouping studies according to differences in the way the research was carried out. In essence, then, these moderator analyses are testing for interaction effects—that is, they ask whether the main-effect relationship is different depending on the level or categories of a third variable, in this case a characteristic of the study. This leads us to consider an important distinction between the types of evidence contained in research syntheses.

There are two different sources of evidence about relationships contained in research syntheses. The first type is called *study-generated evidence.* Study-generated evidence is present when a single study

contains results that directly test the relation being considered. Research syntheses also contain evidence that does not come from individual studies, but rather 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 to one another.

There is one critical distinction between study-generated and synthesis-generated evidence: *Only study-generated evidence based on experimental research allows the synthesist to make statements concerning causality.* An example will clarify the point. With regard to choice and motivation studies, suppose we were interested in whether the number of options a participant is given to choose among influences the effect of choice on motivation. Suppose also that 16 studies were found that directly assessed the impact of number of options by randomly assigning participants to experimental conditions, one in which participants chose between only two alternatives and another in which more than two alternatives were available. The accumulated results of these studies could then be interpreted as supporting or not supporting the idea that the number of choice options *causes* increases or decreases in motivation. Now, assume instead that we uncovered eight studies that compared only a two-option choice condition to a no-choice control group, and eight other studies that compared a multiple-option (more than two) choice condition to a no-choice control group. If this synthesis-generated evidence revealed larger effects of choice on motivation when more (or fewer) options were given, then we could infer an association but not a causal relation between the number of options and motivation.

Why is this the case? Causal direction is not the problem with synthesis-generated evidence. It would be foolish to argue that the amount of motivation exhibited by participants caused the experimenters' decision about the number of options. However, still problematic is another ingredient of causality—the absence of potential third variables causing the relationship. A multitude of third variables are potentially confounded with the original experimenters' decisions about how many choice options to give participants. For instance, the participants in multiple-option studies may have been more likely to be adults while two-option studies were more likely to be conducted with children. Age might be the true cause (could children be thrilled to get choices while adults are unmoved by them?).

Synthesis-generated evidence cannot legitimately rule out as possible true causes any other variables confounded with the study characteristic of interest. This is because the synthesists did not randomly assign the number of choice options to experiments. It is the ability to employ random assignment of participants that allows primary researchers to assume that third variables are represented equally in the experimental conditions. So, a synthesis encompassing studies that all compared varying choice-option conditions to a no-choice control group can make causal statements about the effect of choice per se but not about the effect of the number of options on the effect of choice. Here, an association can only be claimed.

Summary

It is important for synthesists to 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, the lesser strength of synthesis-generated evidence with regard to causal inferences does not mean this evidence should be ignored. The use of synthesis-generated evidence allows you to test relations that may have never been examined by primary researchers. For example, it may be the case that no previous primary study has examined whether the relation between homework and achievement is different for assignments of different length, or whether different types of aerobic interventions differ in their effects on subsequent cognitive functioning. By searching across studies for variations in assignment length or intervention type and then relating this to the effect of homework on achievement or interventions on memory synthesists can produce the first evidence about these potentially critical moderating variables. Even though this evidence is equivocal, it is a major contribution of research synthesis and a source of potential hypotheses for future primary research.

Arguing for the Value of the Synthesis

All research syntheses should be placed in a theoretical, historical, and/or practical context. Why are attitudes toward rape important? Do theories predict how and why particular individual differences will relate to rape attitudes? Are there conflicting predictions associated with different theories? Why do older adults need aerobic activity? Where did the idea for activity interventions come from? Are intervention components grounded in theory or in practical experience? Are there debates surrounding the utility of exercise programs?

Contextualizing the problem of a research synthesis does more than explain why a topic is important. Providing a context for the problem also provides the rationale for the search for moderators of the overall findings. It is an important aid in identifying variables that you might examine for their influence on outcomes. For example, self-determination theory proposes that having a choice will improve intrinsic motivation to engage in a task but providing rewards will undermine future task motivation. This suggests

that studies of choice that also provide rewards might produce different results from studies with no rewards. The presence of rewards, then, should be examined as a potential moderator of the overall relationship.

Also, many social interventions, such as assigning homework, have claims associated with them that suggest they will influence more than one outcome variable. For example, homework proponents provide a list of claimed positive effects that include academic (e.g., improved study skills) and nonacademic outcomes (e.g., better time management). Likewise, homework opponents provide their own list of possible negative effects (e.g., less time for other activities that promote positive life skills). It is important that research synthesists examining the effects of an intervention provide a list of possible intervention effects, both positive and negative, that have been proposed as outcomes. These effects might have been offered by theorists, researchers, practitioners, and pundits.

Again, both quantitative and qualitative research can be used to place the research problem in a meaningful context. Narrative or qualitative descriptions of relevant events can be used to discover the salient features of the problem at hand. These can be the source of important queries for research synthesists to ask of the quantitative evidence. Quantitative surveys also can answer specific questions across a broader array of problem instantiations. In addition to establishing the importance of the problem, surveys can answer questions such as, "How available are aerobic exercise intervention programs for adults?" and "What are the characteristics of participants in these intervention programs?"

If a Synthesis Already Exists, Why Is a New One Needed?

Sometimes, the value of a synthesis is easy to establish: A lot of past research has been conducted and it is yet to be accumulated, summarized, and integrated. However, if a topic has a long history of research, it is not surprising to find that previous attempts to summarize it already exist. Obviously, these efforts need to be scrutinized carefully before the new synthesis is undertaken. Past syntheses can help establish the necessity for a new synthesis. This assessment process is much like that used in primary research before undertaking a new study.

There are several things you can look for in past syntheses that will help your new effort. First, previous syntheses can be used, along with the other background documents you find, to identify the positions of other scholars in the field. In particular, the past syntheses can be used to determine whether conflicting conclusions exist about what the evidence says and, perhaps, what has caused the conflict.

Second, an examination of past syntheses can assess the earlier efforts' completeness and validity. For example, the synthesis on aerobic exercise interventions found one narrative review and four meta-analyses of past research. However, these past efforts disagreed about the magnitude of improvement on neurocognitive functioning that resulted from the interventions.

Past syntheses also can be an important aid in identifying interacting variables that you might wish to examine. Rather than restart the compilation of potential moderating variables, previous synthesists (along with primary researchers, both quantitative and qualitative) will undoubtedly offer many suggestions based on their own research and reading of the literature. If more than one synthesis of an area has been conducted, the new effort will be able to incorporate all the suggestions.

Finally, past syntheses allow you to begin the compilation of a relevant bibliography. Most syntheses will have fairly lengthy bibliographies. If more than one synthesis exists, their citations will overlap somewhat, but not perfectly. Along with other techniques described in the next chapter, the research cited in past syntheses provides an excellent place for you to start the literature search.

The Effects of Context on Synthesis Outcomes

Differences in how a problem is placed in its theoretical or practical context affects the outcomes of syntheses by leading to differences in the way study operations are treated *after* the relevant literature has been identified. Synthesists can vary in the attention they pay to theoretical and practical distinctions in the literature. Thus, two research syntheses conducted using identical conceptual definitions and the same set of studies can still reach decidedly different conclusions if one synthesis examined information about theoretical and practical distinctions in studies to uncover moderating relationships that the other synthesis did not examine. For example, one synthesis might discover that the effect of homework on achievement was associated with the grade level of students, whereas another synthesis never addresses the question. Thus, to evaluate whether (a) the importance of the problem has been established and (b) a list of important potential moderators of findings has been identified, the next question to ask about your research syntheses is,

Is the problem placed in a meaningful theoretical,historical, and/or practical context?

Notes

1. Here, I use the term *participant* in the broader sense: the participant might be an individual person or animal, or a group of such units. For ease of exposition, I will continue to use the term *participant* in place of the more cumbersome term *units under study.*

2. The problem of nonstandard measurements is lessened when study characteristics are tested as *third* variables because the bivariate relationships *within* the studies can be transformed into standardized effect size estimates, thus controlling for different scales (see Chapter 6).

3 Step 2 Searching the Literature

What procedures should be used to find relevant research?

<div style="border:1px solid #000; background:#e0e0e0; padding:1em;">

Primary Functions Served in the Synthesis

1. To identify places to find relevant research (e.g., reference databases, journals)
2. To identify terms used to search for relevant research in reference databases

Procedural Variation That Might Produce Differences in Conclusions

1. Variation in searched sources might lead to systematic differences in the retrieved research.

Questions to Ask When Evaluating the Literature Search in a Research Synthesis

1. Were complementary searching strategies used to find relevant studies?
2. Were proper and exhaustive terms used in searches and queries of reference databases and research registries?

This chapter describes

- Objectives of a literature search
- Methods for locating studies relevant to a synthesis topic
- Researcher-to-researcher, quality-controlled, and secondary channels for obtaining research reports
- How research enters different channels
- How searchers access different channels
- What biases may be present in the kinds of information contained in different channels
- Problems encountered in retrieving studies

</div>

In primary social science research, participants are recruited into studies through subject pools, advertisements, Internet websites, schools, doctors' offices, and so on. In research synthesis, the studies of interest are found by conducting a search for reports describing past relevant research. Regardless of whether social scientists are collecting new data or synthesizing results of previous studies, the major decision they make when finding relevant sources of data involves defining the target population that will be the referent of the research (Fowler, 2014). In primary research, the target population includes those individuals or groups that the researcher hopes to represent in the study. In research synthesis, the target population includes all the studies that test the hypothesis or address the problem.

The sample frame of an investigation in the case of primary research includes those individuals or groups the researcher pragmatically could obtain. In the case of research synthesis, it includes obtainable study reports. In most instances, researchers will not be able to access all of a target population's elements. To do so would be too costly because some people (or documents) are hard to find or refuse to cooperate.

Population Distinctions in Social Science Research

Both primary research and research synthesis involve specifying target populations and sampling frames. In addition, both types of investigation require the researcher to consider how the target population and sampling frame may differ from one another. The trustworthiness of any claims about the target population will be compromised if the elements in the sampling frame differ in systematic ways from the target population. Because it is easier to alter the target of an investigation than it is to locate hard-to-find people or studies, both primary researchers and research synthesists may find they need to respecify their target population when an inquiry nears completion.

The most general target population for social and behavioral science research could be characterized roughly as "all human beings," either as individuals or in groups. Most topics, of course, delineate the elements to be less ambitious, such as "all students" in a study of the effects of homework or "all adults over 50 years of age" in a study of the effects of exercise interventions.

Sampling frames in social and behavioral science research typically are much more restricted than targets. So, participants in an exercise intervention might all be drawn from a similar geographic area. Most researchers are aware of the gap between the diversity of participants they hope their research results refer to and those people actually available to them. For this reason, they discuss limits on generalizability in their discussion of the study's results.

As I noted in Chapter 1, research syntheses involve two targets. First, *synthesists hope their work will cover all previous research on the problem*. Synthesists can exert some control over reaching this goal by

how they conduct their literature search—that is, through their choices of information sources. How this is done is the focus of this chapter. Just as different sampling methods in primary research can lead to differences in who is sampled (e.g., phone surveys and mail surveys reach different people), different literature-searching techniques lead to different samples of studies. Likewise, just as it is more difficult to find and sample some people than others, it is also more difficult to find some studies than others.

In addition to wanting to cover all previous research, *synthesists also want the results of their work to pertain to the target population of people (or other units) that are relevant to the topic.* When we conducted our synthesis of homework research, we hoped that students at grade levels kindergarten through 12, not just high school students, for example, would be represented in past studies. Our ability to meet our goal was constrained by the types of students sampled by primary researchers. If first and second graders were not included in previous homework studies, they will not be represented in a synthesis of homework research. Thus, research synthesis involves a process of sampling samples. The primary research includes samples of individuals or groups, and the synthesist retrieves primary research. This process is something akin to cluster sampling, with the clusters distinguishing people according to the research projects in which they participated.

Also different from primary research, synthesists typically are not trying to draw representative samples of studies from the literature. Generally, they attempt to retrieve an entire population of studies. The formidable goal of finding all studies is rarely achieved, but it is certainly the desired objective.

Methods for Locating Studies

How do you go about finding studies relevant to a topic? There are numerous techniques scientists use to share information with one another. These techniques have undergone enormous changes in recent years. In fact, it is safe to say that the ways scientists transmit their work to one another has changed more in the past three decades than it did in the preceding three centuries, dating back to the late 17th century, when scholarly journals first appeared. The change is primarily due to the use of computers and the Internet to facilitate human communication.

The Fate of Studies From Initiation to Publication

A description of the many mechanisms that searchers can use to find studies will be most instructive if we begin with an account of the alternative possible fates of studies once they have been proposed. My colleagues and I (Cooper, DeNeve, & Charlton, 1997) conducted a survey of 33 researchers who had several years earlier proposed 159 studies to their university's institutional review board. The survey asked the researchers how far along each of the studies had gone in the process from initiation to publication. Figure 3.1 summarizes their responses. Of the 159 studies, 4 were never begun, 4 were begun but data collection was never completed, and 30 were completed but the data were never analyzed. From the point of view of research synthesists, these 38 studies are of little interest because a hypothesis was never tested. However, once a study's data have been analyzed (as happened for about 76% of the proposed studies) then the result is of interest because it represents a test of the study's hypotheses. Not only does the study now include information on the truth or falsity of the hypothesis but what happens to the study next may be influenced by what the data revealed. For example, Figure 3.1 indicates that about 13% of studies with analyzed data produced no written report; the researchers gave several reasons why this was the case. Some of these reasons seem related to the outcome itself, especially the reason that the results were not interesting and/or not statistically significant. This means uninteresting and nonsignificant results may be harder to find. Next, we see in Figure 3.1 that only about half of the written summaries of research were prepared for a journal article, book chapter, or book. And finally, of these somewhere between 75% and 84% eventually found their way into print.

As we examine the different retrieval techniques used by people searching for studies, it will be important to keep in mind that *the difficulty in finding research, and the value of different searching techniques, will be a function of how far along the study went—or currently is, for recently completed work—in the process from data analysis to publication,* as outlined in Figure 3.1. For example, to anticipate the discussion that follows, it is clear that studies that had data analyzed but never were written up will be retrievable only through direct contact with the researchers. Studies that appear in journals will be easier to find, but may overrepresent significant and/or novel findings.

Figure 3.1 Flow Diagram of the Fate of Research From Institutional Review Board Approval to Publication

97.5% of the Total — 97.5% Ran Subjects (n = 155) / 2.5% Did Not Run Subjects (n = 4)

95% of the Total — 97.4% Completed Running Subjects (n = 151) / 2.6% Did Not Complete Running Subjects (n = 14)

76.1% of the Total — 80.1% Analyzed the Data (n = 121) / 19.9% Did Not Analyze the Data (n = 30)

66% of the Total — 86.8% Produced Written Summaries (n = 105) / 13.2% Did Not Produce a Written Summary (n = 16)

34.6% of the Total — 52.4% Prepared Study for a Journal Article or Book (n = 55) / 47.6% Did Not Prepare the Study for Publication (n = 50)

5.5% Resulted in Book Publications (n = 3) / 9.1% Are Under Revision or are Intended for Submission (n = 5)

23.9% of the Total — 69.1% Were Accepted Into a Journal (n = 38) / 16.4% Were Not Accepted Into a Journal (n = 9)

Study Begins With IRB Approval

10 studies were lost due to researchers who could not be found and 9 were lost due to researchers who declined to participate

Why no subjects were run
75% No Funding
25% Not Enough Time
25% Student Investigator Left

Why running of subjects was not completed
50% Design or Operational Problems
25% Design Changes
25% Initial Results Were Not Interesting
25% No Significant Results
25% Assistant Left
25% Assistant Left
25% Researcher Did Not Recall

Why the data was not analyzed
90% Researcher Did Not Recall
3.3% Design or Operational Problems
3.3% Assistant Left
3.3% Student Lost Interest

Why a written summary was not produced
50% Design or Operational Problems
50% Results Were Not Interesting
31.3% No Significant Results
37.5% Publication Not an Aim
12.5% Assistant Lost Interest
6.7% Purpose Was to Design Materials
6.7% Study Was Only Exploratory
6.3% Others Lost Interest

Why the study was not prepared for a journal article
48% Publication Not an Aim
30% Class Project Only
26% Assistant Lost Interest
22% No Significant Results
20% Results Were Not Interesting
12% Design or Operational Problems
6% Researcher Did Not Recall
2% Others Lost Interest

Why the article was not accepted into a journal
66.7% Rejected by a Journal
33.3% Design or Operational Problems
33.3% Assistant Lost Interest
11.1% Didn't Fit With Journal
11.1% Needed Design Change
11.1% No Significant Results
11.1% Others Lost Interest
11.1% Results Were Not Interesting

SOURCE: From "Finding the Missing Science: The Fate of Studies Submitted for Review by a Human Subjects Committee," by H. Cooper, K. DeNeve, & K. Charlton, 1997, Psychological Methods, 2, pp. 448–449. Copyright 2001 by the American Psychological Association.

Some Ways Searching Channels Differ

The section that follows will present descriptions of the major techniques you can use to find research. I will attempt to evaluate the kind of information found using each technique by comparing search results that used it exclusively to that of the target population "all relevant research," or, put differently, "all relevant studies for which data were analyzed." Regrettably, there are only limited empirical data on differences in scientific information obtained using different search techniques, so many of my comparisons will involve some speculation on my part. The problem is complicated further by the fact that the effect of a searching technique's characteristics on its outcomes probably varies from topic to topic.

Also, the proliferation of ways to share information makes it increasingly difficult to find just a few descriptors that help us think about how the search techniques differ and relate to one another. Mechanisms for communication have arisen in a haphazard fashion, so no descriptive dimension perfectly captures all their important features. Still, there are several features that are useful in describing the different search techniques. One important feature that distinguishes scientific communication techniques relates to *how research gets into the channel*. Channels can have relatively open or restricted rules for entry. Open entry permits the primary researcher (the person who wants to put something in the channel) to enter the channel directly and place his or her work into its collection of information. Restricted entry requires primary researchers to meet the requirements of a third party—some person or entity between the researcher and the person searching for information—before their work can be included. The most important of these requirements is the use of peer review in scientific journals to ensure that research meets certain standards of relevance, quality, and importance. In fact, all channels have some restrictions on entries, but the type and stringency differ from channel to channel. It is these restrictions that most directly affect how the research in the channel differs from all relevant research.

A second important feature of search techniques concerns *how searchers obtain information from the channel*. Channels have more or less open or restricted requirements regarding how to access their content. A channel is more restricted if it requires the searcher (the person seeking information from the channel) to identify very specifically what or whose documents they want. A channel is more open if the searchers can be more broad or general in their request for information. These access requirements also can influence the type of research a searcher will find in a channel.

The importance of these distinctions will become clear as I describe how they relate to specific search techniques. For purposes of exposition, I have grouped the techniques under the headings "Researcher-to-Researcher Channels," "Quality-Controlled Channels," and "Secondary Channels."

Researcher-to-Researcher Channels

Researcher-to-researcher techniques for obtaining study reports are characterized by the fact that searchers are attempting to locate investigators who may or may not have relevant studies rather than to locate the reports themselves. There are no formal restrictions on the kinds of requests that can be made through such contact or who can exchange information. The request to the researcher can be very general (e.g., "Have you conducted or are you aware of any studies involving aerobic exercise?") or very specific (e.g., "Have you conducted or are you aware of any studies involving aerobic exercise as interventions on older adults that measured cognitive performance?"). In all but one case, there is no third party that mediates the exchange of information between the searcher and researcher. The principal forms of researcher-to-researcher communication involve personal contacts, mass solicitations, traditional invisible colleges, and electronic invisible colleges. The distinctions between these forms of communication are described in the following paragraphs and summarized in Table 3.1.

Table 3.1 Researcher-to-Researcher Channels for Locating Studies

Channel	Restrictions on How Research Gets In	Restrictions on How Searcher Gets In	Restrictions on Types of Information
Personal contact	Researcher must be known to searcher.	Searcher must know how to contact researcher.	Studies may be relatively homogeneous in methods and results.
Mass solicitation	Researcher must hold status known to searcher (e.g., member of a relevant organization).	Searcher must have mechanism for contacting researchers based on their status.	Studies may be homogeneous in methods and results, if solicitation is based on membership in organization with particular bias.
Traditional invisible college	Research must be "approved" by prominent researchers.	Searcher must know who prominent researchers are and how to contact them.	Studies may be homogeneous in methods. Results may be overly consistent with perspective of prominent researchers.
Electronic invisible college	Researcher must subscribe to distribution list or other electronic distribution mechanism (e.g., Facebook page).	Searcher must subscribe to the same distribution list or make a request to someone who is a subscriber.	Studies may be homogeneous in methods and results if the distribution list is based on membership in organizations with particular bias.

Personal Contact

The first information available to searchers is, of course, their own research. Before anyone else sees research results, the primary investigators see it themselves. So, we began our search for studies about the effects of homework on achievement by including our own studies that were relevant to the issue.

Although this source may seem almost too obvious to mention, it is a critical one. It is important for research synthesists to keep the role of their own work in proper perspective. Primary research that synthesists personally have conducted has a strong impact on how they interpret the research literature

as a whole (Cooper, 1986). Typically, we expect that all research should come to the same conclusions as our studies. However, any researcher's own studies on a topic could differ markedly on a number of important dimensions from other research, with many of the differences in how research was conducted, possibly influencing results. Each researcher is likely to repeat some of the same operations across studies, using only a few measurement devices and/or instructions to participants. For example, studies of homework that one researcher conducts might exclusively use students' class grades as the measure of achievement. Other researchers might use textbook unit tests or standardized tests but not class grades. Also, participants in one researcher's studies might be drawn from the same institutions (e.g., a researcher always uses students in a nearby school district) and geographical area. This makes participants homogeneous on some dimensions (e.g., SES) and different from participants in other researchers' studies. Even research assistants will be more homogeneous within the same laboratory in potentially relevant ways (e.g., how well they are trained) than a random sample of all research assistants working on studies related to the topic.

Other one-on-one contacts—that is, people you contact directly or who contact you to share their work because they know the things you are interested in—take you outside your own laboratory, but perhaps not far outside. Students and their professors share ideas and pass on to one another papers and articles they find that are of mutual interest. Colleagues who have collaborated in the past or have met and exchanged ideas previously also will let one another know when new studies become available. A colleague down the hall might run across an article in a journal or conference program and, knowing of your interest in the topic, might pass it on to you. Occasionally, readers of a researcher's past work will point out literature they think is relevant to the topic but is not cited in the report. This sometimes happens after the research report appears in print, but also can happen as part of the manuscript review process. It would not be uncommon for a peer reviewer of a homework manuscript I submitted for journal publication to suggest some additional relevant articles that were not referenced in my work. These would be added to our list of relevant research as we begin our homework synthesis.

Limitations of information obtained by personal contact.

Personal contact is generally a restricted communication channel. A searcher must know of and individually contact the primary researchers to obtain relevant information. Or the primary researchers must know the searcher is interested in what they do in order to initiate the exchange of information. So, much like a researcher's own work, information found through personal contacts, be they friends or colleagues, generally will reflect the methodological and theoretical biases of the searcher's informal social system. It most likely will be more homogeneous in findings than "all relevant research." That is not to say that personal contacts will rarely reveal to searchers findings that are inconsistent with their expectations. However, personal contacts are less likely to result in inconsistencies than they are to reveal research that confirms expectations (and looks like the kind of research the colleagues do). Therefore, *personal contacts with friends and colleagues must never be the sole source of studies in a research synthesis*. Research synthesists who rely solely on these techniques to collect relevant work are acting much like surveyors who decide to sample only their friends. That said, Figure 3.1 also suggests that these personal contacts may be the only way to obtain studies in which the data were analyzed but never resulted in a written research report.

Mass Solicitations

Sending a common solicitation to a group of researchers can produce less-biased samples of information. These contacts require that you first identify groups whose individual members might have access to relevant research reports. Then, you obtain lists of group members and contact the members individually —typically by e-mail—even if you do not know them personally. For example, for our homework search we contacted the dean, associate dean, or chair of 77 colleges, schools, or departments of education at institutions of higher education. We asked them to transmit to their faculty our request that they share with us any research they had conducted or knew of that related to the practice of assigning homework.

When you write an e-mail to a group of mostly strangers to ask for help it is important that your message be short, courteous, and transparent. It should say:

- Who you are.
- What you are studying. Be general but not too broad. For instance, do not say "studies on motivation" but rather "studies on the effects of choice on motivation." A very broad request will lead to nonresponse. A very narrow request will lead responders to think something is irrelevant when it is.
- Why you need this information (you are doing a literature search and want to be as exhaustive as possible).
- That you are willing to reimburse them for any expenses.
- That you will share with responders the final report of your project, regardless of whether they have relevant reports.
- A sincere "thanks in advance."

My experience suggests that while the hit rate for mass mailing is generally low, those who do respond are very interested and often provide material that is not yet publicly available. It is also a good way for you to introduce yourself to people who might share your interests. One indirect benefit is that you could make some new professional contacts.

Limitations of information obtained through mass solicitation.

Mass solicitations can reveal a more heterogeneous sample of studies than personal contacts depending on the technique used to generate the mailing list. For example, it is hard to see how our strategy of contacting deans, associate deans, and department chairs would lead to a terribly biased sample of studies (although we did not know exactly which deans actually forwarded our e-mail, and this might have been related to what types of information they thought we would be "happiest" to receive). With regard to Figure 3.1, I suspect that information on studies that were stopped after the data were analyzed but no report was written is less likely to be retrieved by mass mailing than it is by personal contact. In mass mailings, the searcher is less likely to be known to the recipient of the solicitation.

Traditional Invisible Colleges

Another channel of direct communication, a bit less restrictive than personal contacts, is called the *invisible college.* According to Crane (1969), invisible colleges are formed because "scientists working on similar problems are usually aware of each other and in some cases attempt to systematize their contacts by exchanging reprints with one another" (p. 335). Through a sociometric analysis, Crane found that most members of invisible colleges were not directly linked to one another but were linked to a small group of highly influential members. In terms of group communication, traditional invisible colleges are structured like wheels: influential researchers are at the hub and less-established researchers are on the rim, with lines of communication running mostly between the hub and the rim, and less often between or among members along the rim.

The structural characteristics of the traditional invisible college are dependent on the fact that in the past the informal transmission of information between scientists occurred one on one, primarily through printed mail and by telephone. These two mediums required that only two people at a time could exchange information (though multiple two-way communications might occur in parallel through, say, mass mailings). Also, the two communicators had to know and choose to talk to one another. Thus, influential researchers acted as hubs, both restricting the input (entry) and directing the output of (access to) information to a group of researchers known to them.

Today, traditional invisible colleges still exist but they have lessened in importance because of the ease and speed with which researchers can communicate with one another. For example, for our homework search, we sent similar e-mails to 21 scholars who our reference database search (discussed in a following section) revealed had been the first author on two or more articles on homework and academic achievement between 1987 and the end of 2003. Among these 21 researchers, there were about a half dozen we already knew were active homework researchers. So, you might say that our decision to identify homework researchers by finding those who had multiple publications in recent years was a strategy to find people likely to be the hubs of homework wheels. Prominent researchers who publish frequently in an area are likely to get contacted more often than researchers just starting out. Our requests to these hubs were not only that they send us their research, but also that they send us other research they were aware of and to suggest other researchers we should contact.

Limitations of information obtained through traditional invisible colleges.

The influence of prominent researchers over the information communicated through traditional invisible colleges holds the key to assessing the biases in the information transmitted through this channel. Synthesists gathering research solely by contacting prominent researchers will probably find studies that are more uniformly supportive of the beliefs held by these central researchers than are studies gathered from all sources. This is because new or marginal researchers who produce a result in conflict with that of the hub of an invisible college would be less likely to try to enter their work into this channel. If the disconfirming researchers do try to enter the invisible college, they are less likely to see their work widely disseminated throughout the network. Disconfirming findings may lead a researcher already active in an invisible college to leave the network. Also, because the participants in a traditional invisible college use one another as a reference group, it is likely that the kinds of operations and measurements used in their research will be more homogeneous than those used by all researchers who might be interested in a given topic.

Electronic Invisible Colleges

While traditional invisible colleges still exist today, there exists also a newer type of invisible college. This is really a hybrid of the invisible college and mass mailings. With the Internet, the need has diminished for communication hubs that hold together groups of scientists interested in the same topic. Instead, the

Internet does it for the group. The Internet allows searchers to send the same information request simultaneously to a group whose members share an interest but may be largely unknown to one another.

Electronic invisible colleges operate through the use of computerized list management programs. These programs maintain mailing lists and automatically send e-mail messages to members. So, for our homework synthesis, we identified a group called *the National Association of Test Directors,* composed of the directors of research or evaluation in over 100 school districts. We contacted the manager of the distribution list and asked this person to send our request for studies to members. If you are a member of an organization that is relevant to your topic, you may be able to make this request directly of the other members.

Sometimes groups of researchers may not be associated with a formal distribution list maintained by an organization but rather communicate through informal lists. In other instances, researchers may be members of a growing number of Internet vehicles that allow like-minded individuals to share information. These include electronic bulletin boards or discussion groups, Facebook, LinkedIn, ResearchGate, and the e-mail lists on which members hold electronic discussions by submitting topics or questions and receiving comments from other subscribers. Any of these can be used to make requests for research reports.

How do literature searchers know what electronic invisible colleges are out there? The best way to find these is to do an Internet search including keywords about your area of interest and the terms *discussion group, electronic bulletin board,* or *e-mail list* and descriptors of the topic of interest. Lists also can be found by visiting Internet sites of research organizations.[1] Many organizations now support special interest groups that bring together researchers with common interests, further blurring the line between mass correspondence and invisible colleges.

Limitations of information obtained through electronic invisible colleges.

A large majority of subscribers to distribution lists or discussion groups who receive messages asking for help in identifying studies relevant to a particular topic probably could not help you find studies and will not respond. But if even a few do know of studies, this can be very helpful. Especially, these channels can help you locate new research—perhaps in report form but not yet submitted for publication or in the publication queue but not yet published—or old research that never made its way into another communication channel.

Electronic invisible colleges, unless they are associated with stable organizations, can be temporary, informal entities that often deal with special problems. They can vanish when the problem is solved or the focus of the discipline shifts. They can become out of date by including researchers whose interests have moved on from the topic. They can exclude new researchers who have recently entered the field and do not yet know of the invisible college's existence. That is why *it is good practice to use electronic invisible colleges along with the more direct personal contacts* described previously.

Electronic distribution lists can be less restrictive than a traditional invisible college because while an individual may act as the list coordinator (the hub) many lists are not moderated by individuals at all. Instead, the computer often acts as the hub of the communication wheel. It disseminates the communications that come to it without imposing any restrictions on content. In moderated mailing lists, the list of members can be held privately and admittance and/or content may be screened, so these can function more like traditional invisible colleges.

Anyone can join many distribution lists, once they know that the list exists, by sending a simple command to the list's host computer. Other lists require more-formal membership. So, I could not join the National Association of Test Directors e-mail list because I am not a test director. (We had to contact the list coordinator and ask that person to send the request on to members.) Generally, however, literature searchers who use these channels to gather research should obtain a more heterogeneous set of studies than would be the case using a traditional invisible college or personal contacts.

Still, distribution lists will not produce studies as diverse in method and outcome as "all relevant research." Subscribers may still share certain biases. For example, I might try to gather research investigating homework by contacting the e-mail list of the American Psychological Association's (APA's) Division of Educational Psychology. Subscribers to this list might overrepresent researchers who do large-scale surveys or experiments and underrepresent researchers who do ethnographic studies. And, of course, in order to use these lists, you must know they exist, suggesting that less-established researchers are less likely to know of and contribute to them.

In sum, then, all the researcher-to-researcher channels share an important characteristic: There are no restrictions on what two colleagues can send to one another. Therefore, samples of studies found through personal contacts, mass solicitations, invisible colleges and the like are more likely to contain studies that have not undergone scrutiny by others (e.g., peer review) than will some other methods for retrieving studies. Because of the reasons suggested in Figure 3.1, many of the studies found through direct contact may never appear in more-restricted communication channels. In addition, many of the researcher-to-

researcher channels for scientific communication are likely to retrieve studies that are more homogeneous in methods and results than all studies that are relevant to the topic.

Quality-Controlled Channels

Quality-controlled channels of communication require research to meet certain criteria related to the way the research was conducted before the reports can gain entry. Whether or not the criteria are met typically is judged by other researchers who are knowledgeable about the research area, so in this way this channel resembles the traditional invisible college. It is different from the invisible college, however, in that in most instances a report submitted for inclusion in a quality-controlled channel will likely be judged by more than one person. The two major quality-controlled channels are conference presentations and scholarly journals. Their characteristics are summarized in Table 3.2.

Table 3.2 Characteristics of Quality-Controlled Channels

Channel	Restrictions on How Research Gets In	Restrictions on How Searcher Gets In	Restrictions on Types of Research Getting In
Professional conference paper presentations	Research must pass weak peer review.	Searcher must be aware that conference may have pertinent research.	Statistically significant and interesting results are more likely to appear. Research that has been previously published typically is not eligible.
Peer-reviewed journals	Research must pass strict peer review.	Searcher must subscribe to or be aware of the journal.	Statistically significant and interesting results are more likely to appear. Journals are likely to be homogeneous in methods.

Conference Presentations

There are a multitude of social science professional societies, structured both by professional concerns and topic areas, and many of them hold yearly or biannual meetings. By attending these meetings or searching the Internet for the papers given at them, you can discover what others in your field are doing and what research has recently been completed.

As an example of a search for conference presentations, in preparing this chapter I visited the website of the American Educational Research Association (AERA) and followed the link to the 2015 convention program. Along the way, I had to identify myself as a member or guest, and I had different privileges depending on what my status was. Appropriately, none of these privileges related to my access to the program proper; however, different organizations may have different rules and may restrict access to programs.

Next, I entered the search term *homework* and received the titles of 23 presentations, along with information about the session at which the paper was scheduled to be presented, and a brief abstract of the presentation. Another link took me to a description of all the papers in the session and the sponsoring division of the organization. A separate link for each presentation then took me to a page with the titles, authors, and the authors' professional affiliations.

As is typical of most websites for convention programs, there was no link to a complete paper or to specific contact information for the authors. Still, with their name and affiliation, I could easily search for an author's contact information through the AERA convention website or elsewhere on the Internet and send each author a request for a copy of the paper (and for other related papers they may have). Also, depending on the type of organization or conference, it is becoming more common for authors to be asked to submit more lengthy summaries or even complete papers along with the abstract of their presentation.

I could do similar searches separately for each AERA meeting program back to 2005. I also could conduct

similar searches for papers presented at other related meetings (e.g., the Society for Research in Child Development) as well as regional educational research associations. Or, if I wanted to do a more general search of conference proceedings, I could use the databases PapersFirst or ProceedingsFirst (available through my institutional library). These databases contain papers presented at conferences worldwide.

Limitations of information obtained through conference proceedings.

In comparison to personal contacts, the research found through conference proceedings is less likely to reveal a restricted sample of results or operations and more likely to have undergone peer review. However, the selection criteria for meeting presentations are usually not as strict as those required for journal publication; in general, a larger percentage of presentations submitted to conferences are accepted than are manuscripts submitted to peer-reviewed journals. Also, the proposals that researchers submit for evaluation by a conference committee are often not very detailed. Finally, some researchers are invited to give papers by the people who put together the meeting agenda. These invited addresses generally are not reviewed for quality: they are assumed to be high quality based on the past work of the invitee.

A search for presentations complements a search of published studies because some presentations given at meetings will describe data that never will be submitted for journal publication. Or, the data will be relatively new and will have not yet made their way through the publication process. Researchers may never follow up a presentation by preparing a manuscript or they may present a paper before a publishable manuscript has been written, reviewed, or accepted. (It is also the case that research that has already been published typically is not permitted to be given as a paper presentation at most large organization conferences.) Journals also often have long lag times between when a manuscript is submitted and when it is published. McPadden and Rothstein (2006) found that about three-quarters of the best papers presented at Academy of Management conferences eventually were published and the average time to publication was about two years after submission. Nearly half of the published papers included more or different data than were described in the conference proceedings. These new data included the addition of more outcome variables, an important component of a thorough research synthesis. A less-selective sample of all papers presented at annual conferences of the Society for Industrial and Organizational Psychology revealed that only about half were eventually published and 60% of these contained data that were different from those reported in the conference paper. For this reason, if you find a paper presentation that is relevant to your search but the conference occurred some time in the past, it is good to contact the author to see if a more complete and up-to-date description of the research is available.

Scholarly Journals

Synthesists can learn of research done in a topic area by examining the journals they themselves subscribe to, or those they believe are relevant and have access to through colleagues or their library. Journal publication is still the core of the formal scientific communication system. Journals are the traditional link between primary researchers and their audience.

Limitations of information obtained through journals.

There would be some serious biases in a literature search that used personal journal reading as the sole or major source of research. The number of journals in which relevant research might appear is generally far greater than those that a single scientist examines routinely. As early as 1971, Garvey and Griffith noted that scholars had lost the ability to keep abreast of all information relevant to their specialties through personal readings and journal subscriptions. Thus, scientists tend to restrict the journals they read routinely to ones that operate within networks of journals (Xhignesse & Osgood, 1967). *Journal networks* comprise a small number of journals that tend most often to cite research published in other network journals.

Given that personal journal reading is likely to include journals in the same network, it would not be surprising to find some commonalities shared by network members. As with personal contacts and traditional invisible colleges, we would expect greater homogeneity in both research findings and operations within a given journal network than in all the research available on a topic area.

The appeal of using personal journal subscriptions as a source of information lies in their ease of accessibility. The content of these journals also will be credible to the reference group the synthesists hope will read their work. So, *personal journal readings should be used to find research for a synthesis, but this should not be the sole source of studies.* One criticism of research syntheses in the past was that they relied too heavily on personal contacts and the synthesist's own journal network. It should be obvious now why using just these two search channels can produce a biased sample of studies.

Online journals.

The journals researchers routinely consult for work related to their interests can come to them on printed

pages or online. Online journals are rapidly replacing print journals. Online journals disseminate and archive full-text reports of scholarly work using computer storage media (see Peek & Pomerantz, 1998, for the early history of electronic journals). Many journals appear in both print and online form. Other journals are strictly paper or strictly online.

There are two characteristics of online journals that distinguish them from print journals. First, far fewer online journals than print journals use peer review procedures to screen the work they publish. It is critical for you to know which journals you have accessed do and do not evaluate submitted articles so you can use this information to assess both the potential methodological rigor of the studies and the likelihood of bias against null findings (see the following section). Second, relative to print journals online journals can have much shorter times between when a paper is accepted and when it is published. In fact, journals that appear in print and online often make the online version available weeks or months in advance of the print version. For example, the APA uses Online First to electronically publish issues of journals before they appear in print.

Just as the Internet is replacing the need for (and is removing some of the biases of) the invisible college, it is also dissolving journal networks. Two developments have opened the journal searching process in ways that help bring all sorts of journal articles to searchers. The first involves alert systems. Many journals now have systems that will inform you of the contents of current and upcoming published articles. What you need to do is visit the websites of journals that publish articles that interest you and set up an account that will put you on the alert e-mail list. These journals can span multiple disciplines. You can do this for as many journals as you wish. There are even alert services that will send you an e-mail when an article contains keywords you have designated, cutting down on content that is irrelevant to your interest. Finally, once you have published an article, you may receive an invitation to join a service that will alert you whenever your article is cited or when articles with similar content appear.

The second development involves open access journals; these journals make their articles available to readers on the Internet free of charge. The expense of preparing the manuscript and distributing it (which can be much less than print journals) is borne by the author or an institution supportive of open access. Thus, you do not have to subscribe to these journals in order to download entire articles from the Internet. With the cost of subscriptions no longer an issue, these journals can broaden your reading habits well beyond just a few journals in a journal network.

Finally, you should check with your university or place of employment to see if it has any agreements with database services that provide full text articles. These will include not only open access journals, but also other journals from publishers who charge subscription fees but agree to a single fee (paid by your employer) that allows their employees free access.

With regard to open access journals, one potential drawback for literature searchers is that, while obtaining complete journal articles is easier for readers, access for researchers may be more restricted. Because the researcher must bear the publication costs, open access journals may overrepresent (a) researchers at large institutions that may have funds available for this purpose and (b) researchers who have grants with publication costs built into the budget. While it is tempting to suggest that these restrictions provide a quality control, this is far from a forgone conclusion. For example, an intervention to increase aerobic exercise is an expensive undertaking likely requiring some form of university or external support. However, a survey on the correlates of attitudes toward rape or a laboratory study on the effect of choice on motivation is relatively inexpensive. Unfunded researchers at small institutions can conduct excellent studies on these topics. But the researchers might then shy away from publishing in open access journals; the publication costs might be the most expensive part of their studies.

Peer Review and Publication Bias

Most scientific journals (and conference programs) use peer review to decide whether to publish a particular research report. Upon submission, the journal editor sends the report to peer reviewers, who judge its suitability for publication. The primary criteria used by peer reviewers will be the methodological quality of the research and the presence of safeguards against inferential errors. However, journal reviewers will also consider the correspondence of the manuscript's content to the substantive focus of the journal and whether the article makes an important contribution to the particular research literature. Largely, these last two criteria are irrelevant to the objectives of a research synthesist. As a synthesist, you want articles related to your topic regardless of the foci of the journals you read. Also, a report of a study that is not terribly significant in its contribution, perhaps because it reports a direct replication of earlier findings, might not meet a journal's criterion for importance but it still can be very important to include in a synthesis.

The major concern raised by the fact that one criterion for publication might be the importance of the study's contribution to the field is that *research published in many journals is more likely to present statistically significant findings—that is, findings that reject the null hypothesis with a probability of p <.05 (or some other significance criterion)—than all research on the topic.* This bias against null findings is present in the decisions made by both reviewers and primary researchers. To demonstrate, Atkinson,

Furlong, and Wampold (1982) conducted a study in which they asked consulting editors for two APA journals in counseling psychology to review manuscripts. The manuscripts were identical in all respects except whether the hypothesized relation was statistically significant. Atkinson et al. found that significant results were more than twice as likely as nonsignificant results to be recommended for publication. Furthermore, they reported that the manuscripts with statistically significant results were rated to have better research designs than those with nonsignificant results, even though the methods were the same.

Primary researchers also are susceptible to bias against null findings. Greenwald (1975) found that researchers said they were inclined to submit significant results for publication about 60% of the time. On the other hand, researchers said they would submit the study for publication only 6% of the time if the results failed to reject the null hypothesis. Examining actual decisions by researchers, Figure 3.1 reveals a similar bias. Researchers' decisions not to submit statistically nonsignificant results are probably based on their beliefs that nonsignificant findings are less important and interesting than statistically significant ones. Also, they probably believe that journals are less likely to publish null results. A study by Coursol and Wagner (1985) replicated the effect of significant findings on both researchers and peer reviewers. If there is a positive note, it appears that bias against the null hypothesis has waned a bit and is taken more seriously in recent years (Rothstein, Sutton, & Borenstein, 2005).

The bias against null findings in journal publication (and conference presentations) ensures that the size of correlations or differences between the mean scores of groups reported in published works will be larger than the differences you would be likely to find in all relevant research. Lipsey and Wilson (1993) empirically demonstrated the bias against null findings. They examined 92 meta-analyses that presented separate estimates of a treatment's effect found in published and unpublished research reports. The published estimates were about one-third greater than the unpublished ones. I will return to methods for detecting and adjusting for publication bias in Chapter 6.

Bias against null findings is not the only source of bias that influences the results of published research. For example, researchers believe that their work will fare better in the peer review process if their results are consistent with the beliefs of prominent members of the field (see Suhls & Martin, 2009, for a review of numerous concerns about flaws in peer review). This phenomenon has been labeled *confirmatory bias* (Nickerson, 1998).

The existence of bias against null findings and confirmatory bias means that quality-controlled journal articles (and conference presentations) should not be used as the sole source of information for a research synthesis unless you can convincingly argue that these biases do not exist in the specific topic area.

Secondary Channels

The providers of secondary channels for obtaining research do so by gathering document information from other sources—such as journals, government agencies, and even from researchers directly—and then creating databases for searchers to use. They are constructed by third parties for the explicit purpose of providing literature searchers with lists of studies relating to a topic. The major secondary channels, summarized in Table 3.3, are research report reference lists, research bibliographies, prospective research registers, the Internet, and reference databases, including citation indexes.

Table 3.3 Quality-Controlled Channels for Locating Studies

Channel	Restrictions on How Research Gets In	Restrictions on How Searcher Gets In	Restrictions on Types of Information
Research report reference lists	Previous research must be known to the report's authors.	Searcher must be aware of the report's existence and be able to obtain a copy.	Studies in the same network are more likely to be cited, producing homogeneity in methods and results.
Research bibliographies	Compiler must be aware of research.	Searcher must be aware of bibliography.	There are few restrictions, but there is possible bias toward particular methodologies.
Prospective research registers	Researcher must know about the register or be obligated to have research listed in it.	Searcher must be aware of register.	Prospective registers likely overrepresent large-scale and/or funded research.
The Internet	Researcher must make work available on the Internet.	Searcher must choose search terms that will capture the website.	There are few, if any, restrictions. Search terms will limit the documents retrieved.
Reference databases	Research must be in a source (e.g., journal) covered by the database.	Searcher must be aware of database and search it using terms that capture the document.	Depending on the database, it may favor published research. Recent research is missing. Search terms will limit documents retrieved.
Channel	Restrictions on How Research Gets In	Restrictions on How Searcher Gets In	Restrictions on Types of Information
Citation indexes	Research must be cited in a covered source.	Searcher must know of a target document (e.g., a relevant study) that is likely to be cited by other relevant documents.	Cited research is most likely published research. Recent research is missing. Search terms will limit documents retrieved.

Research Report Reference Lists

Using the reference lists at the end of research reports to locate other reports that might be relevant to a search is sometimes called the *backward search* or *ancestry approach,* or, more informally, *footnote chasing.* It involves examining the research reports you have already acquired to see if they contain references to studies as yet unknown to you. Then, you judge the list entries (based on their title and what is written about them) for their relevance to your problem. If a reference may be relevant, you

retrieve its abstract or full report. The reference lists of these reports then can be scrutinized for further leads. In this way, you work your way back through a literature until either the important concepts disappear or the studies become so old you judge their results to be obsolete.

There is another way to chase references made possible through the use of a secondary source called *the Web of Science.* When you are viewing the full record of an article in the Web of Science, there is a link on the page: "View Related Records." Clicking on the link will display all articles that refer to at least one of the same earlier papers cited in the paper whose page you were on. Fortunately, the papers with shared references are listed high to low by the number of references they share. The assumption here is that articles sharing more references are more likely to be related to one another. For example, when I used Web of Science to retrieve the full record for an article on the association between homework and achievement that my lab had published (Cooper, Jackson, Nye, & Lindsay, 2001), the record told me that the article contained 18 references. When I clicked on the "View Related Records" link, the new page told me that there were 15,805 articles that shared at least one reference in common with our article. The first two articles listed shared six references, the next seven shared five references, and so on. I could then click on the link to each of these articles to find its full record.

Limitations of information obtained through report reference lists.

Reference lists in primary research reports are rarely exhaustive compendia of the relevant research. In fact, authors are often advised to keep such lists to a minimum and cite only the most directly related material. They are meant to provide context for interpreting the new primary research. Furthermore, primary report reference lists will tend to cite other work available through the same outlet or the small group of outlets that form an exchange network, like a journal network. Also, studies referred to in other study reports seem to be more likely to have statistically significant results (Dickerson, 2005). Therefore, you should expect more homogeneity in research methods and results found through primary report reference lists than would be the case in all relevant studies.

Another form of reference list is provided by previous research synthesists. Obviously, these can be especially helpful sources of relevant studies and will likely not contain the number and networking restrictions associated with references in primary research reports. However, even though these will be more comprehensive, you should not assume previous syntheses are based on all relevant research. To determine this, you would need to (a) read and evaluate the literature search strategies used by the synthesists and (b) determine whether their inclusion and exclusion criteria match your own. They also may be dated, thus missing the most recent research.

In sum, searching reference lists, either through the ancestry approach or through related records, will overrepresent published research because it is generally easier to find than unpublished work. Also, the most recently completed research will not appear on these lists because of the lag between when a final manuscript is submitted and when it is published. However, while reference lists in reports should not be used as a sole means for finding studies, they are generally productive sources of relevant research. Although we did not keep track of the precise numbers, we found many homework research articles by examining report reference lists.

Research Bibliographies

Research bibliographies can be either evaluative or nonevaluative listings of books, journal articles, and other research reports that are relevant to a particular topic area. Bibliographies are sometimes maintained by individual scientists, groups of scientists within a particular research area, or formal organizations. For example, while I am not aware of individuals or organizations that maintain bibliographies on homework, I do know that the Harvard Family Research Project maintains a database called the *Out-of-School Time Program Research and Evaluation Database and Bibliography*. This database contains profiles about research and evaluations conducted on both large and small out-of-school programs and initiatives. Each profile contains an overview of the program or initiative as well as detailed information about each report produced about that program.

Limitations of information obtained through bibliographies.

The use of bibliographies prepared by others can be a tremendous time saver. The problem, however, is that most bibliographies are likely to be of much greater breadth than the searcher's interest. Also, it is important to check when the bibliography was last updated. Even with these precautions, comprehensive bibliographies generated by interested parties can be a great help to you. The compilers have spent many hours obtaining information, and the biases involved in generating the bibliographies may counteract biases that exist in the other techniques you use to find research.

Prospective Research Registers

Prospective research registers are unique in that they attempt to include not only completed research,

but also research that is in the planning stage or is still underway (see Berlin & Ghersi, 2005). Today, such registers are more commonly available in the medical sciences than in the social sciences. Still, there are ways to find lists of social science research projects that are currently underway or have recently been completed. For example, the websites of many private foundations or government agencies that sponsor research can be visited to seek a list of current or recent research grants. For a topic such as homework, I might visit the websites of the W. T. Grant Foundation, the Spencer Foundation, the U.S. Department of Education's Institute for Education Science, and the U.S. National Institutes of Health.

As with bibliographies, a difficulty in finding pertinent research registers is in knowing where to start. Librarians and knowledgeable colleagues can be a big help if you do not know of funders in your area of interest.

Limitations of information obtained through prospective research registers.

From the searcher's point of view, identifying a prospective research register with relevant studies can provide access to ongoing and unpublished research that is not filtered through personal allegiances as is personal contact. In the case of funded research, you know the results will be available regardless of the study's outcome. This can be a great complement to other search channels.

That said, prospective research registers are likely to overrepresent large-scale and funded research projects. My examples of foundations and government projects make this clear. Also, the comprehensiveness of the register is of greatest importance to the literature searcher. Therefore, it is critical that searchers determine (a) how long a register has been in existence and (b) how the research included in the register got to be there.

The Internet

The capability of the Internet to assist in the transfer of information has revolutionized modern society; scientific communication has been no less affected than other areas of human interaction. The critical task for research synthesists using the Internet is to develop a strategy for finding websites with information that addresses their problem. Search engines such as Google, Yahoo, and Bing can be used for this purpose. However, it is important to keep in mind that studies of the overlap in search results when different search engines are used suggest that a large portion of the pages in the major search engine's database exist only in that database and first page results overlap minimally across the search engines (http://searchengineland.com/070601-094554.php). So, *it is good practice to use more than one search engine to be certain you are conducting a thorough search of the Internet*. A good list of search engines for conducting academic-related searches can be found at Te@chthought (http://www.teachthought.com/technology/100-search-engines-for-academic-research/).

Also, you should keep in mind that search engines do not make judgments about the quality of the material contained on websites. Especially when research is involved, where a high degree of expertise is needed to carry out a credible study it is important to make sure the information you glean from the Internet is reliable. This means you should not rely on secondary sources of information about studies. If you find a secondary source that describes a study that interests you, contact the researchers directly. Sometimes you will find full reports of research posted on websites.

Internet searching is now a way of life. You provide a search term, phrase, or set of terms and phrases to the search engine. All the search engines in some way permit the use of Boolean syntax operators to expand or restrict the search. Boolean operators allow the searcher to use set theory to help define the items that will be retrieved by a search. However, how websites are represented and the precise commands used to do the Boolean syntax search will differ somewhat for each search engine. All three search engines mentioned above provide online assistance to help you learn how to use them.

The result of a search will be a list of websites that fit the keyword description, most often because the website contains the keyword or words somewhere on the web page. The order in which the websites appear on the results page will be determined by the search engines algorithms, usually a function of the degree of match between search terms and the website content, as well as how frequently the website is viewed.

As an Internet search example, while preparing this chapter I began a search for homework research by asking a search engine to list all websites that included the term *homework*. Bad idea: Google found about 167,000,000 websites. Of course, many of these websites include homework assignments posted by teachers on the Internet, tips on how to do homework, newspaper articles about homework, and so on. The numbers were no less daunting when I added the term *research* to the search and required that both be present for the website to be retrieved (just 100,000,000). Even requiring that the two terms appear adjacent to one another led to 218,000 sites in Google.

As these results suggest, using the Internet to find scientific research on a specific topic can be overwhelming and time consuming. It would be nonsensical to go to each of these websites to see if it

reported research relevant to the research synthesis (and not found through other channels). The Internet contains much more than research information. An Internet search using the terms *research engines* and *social science* will lead to other sites that list search engines more specific to your purposes and potentially related to your topic. The search engines listed in these sites primarily provide computer access to research registers and to the reference databases that I will describe shortly.

The strategies for searching the Internet I have described are only some examples of numerous approaches. I am being deliberately general here because these resources change quickly. With practice, you will become more familiar with the resources available to you and how to construct searches that produce relevant material.

Limitations of information obtained through the Internet.

Internet websites can be constructed by anyone who has (or knows someone who has) the required expertise. Thus, there is little restriction on what information can be made available on websites. And, of course, this can be both a good thing and a bad thing because the amount of information can be exhaustive but overwhelming, and with no quality check on content.

Reference Databases

Finally, the sources of information likely to prove most fruitful to research synthesists are called *reference databases*. These are indexing services maintained by both private and public organizations associated with social science (or other) disciplines.

Our search for homework studies used four reference databases. We searched the Education Resource Information Center (ERIC), PsycINFO, Sociological Abstracts, and Dissertation Abstracts electronic databases for documents cataloged between January 1, 1987, and December 31, 2003. Because these databases and their interfaces are constantly being updated, I recommend that you visit their web pages or your library's resource pages to get the most current information about them and many other databases. One good general source of information on databases is the Gale Directory Library (http://www.gale.cengage.com/DirectoryLibrary/).

Another reference database used frequently across all the sciences is the Web of Science Core Collection (http://wokinfo.com/). When I ran my search, using the term *homework* with no restrictions on dates or indexes searched but restricting my record field to "topic" (excluding authors and publications named "homework") I retrieved 3,174 documents. The core collection searches several separate databases related to different disciplines, including the Science Citation Index Expanded, Social Sciences Citation Index, Arts & Humanities Citation Index, as well as several indexes that include conference proceedings and books. If I restricted my search to only the time span 1987 to 2015 and only the Social Science Citation Index, the result was reduced to 1,902 documents. The Web of Science search home page links to a tutorial that gives you a quick tour of what it has to offer, and to tips on searching and training materials on some of its advanced procedures (e.g., citation reports and maps).

A relatively new reference database that is broad in its reach and free to the public is Google Scholar (http://scholar.google.com/intl/en/scholar/about.html). This search engine is restricted to scholarly documents and permits the specification of searches that can take many of the irrelevant documents out of the search results. So, searching "homework" in Google Scholar retrieved 457,000 documents. But when I used the Advanced Search feature to specify that both the terms *homework* and *effects* had to appear in the title of a document appearing between 1995 and 2015, I found 245 entries.

Limitations of information contained in reference databases.

Even though reference databases are superb sources of studies, they have limitations. First, there can be a time lag between when a study is completed and when it will appear in the reference database, though technology has reduced this lag dramatically. Still, the study must be written up and submitted, accepted into its primary outlet, appear in print or online, and then cataloged into the reference database. So, the most recently completed research—the type you would find by contacting researchers or research contained in prospective registers (and perhaps on the Internet)—will not appear in reference databases. Second, each reference database contains some restrictions on what it allows to enter the system based on topical or disciplinary boundaries. Therefore, if you are interested in an interdisciplinary topic you will need to access more than one reference database. For example, studies about homework certainly interest education researchers but might also appear in psychology or sociology journals. Third, some reference databases contain only published research, others both published and unpublished research, and others just unpublished research (e.g., dissertation abstracts). So, if you want to minimize publication bias, it is important to find out the coverage of the databases you plan to use and try to include databases that assemble unpublished as well as published documents.

Citation indexes.

A citation index is a unique kind of reference database that identifies and groups together all published articles that have referenced (cited) the same earlier publication. In this way, the earlier publication becomes the indexing term for the more recent articles. In contrast to using research report reference lists to look backward for the ancestors of a report, a citation index does a *forward search* using a *descendance approach*. Three citation indexes produced by the Institute for Scientific Information are available through subscription or most research libraries and can be entered through the Web of Science. It provides access through the Web of Science Core Collection to the Science Citation Index Expanded (which includes published articles in science journals from 1900 to the present), the Social Sciences Citation Index (beginning in 1956), and the Arts & Humanities Citation Index (beginning in 1975). As noted, the Web of Science also provides cited reference searching and thus allows users to move forward and backward through the literature.

Past research is not the only way to track down descendants. The search for studies concerning individual differences in rape attitudes also used the Social Sciences Citation Index to great effect even though we could not identify seminal research articles. Here, five frequently used measures of attitudes toward rape were identified and the articles in which the measures were originally described were used to access the citation index. We found 545 citations of the five scales and examined their abstracts to determine if the studies were relevant to the study of individual differences.

Limitations of information contained in citation indexes.

Citation indexes limit entry to references in published research, both journals and books. Therefore we can expect a bias against the null hypotheses in citations in the same way we expect this in the references we find in research reports. However, the coverage of the Social Science Citation Index is quite exhaustive within these categories. Also, citation indexes will miss more-recent publications because of the time it takes to index documents.

Database vendors.

All major research libraries have numerous reference databases available. Reference librarians can help you identify the databases most appropriate for your search and can provide the introductory instructions needed to access them. Once you have identified the databases of relevance to your search, the database interface will contain step-by-step, menu-driven instructions that make them easy to use. However, the same database can be provided to your library by more than one vendor and the results of your search using the same database may be slightly different depending on which vendor you choose. This likely happens because of differences in the frequency with which the databases are updated.

Conducting Searches of Reference Databases

Research libraries employ trained specialists who can conduct your search or help you through the process. It is good practice to discuss your search with a trained research librarian before you begin; they are likely to have suggestions about places to look that you have not thought of. Also, there are many publications that can help you start thinking about a search. Reed and Baxter (2009) provide a more in-depth treatment of reference database searching strategies in the context of research synthesis.

Typically, you begin your reference database search by deciding which databases to access. So, we started a search for homework research by searching PsycINFO, ERIC, Sociological Abstracts, and Dissertation Abstracts. The synthesis on aerobic exercise searched 13 databases, including not only PsycInfo and ERIC but also databases covering research on medicine, health, exercise, and aging. You choose the databases based on your understanding of how likely they are to include documents relevant to your search. You should also be familiar with the database catchment of documents. Does it include journals that are likely to include relevant research? Does it contain only journal articles or also paper presentations, dissertations, and other kinds of reports? If a certain type of report is lacking, does another database include these reports?

Before entering any of my choices of databases, I checked to see if more than one vendor offered them through my university. I found that my university offered PsycINFO through two vendors. One vendor permitted me to search both PsycINFO and ERIC at the same time. Therefore, using this vendor would save me the effort of removing duplicate documents from separate searches of each database. I also found little difference in the number of documents retrieved using the term *homework* in the ERIC database using the two different vendors. (I chose ERIC to run this test because I anticipated that it is the database that would reveal the most relevant documents.) So, I chose the vendor that permitted me to search jointly ERIC and PsycINFO.

Keywords and other search parameters.

Next, I chose the terms I used to search for documents. A searcher can browse through thesauri that accompany the different databases to identify terms that might not have initially come to mind. You can

also use examples of documents you hope to retrieve and see what terms are used to index these documents or that appear in their title or abstract. This gives you some concrete idea about how to access the material you want. Also, if you run a search and it does not capture documents you know are relevant, something has gone awry.

Regardless of how you identify terms to use in a search, when you evaluate the search procedures used in a research synthesis, you should ask the question,

> Were proper and exhaustive terms used in searches and queries of reference databases and research registries?

I began my search simply with the term *homework* because I first wanted to explore whether related terms exist. My search engine had a link to the ERIC Thesaurus, which relates the term *homework* to the terms *assignments* and *home study.* I can then use an explode function to expand these terms and examine yet more terms. These terms seemed too far afield (*instruction* comes up as a term related to assignments, and *distance education* as a term related to home study), so I decided to use just "homework" without much concern that too many relevant documents will be missed. The thesaurus for PsycINFO told me that the term *homework* was added in 1988. It also told me that the definition of homework used in the database was an "assignment given to students or clients to be completed outside regular classroom period or therapeutic setting" (APA, 2015). Here, then, I encountered an instance in which the same term is used in two very different contexts, one academic and one therapeutic. This alerted me to the possible need to restrict my search in some way to exclude therapeutic homework in a clinical situation. The PsycINFO thesaurus offered three related terms: *note taking, psychotherapeutic techniques,* and *study habits.* I decided that none of these would likely add many relevant studies if the term *homework* is already in the search.[2]

It is also possible to truncate your keywords so you can capture variations on the theme. For example, the search for aerobic exercise intervention effects used the truncated keyword "cogniti*." That way, the search would pick up words such as "cognition," and "cognitive," as well as any other words with the same first letters.

Next, I set my search parameters. I decided to use the search terms *homework* and *achievement*. I wanted both terms to appear. So I used the Boolean operator AND. The search engines will also let you use the OR operator, and sometimes the NOT operator, when you do not want documents containing a keyword. For example, I might have wanted to exclude homework reports that included the word "college" in the title. The addition of "achievement" to the search should exclude all or most of the research in which homework is part of a therapeutic regimen. My search engine then gave me a series of other decisions to make regarding whether I wanted to restrict my search to, for example, only journal articles, only articles intended for specific audiences, and only studies using particular research methodologies. Consistent with my problem definition, I decided to leave the search unrestricted except for two parameters. First, I wanted documents that pertained only to school age children (6 to 12 years old) and adolescents (13 to 17 years old), not early childhood or adulthood. Second, I only wanted documents that appeared since 2006, the last time we synthesized the research on homework.

Finally, I decided that I wanted to see only documents that used the two terms *homework* and *achievement* in the abstract. I could restrict documents to those that used homework in the title, but that seemed too restrictive. I am aware that there are some studies that use homework as one of many predictors of achievement and therefore it is more likely to be mentioned in the abstracts of these articles but not in the title. Including all documents that mention homework anywhere in the text seemed too inclusive.

It is also possible to run more than one search and combine them in some fashion. I could have run a search on "homework" and another on "achievement." I could then have asked the search engine to combine these into a third search or to create a third search that included documents that appeared in one search but not the other. This can be a good strategy early in a search when you are still considering the breadth of your conceptual definitions and whether broader definitions are possible. It can be used to tell you how many documents are added or removed depending on the breadth of your concepts.

In making these decisions, I was making trade-offs between the recall and precision of my search. The term *recall* relates to the percentage of all relevant documents that my search uncovers. I want high recall so I do not miss documents. However, the higher the recall of my search (the more and broader the keywords were that I use in the search), the more likely it is that I will retrieve many document records that are irrelevant to my search. The term *precision* relates to the percentage of all retrieved documents that are relevant to my search. The more precise my search, the more likely it is that I will miss some relevant studies. Obviously, as the recall of a search goes up, the precision goes down. The keywords you choose will determine the recall and precision of your search.

On the day I conducted this search, I found about 150 documents that met the inclusion criteria. I could

then repeat my search with the other databases, keeping my search parameters as similar as possible.

Let me also illustrate a search of a citation index. I did a search for my book *Homework* (Cooper, 1989) using Google Scholar. First, I opened the "Advanced Scholar Search" by clicking on the down arrow in the search box. I provided the last name of the author (Cooper), the keyword (homework), indicated that I wanted the keyword in the title of the work and that the work should have appeared in 1989. This is a very restricted search, as it is looking for citations to only a single publication. Had I entered only "Cooper H" and left the cited work and year unspecified, I would retrieve citations to all documents authored by scholars who share this last name and first name initial. My search gave me 11 results with my book at the top of the list. Under a brief abstract was a link, "Cited by 528." Clicking on this link brought me to a page with the titles, authors, location (e.g., the journal the article was published in), and a brief abstract of each document that cited my book.

As my example reveals, another limitation on the exhaustiveness of searches based solely on reference databases derives not from what they contain but from how they are accessed by searchers. *Even if a database were to have exhaustive coverage of the documents that are relevant to your topic, you will not necessarily be able to describe your topic in a manner that ensures you uncover every relevant article in it.* The search may not recall all the wanted information. Like searching the Internet, searchers must enter the database by specifying search terms associated with particular research topics. Searchers who are unaware or omit terms that apply to documents relevant to their interests are likely to miss articles. All searchers make trade-offs between the likelihood of (a) missing relevant documents and (b) including lots of irrelevant documents.

Determining the Adequacy of Literature Searches

The question of which and how many sources of information to use in a search has no general answer. The appropriate sources will be a function partly of the topic under consideration and partly of the resources available to you. As a rule, however, *searchers must always use multiple channels with different entry and access restrictions so that they minimize any systematic differences between studies that are and are not found by the search.* If a searcher has uncovered different studies through channels that do not share similar entry and access restrictions, then the overall conclusions of the synthesis should be replicable by someone else using different, but also complementary, sources for primary research. This rule embodies the scientific principle of making results replicable. So, an important question to ask about the adequacy of the search strategy used in a research synthesis is,

Were complementary searching strategies used to find relevant studies?

Reference databases and research registers, if they are available, should form the backbone of any comprehensive literature search. These sources probably contain the information most closely approximating all research. Typically, they cast the widest net. Their restrictions are known and can be compensated for by the use of other complementary search strategies.

Earlier in this chapter, I mentioned that concentrating on only quality-controlled sources would produce a set of studies that overrepresented statistically significant results. However, because these sources involve peer review, it could be argued that this research has undergone the most rigorous methodological appraisal by established researchers and probably is of the highest quality. As we shall see in Chapter 5, publication does not ensure that only studies of high quality will be included in the synthesis. Faulty studies often make their way into journals. Also, well-conducted studies may never be submitted for publication.

A focus on only published research might be legitimate in two circumstances. First, published research often contains several dozen, or in some cases hundreds, of relevant studies. In such an instance, it is likely that while the published research may overestimate the certainty with which a null hypothesis can be rejected and the size of the relationship, it probably will not incorrectly identify the direction of a relationship. The suggested magnitude of the relation can be adjusted for the possibility of bias against null results. (I will return to this in Chapter 7.) Also, enough instances of a hypothesis test will be covered to allow a legitimate examination of which study characteristics co-vary with study outcomes.

Second, there are many hypotheses that have multiple tests in the literature that were not the primary focus of the research. For instance, many psychological and educational studies include the participants' sex as a variable in the data analysis and report hypothesis tests of sex differences, although these are only an ancillary interest of the primary researchers. The bias toward significant results in publications probably does not extend much beyond the primary hypothesis. Therefore, a hypothesis that appears in many publications as a secondary interest of the researchers will be affected by bias against null results to a lesser degree than the researcher's primary focus.

Generally speaking, however, focusing on only published studies is not advisable. The possibility of bias

against the null hypothesis is too great. In addition, *you should not restrict your search to published outlets even if you ultimately decide to include only published work in your synthesis.* To make a well-informed choice about what to put in and leave out, and even to help you decide what the important issues are in a field, you need to have the most thorough grasp of the literature.

Finally, the information contained in channels involving personal contact with researchers is not likely to reflect information gleaned from all potential sources. However, research found by contacting researchers directly likely will complement that gained through other channels because it is likely to uncover research that is more recent.

Problems in Document Retrieval

Depending on the databases you use and the nature of your topic (especially the age of the research you want to retrieve), once you have your search results, you will be able to retrieve relevant articles through printed sources and the Internet. For some documents, typically older ones, you might also have to use microfiche records, though their use is becoming rare (if it is still available at all) because more and more early research reports are being digitized. Digitization of documents has made retrieval much easier, and as more documents are stored and accessed online, you will find retrieval is just a few keystrokes away. And, as I mentioned above, open access journals and institutional subscriptions to journals are making online access to research even easier.

However, some deficiencies in document retrieval procedures will frustrate you regardless of how thorough and careful you try to be. Some potentially relevant studies do not become public and defy the grasp of even the most conscientious searchers. Other documents you will become aware of but you will be unable to obtain.

Every research synthesist will find that some documents of potential relevance (based on their title or abstract) cannot be obtained from their personal journals, institutional library's print or microfiche collections and will not be available electronically. To what lengths should you go to retrieve these documents? The use of interlibrary loans is a viable route. Dissertations and master's theses can be obtained through interlibrary loans, and dissertations can be purchased through Proquest UMI. Contacting the primary researchers directly is another possibility, although personal contact often results in only a low rate of response. Whether or not a primary researcher can be located and induced to send a document is influenced in part by the age of the requested material, whether it is digitized, and the status of the requester.

In general, when deciding how much effort should be expended trying to retrieve documents that are difficult to obtain, you should consider (a) the likelihood that the needed document actually contains relevant information, (b) the percentage of the total known documents that are difficult to find and how their results might differ from the results of studies you have, (c) the cost involved in undertaking extraordinary retrieval procedures (e.g., interlibrary loan is cheap, buying dissertations is expensive), and (d) any time constraints operating on you.

The Effects of Literature Searching on Synthesis Outcomes

At the beginning of this chapter I mentioned that literature searches have two different targets—previous research and individuals or groups relevant to the topic area. Therefore, it is necessary for you to address the adequacy of your accessed studies with respect to each of the targets. You must ask (a) how the retrieved studies might differ from all studies and (b) how the individuals or groups contained in retrieved studies might differ from all individuals or groups of interest.

Much of this chapter has dealt with how to answer the first of these questions. Not every study has an equal chance of being retrieved. It is likely that studies easily obtained through your retrieval channels are different from studies that never become available. Therefore, you must give careful thought to what the results of inaccessible studies might be and how this might differ from what is found in studies that have been retrieved (again, I return to this topic in Chapter 7).

The synthesist's second population of interest, referring to individuals or other basic units of analysis, injects a note of optimism into the discussion. There is good reason to believe research syntheses will pertain more directly to a target population than will the separate primary research efforts in the topic area. The overall literature can contain studies conducted at different times, on units with different characteristics, and in different locations. A literature can also contain research conducted under different testing conditions with different methods. For certain problem areas containing numerous replications, the diversity of samples accessible to a synthesist should more closely approximate the target population of the primary researcher.

Of course, we must bear in mind that the biases against null results and contradictory findings may affect the available samples of people as well as studies. *To the extent that more retrievable studies are*

associated with particular subpopulations of elements, retrieval biases will be associated not only with the outcomes of studies, but also with the characteristics of study samples.

The best way to ensure that the sample of studies in your syntheses is representative of all research on your topic is to conduct a broad and exhaustive search of the literature. While the law of diminishing returns applies here, a complete literature search has to include at least

- A search of reference databases,
- A perusal of relevant journals,
- The examination of references in past primary research and research syntheses, and
- Personal contacts with active and prominent researchers.

The more exhaustive a search, the more confident you can be that other synthesists using similar, but perhaps not identical, sources of information will reach the same conclusions. Table 3.4 presents an example of a log that can be used to keep track of the techniques you used to search the literature. It is important to keep track of this information because you will need much of it when you write up your synthesis report.

Also, in your analysis of your synthesis' results, you should present indices of potential retrieval bias, if they are available. For instance, many research syntheses examine whether any difference exists in the results of studies that are published versus those that are unpublished. Others examine the distribution of results to see if they suggest that some results are missing. Techniques for conducting these analyses are discussed in Chapter 7.

Table 3.4 Log Example for Literature Search

Researcher-to-Researcher Search Techniques	Used? When?	Who Was Contacted?	Date Sent	Date Reply Received	Nature of Reply
Personal contact	__ Yes Date_____ __ No Reason:	Researcher Names _____ _____ _____ _____	_____ _____ _____ _____	_____ _____ _____ _____	_____ _____ _____ _____
Mass solicitation	__ Yes Date_____ __ No Reason:	Organization Names _____ _____ _____ _____	_____ _____ _____ _____	_____ _____ _____ _____	_____ _____ _____ _____
Traditional invisible college	__ Yes Date_____ __ No Reason:	_____ _____ _____ _____	_____ _____ _____ _____	_____ _____ _____ _____	_____ _____ _____ _____

Table 3.4 (Continued)

Electronic invisible college	__ Yes Date_____ __ No Reason:	Organization Names ———————— ———————— ———————— ————————	———— ———— ———— ————	———— ———— ———— ————	———— ———— ———— ————
Quality-Controlled Search Techniques	**Used?**	**Organization Names or Journal Titles**	**Years Searched**	**Number of Documents Examined**	**Number of Relevant Documents Found**
Professional conference paper presentations	__ Yes Date_____ __ No Reason:	Organization Names ———————— ———————— ———————— ————————	———— ———— ———— ————	———— ———— ———— ————	———— ———— ———— ————
Peer-reviewed journals	__ Yes Date_____ __ No Reason:	Journal Titles ———————— ———————— ———————— ————————	———— ———— ———— ————	———— ———— ———— ————	———— ———— ———— ————
Secondary Search Techniques	**Used?**		**Years Covered**	**Number of Documents Examined**	**Number of Relevant Documents Found**
Research report reference lists	__ Yes Date_____ __ No Reason:	# of Reports Examined ————————	——–——	————	————————
Research bibliographies	__ Yes Date_____ __ No Reason:	Source (Name) of Bibliography ———————— ———————— ———————— ————————	——–—— ——–—— ——–—— ——–——	———— ———— ———— ————	———————— ———————— ———————— ————————
Prospective registers	__ Yes Date_____ __ No Reason:	Register Names ———————— ———————— ———————— ————————	——–—— ——–—— ——–—— ——–——	———— ———— ———— ————	———————— ———————— ———————— ————————

Table 3.4 (Continued)

Secondary Search Techniques	Used?		Years Covered	Search Terms	Other Restrictions
Internet	__ Yes Date_____ __ No Reason:	Search Engines _____ _____ _____ _____	___–___ ___–___ ___–___ ___–___	_____ AND/OR _____ AND/OR _____ AND/OR _____	_____ _____ _____
Reference databases	__ Yes Date_____ __ No Reason:	Database Names _____ _____ _____ _____	___–___ ___–___ ___–___ ___–___	_____ AND/OR _____ AND/OR _____ AND/OR _____	_____ _____ _____
Citation indexes	__ Yes Date_____ __ No Reason:	Index Names _____ _____ _____ _____	___–___ ___–___ ___–___ ___–___	_____ AND/OR _____ AND/OR _____ AND/OR _____	_____ _____ _____

Exercises

1. Using the topic area you identified in Chapter 2, conduct a search of a reference database. Perform a parallel search of another database or the Internet. How are the outcomes different? Which was more useful and time- and cost-effective?
2. For a topic of your choice, choose the channels you would use to search the literature and the order in which you would access them. For each step in the search, describe its benefits, limitations, and cost-effectiveness, given your topic.

Notes

1. Another strategy would be to start a distribution list related to the topic of interest. This strategy would take longer to pay off but might reap great rewards when it did so.

2. In some databases you might run into a distinction between natural language key words or search terms and controlled vocabulary. Natural language consists of the words researchers and searchers use to describe research. Controlled vocabulary consists of terms added to document records by the database constructors to describe documents. Today, the distinction will not much change what you do, but you may be happy that the controlled vocabulary has been added to the record because it tends to diminish the scatter of a literature.

4 Step 3 Gathering Information From Studies

What procedures should be used to extract information from each study report?

<div style="border: 1px solid black; padding: 10px;">

Primary Functions Served in the Synthesis

1. To create a coding frame for obtaining information from studies
2. To train coders
3. To assess the accuracy of extracted information

Procedural Variation That Might Produce Differences in Conclusions

1. Variations in the information gathered from each study might lead to differences in what is tested as an influence on cumulative results.
2. Variations in coder training might lead to differences in entries on coding sheets.
3. Variation 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.

Question to Ask When Evaluating the Information Gathered From Each Study to Be Included in a Research Synthesis

Were procedures used to ensure the unbiased retrieval of information from study reports?

This chapter describes

- How to construct a coding guide that will gather the important information about studies to be included in a research synthesis
- How to train coders so the information about studies will be gathered reliably
- Issues in judging whether separate outcomes from the same study should be considered independent outcomes
- What to do when information about a study is missing

</div>

So far, you have formulated the problem you want to explore in your research synthesis. You know the crucial issues that have come to the attention of theorists, researchers, and previous synthesists. And your literature search is underway. The next step in your synthesis is to begin the construction of your coding guide. The coding guide is the device you (and those who are assisting you) will use to gather information about each study. Most of this information will come from the study report itself, but some information may come from other sources as well.

Inclusion and Exclusion Criteria

I touched on how you make judgments about the relevance of studies when I discussed how a problem gets defined: you tie conceptual variables to observable research operations and measurements. Broadly defined concepts in a research synthesis will encompass more operational definitions than narrowly defined concepts. After the initial screening of studies, the coding guide you devise will direct the retrieval of information from studies. The guide needs to tell coders what characteristics of studies need to be present for a study to be included in the synthesis. It is where the conceptual rubber hits the operational road.

But conceptual relevance might not be the only criterion you wish to use for inclusion of studies. You might decide that a study that examines the hypothesis or intervention of interest to you conceptually does not match up with other criteria you want studies to meet. For example, you might want to limit studies based on when they were conducted. Our homework research synthesis excluded studies conducted before 1987. We used this criterion because an early synthesis ended with that year and we did not want our synthesis to cover overlapping research. The synthesis on aerobic exercise was limited to studies that used random assignment of participants to treatments. These were plentiful enough that limiting the synthesis to this type of research design, the one that allows the strongest causal inferences, was feasible (more on this in [Chapter 5](#)). In addition to timeframe and study design, other possible inclusion or exclusion criteria include characteristics of the study's context (e.g., its authors, dissemination outlet, funding source), participant sample (e.g., age, sex, economic status, geographic location), and outcomes (types of measures and their psychometric characteristics).

Sometimes inclusion and exclusion criteria other than conceptual relevance can be applied before the coding of studies even begins; it is easy to identify and exclude studies that are older than you wish. It is also possible, however, that you will begin coding studies and decide after the fact that additional screens need to be added. The coding sheet should allow you to do this. Also, you may decide that rather than exclude studies based on, say, the country in which they were conducted, you will use this variation in

study context as a possible moderator of study outcomes.

Developing a Coding Guide

If the number of studies involved in your synthesis is small, it may not be necessary before you begin to examine the literature to have a precise and complete idea about what information to collect about the studies. The relevant reports, if only a dozen or so exist, can be retrieved, read, and reread until you have a good notion of what aspects of the studies would be interesting to code or how often the important characteristics suggested by others actually appear in the studies. For example, you might be interested in whether the effect of homework is moderated by the SES of students but you find that very few studies report the SES of the students taking part.

Of course, *if you read the entire literature first and then decide what information to code about each study, your choices of codes are post hoc and should not be solely dictated by what your reading suggested will be significant predictors of results*. If you do this, the proportion of significant results you get might be greater than if you chose predictors based solely on their theoretical or practical importance. Still, small sets of studies allow you to follow up on ideas that emerge only after the studies have been read. Then, you can return to previously read studies to code the new information you did not realize was important during the first reading.

If you expect to uncover a large number of studies, reading then rereading reports may be prohibitively time-consuming. In this case, it is necessary to consider carefully what data will be retrieved from each research report before the formal coding begins. Of course, reading a few randomly chosen studies can help you think about information to code and is something you should do. In fact, if you are interested in conducting a research synthesis, you are probably already familiar with many studies in the area.

When an area of research is large and complex, the construction of a coding guide can be no small task. The first draft of a coding guide should never be the last. First, you need to list all the characteristics of studies you want to gather. Then, you need to consider what possible values studies might take on each variable. For example, in a research synthesis of interventions to increase aerobic exercise among adults, you would certainly want to gather information on the age of participants and characteristics of the interventions, such as their length and intensity. You might decide that your definition of the term *adults* includes people over the age of 18, but participants still might be much older than this, which might influence the effects of the intervention. So you might want your coding guide to help you gather information on the range in age among participants. You might exclude studies involving adolescents, but the coding guide would still contain a question about the age of the youngest participant in the study, one about the oldest participant, and the mean and/or median age of participants.

After you have this preliminary set of coding questions and response categories, you need to show this first draft to knowledgeable colleagues for their input. They are certain to suggest additional codes and response categories. They will also point out instances in which your questions and responses are ambiguous and thus difficult to understand. After taking their advice, you should code a few randomly selected studies using the coding guide. This will add further precision to questions and response categories.

An important rule in constructing a coding guide for research synthesis is that *when many studies are involved, any information that might possibly be considered relevant should be retrieved from the studies.* Once data coding has begun, it is exceedingly difficult to retrieve new information from studies that have already been coded. Some of the information you gather on the coding sheets may never be examined in your completed synthesis. Sometimes, too few studies will report information about the variable of interest. In other cases, studies will not vary enough across values of a characteristic to allow valid inferences. For example, you might include a question about the health status of participants and discover that most if not all exercise interventions have been conducted with participants who have experienced a health problem. Still, it is much less of a problem to gather more information with your coding guide (by including a question about health status) that you may eventually find useful than it is to have to return to reports to get information that was neglected the first time through.

Information to Include on a Coding Guide

While the content of every research synthesis coding guide will be unique to the question asked, there are certain broad types of information that every synthesist will want to gather from primary research reports. Here, I will classify these types of information into eight categories:

1. The report
2. The predictor or independent variable
    1. If the report describes an experimental manipulation, information about the manipulated conditions—that is, the intervention (such as homework or exercise programs) or the independent variable (if the study is testing basic theoretical predictions, such as the effects of

task choice)
2. If the report describes nonmanipulated predictor variables, information about how these were collected and their psychometric characteristics (e.g., the scales used to measure participants' individual differences and attitudes toward rape)
3. The setting in which the study took place
4. Participant and sample characteristics
5. The dependent or outcome variables and how they were measured (such as level of achievement, amount of physical activity, motivation, or rape myth acceptance)
6. The type of research design
7. Statistical outcomes and effect sizes
8. Coder and coding process characteristics

In this chapter I will focus on six of the eight types of information about a study. I will return to discuss how to code research designs in Chapter 5 and statistical outcomes in Chapter 6, when each of these topics is covered in more detail.

A general coding guide will never capture all the important aspects of all studies. The questions that should guide your construction of the material to be retrieved from studies should include the following:

- Are there any theoretical and applied issues that need to be captured in the coding?
- Do theories suggest what study characteristics might be important and how the studies might differ on these characteristics?
- Are there issues in practical application that suggest that the way studies are conducted could relate to the impact of the intervention or policy?
- Are there any methodological issues that have arisen in the interpretation of past research?
- How might methods vary in ways that could relate to study outcomes?
- Are there disputes in the literature that relate to how studies are conducted?

Finally, completed coding sheets are often characterized by numerous entries left unfilled (I will return to this later) and notes in margins. Coders sometimes will feel as though they are slamming round pegs into square holes. Perfection is never achieved. Therefore, *it is good practice to leave coders space to make notes about on-the-spot decisions they made.* In general, the rules for constructing a coding guide are similar to rules used in creating a coding frame for a primary research effort (Bourque & Clark, 1992); a more-detailed description of the process in research synthesis can be found in Wilson (2009) and Orwin and Vevea (2009).

Characteristics of the report.

Table 4.1 provides an example of a coding guide for report characteristics. My example is set up for coding on printed pages. If coders are coding directly into a spreadsheet, the response column is not needed. It is extremely important, however, that the spreadsheet clearly identify which column is devoted to which code. If you are using a program, such as Access, each coded variable can have its own page and coders can then click the appropriate response box. Also, when coding directly into spreadsheets you may forgo numbering many responses and simply type the coded response into the spreadsheet cell, for example by typing in "journal" rather than "1" for question R4 below.

Of course, if you type responses directly into the spreadsheet, spelling mistakes will appear as separate categories of response. On the positive side, typing in responses can be good for spotting errors. Typing "journal" into a spreadsheet column will make it obvious if the entry is in the wrong column, more so than if you are typing in numbers, which might be used repeatedly in adjacent columns.

Table 4.1 Example Coding Sheet for the Report Identification Section of a Coding Guide

Report Characteristics	
R1. What is the report ID number?	____ ____ ____
R2. What was the first author's last name? (Enter ? if you can't tell.)	_____
R3. What was the year of appearance of the report or publication? (Enter ? if you can't tell.)	__ __ __ __

R4. What type of report was this?
    1 = Journal article
    2 = Book or book chapter
    3 = Dissertation
    4 = MA thesis
    5 = Private report
    6 = Government report (federal, state, country, city)
    7 = Conference paper
    8 = Other (specify) _____
    ? = Can't tell
	_____

R5. Was this a peer-reviewed document?
    0 = Not peer reviewed
    1 = Peer reviewed
    ? = Can't tell
	_____

R6. What type of organization produced this report?
    1 = University (specify) _____
    2 = Government entity (specify) _____
    3 = Contract research firm (specify) _____
    4 = Other (specify) _____
    ? = Can't tell
	_____

Table 4.1 (Continued)

Report Characteristics	

R7. Was this research conducted using funds from a grant or other sponsor?
    0 = No
    1 = Yes
    ? = Can't tell
	_____

R7a. If yes, who was the funder?
    1. Federal government (specify) _____
    2. Private foundation (specify) _____
    3. Other (specify) _____
	_____

Note that an "R" is placed before each question number in the first column. This was done to distinguish questions about the report from questions about other features of the study, which will be given other letters, such as "I" for intervention characteristics and "O" for outcome characteristics. Doing this is really a matter of personal taste: you could also just number the questions successively. Also, note that all possible responses to each question are listed below the question and each response is given a number

that will be entered by the coder into the spaces provided in the second column. Some responses are simply "other." This code will be used if the coder finds a report characteristic that does not correspond to any response listed above it. When the "other" response is used, the coder is asked to provide a brief written description of the characteristic. Some of the questions also provide a "can't tell" response. Coders will use a question mark in the response space for "can't tell." This makes it easy to distinguish missing information from other coded values. I have repeated the "can't tell" response in most of the questions, but to save space coders could be instructed simply to use this convention throughout the coding sheet.

You will want to start by giving each report a unique identification number (Question R1). Later, you will also give unique numbers to each study in a report (if there is more than one study in it), to each unique sample within a study for which separate data are reported, and to each outcome reported within each sample.

Next, you will want to include on your coding sheet enough information about the first author of the study so that if you later want to group studies by their author (perhaps to test whether different authors get different results), you will be able to do so. In Table 4.1, the first author's name is used for this purpose (Question R2). Note that this is one of only two responses in the coding guide example that do not use numbers; the other one asks for the postal code of the state in which the study was conducted.

Third, you will want to know the year in which the report appeared. This might be used later to examine temporal trends in findings, or simply to help uniquely identify the study (along with the first author's name) in summary tables.

Fourth, you will want to describe the type of report and whether the report had undergone some form of peer review before it appeared. This information will later be used to test for the possibility of publication bias. Note here that the response categories are "mutually exclusive" (every report should fall into only one category) and "exhaustive" (every report will have a category).

Finally, you might be interested in what type of organization produced the report and whether the report was done with some type of funding support. This information can be critical if you discover that the funders of some studies might have had a monetary or other interest in whether studies had a particular outcome; an example would be a chain of gymnasiums that supports a study on the value of exercise for older adults. If so, you might want to see if such studies produce different results from unfunded studies. The importance of gathering this information will depend on your research problem.

Experimental conditions, if any.

You will need to describe carefully the details of any experimental conditions—that is, the intervention or independent variable—if these were part of the study. This portion of the coding guide describes the relevant operations that define the experimental conditions and the categories that capture the variations in how the conditions might have been operationalized. What was experienced by people in the experimental condition? What was the intensity and duration of the intervention? As important as it is to describe what happened in the experimental condition, it is equally important to describe how the control or comparison group was treated. Was there an alternate intervention? If so, what was it? If not, what did participants in the comparison conditions do, or how were controls obtained? Differences among studies on any of these variables would be prime candidates for causes of differences in study outcomes.

Table 4.2 provides some examples of the types of information that might be gathered on a coding sheet for studies comparing students who did homework with students who did not do homework. First, note that the homework intervention code sheet gives each intervention described within the report a unique study number. This allows for the possibility that there might be more than one study of homework described in a single report, or that there might be more than one homework intervention within the same study (e.g., some students did an hour of homework, other students did a half hour, and still others did no homework at all).

Table 4.2 Example Coding Sheet for Homework Interventions
(Selected Questions)

Information About the Homework Intervention	
(Complete these questions separately for each homework intervention described in the study report.)	
I1. What is this study's ID number?	_____
I2. Which of the following characteristics were part of the homework intervention? (Place a 1 in each column that applies, 0 if not, ? if not reported.)	(Page found__)
1 = Focuses on academic work	_____
2 = Assigned by classroom teachers (or researcher via teacher)	_____
3 = Meant to be done during nonschool hours or during study time at school	_____
I3. What was the subject matter of assignments? (Place a 1 in each column that applies, 0 otherwise)	
1 = Reading	_____
2 = Other language arts	_____
3 = Math	_____
4 = Science	_____
5 = Social studies	_____
6 = Foreign language	_____
7 = Other (specify) _____	_____

14. How many homework assignments were assigned per week? (Enter ? if not reported.)	_____
15. What was the expected amount of time needed to complete each assignment, in minutes? (Enter ? if not reported.)	_____
16. Were assignments graded? 0 = No 1 = Yes ? = Can't tell	_____
17. Was the homework used in determining class grades? 0 = No 1 = Yes ? = Can't tell	_____
18. Was evidence reported that the homework intervention was or was not implemented in a manner similar to the way it was defined? (An example of when you would answer "not implemented as specified" to this question would be if the report says the homework was meant to be assigned three times a week but was assigned only once a week.) 0 = Not implemented as specified: What information was used to make this determination? _____ 1 = Implemented as specified: What information was used to make this determination? _____ ? = Nothing reported about the fidelity of implementation	_____

Table 4.2 (Continued)

I9. Was there evidence that the group receiving homework might also have experienced a changed expectancy, novelty, and/or disruption effect that the control group did not also experience? 0 = No change in expectancy, etc. 1 = Yes, change in expectancy, etc. ? = Nothing reported about change in expectancy, etc.	_____
I10. How was the comparison group treated? 0 = No homework and no other compensating activity 1 = Other compensating activity (specify) _____ ? = Not reported	_____
I11. Were the homework and comparison group drawn from the same school building and were they in the same grade? 0 = No 1 = Yes ? = Not reported	_____
I11a. If yes to I11, did the students, parents, and/or teachers in either the homework or comparison group know who was in which condition? 0 = No 1 = Yes ? = Not reported	_____

Note that the second cell of the second column of the coding sheet (Question I2) now also asks coders to give the page number on which the information was found in the report. This is a good procedure and speaks in favor of using print coding guides. This is an excellent procedure to follow when the placement of information in the report might not be clear. (I did not do this in Table 4.1 because all the information should be available on the report's title page or front matter.) Later, if coders have concerns about how they coded particular pieces of information or if two coders disagree about a code, having its location reported on the coding sheet will ease the process of finding the information for checking and can save much time. To save space, I have shown this only once in the tables, but it can appear for just about every question. If coders are working from their own copies of reports, you can also ask them to circle or highlight (in a pdf) the place in the report where information was found and put the question number from the coding guide on the report as well. It will then be easy to see where each code came from.

The next question asks whether this homework intervention meets each of three characteristics that define homework. If any of these three are answered "no," it might lead to the study being excluded from the analysis. The next five questions ask about other characteristics of assignments that the synthesists might want to test as moderators of the effects of the homework intervention. There might be more of these. Note that Question I3 (i.e., "What was the subject matter of assignments?") uses numbers to distinguish seven different coded responses and each is given a "applies," or "does not apply), or "not reported" answer. The reason for this is that a homework assignment might cover any one of the six subject matters or any combination of two or more. There are dozens of such combinations. It would be tedious for you and coders if you listed them all out, especially since you know that most of the combinations would never be coded. By coding the subject matters using just these seven codes (which still give you precise information about each study), you can then examine how frequently each combination occurs and have the computer create new variables based on the codes. For example, you might find that most studies cover only one subject matter but a few cover both reading and language arts. So, you could instruct the computer to create a new variable that has eight values, one for each instance in which only one subject matter gets a 1 and the others get a 0, a seventh in which both reading and language arts get a 1, and an eighth for all other combinations.

Question (I8) relates to the fidelity with which the homework intervention was carried out. If the way homework was actually carried out in studies was different from the intended treatment, this might raise questions about whether the study was a fair test of homework's effects. For this question, the coding sheet also provides a note that is meant to help the coder remember a coding convention that was established to clarify how to code a study. In this case, the meaning of "implemented in a similar manner" might be ambiguous, so the note clarifies its meaning. Using these notes will help ensure that different coders use the guide in the same manner, thus reducing differences between them (and also within a particular coder's responses from study to study). The next question (I9) asks whether there was evidence that the homework intervention was confounded with other differences in the way the experimental and control group were treated. If such confounds exist it would compromise the study's ability to draw causal inferences about the effects of homework. Answers to either or both of these questions might lead to the study being excluded from the synthesis, or the information might be used to group studies to see if these characteristics were associated with study outcomes.

Questions I10 and I11 relate to the control participants. Question I10 asks about how the control group was treated and Question I11 tries to get at whether the participants in each condition knew there were other participants in the study who were being treated differently. If so, this might have influenced how they behaved. Each of these questions (I10–I11) relates to the construct validity of the treatment manipulation.

Setting of the study.

This information most often includes the geographic location of the study (e.g., country, state, or part of the country; urban, suburban, or rural community). If the studies have been conducted within an institutional setting—for example, schools, hospitals, or gymnasiums—this information could be gathered as well. Furthermore, some studies will always be conducted within institutional contexts (e.g., homework studies always occur within schools), so differences in institutions might be of interest (e.g., "Was it a public or private school?," "Did the school have a religious affiliation?," etc.). Table 4.3 presents some example questions related to setting that might appear on a homework code sheet.

Participants and samples.

Another type of information typically collected from research reports concerns the characteristics of the participants included in the primary research. This can include the age, race and/or ethnic group, and social class of participants, as well as any restrictions placed by the primary researchers regarding who could participate in the study. Table 4.4 provides some examples of participant and sample characteristics that might be important in a study of the effects of homework. Also, note that yet another unique sample ID number must be provided here because some studies might present information on separate samples within the study. For example, a study might break out its samples and results based on whether students were high achieving or average. To capture this distinction, each sample would get a different sample number and Question P2 would be answered differently for each sample.

Predictor and outcome characteristics.

For studies that do not involve experimental manipulations but rather associate measured variables, one as the predictor of another (e.g., individual differences predicting rape attitudes), or for the outcomes of studies with experimental manipulations (e.g., measures of cognitive functioning by adults after aerobic exercise, or motivation after a choice manipulation), you will want to retrieve information concerning the types of outcomes and whether they were standardized measures, and evidence about the outcomes' validity or reliability, if this information is available.

Table 4.3 Example Coding Sheet for Study Setting Characteristics in a
Homework Synthesis

Setting Characteristics	
S1. Where were the participants? (Place a 1 in each column that applies, 0 if not, ? if not reported.) 1 = In the United States 2 = In a country other than the United States (specify country) _____	_____
S2. What state(s) was the study conducted in? (Use postal service code/s.)	_____
S3. What type of community was the study conducted in? 1 = Urban 2 = Suburban 3 = Rural ? = Can't tell	_____
S4. What type of school was the study conducted in? 1 = Public school 2 = Private school (secular) 3 = Private school with a religious affiliation (specify religious group) _____ ? = Can't tell	_____
S5. What classroom types were represented among the settings? (Place a 1 in each column that applies, 0 if not, ? if not reported.) 1 = Regular education 2 = Special education 3 = Other (specify) _____ 4 = No classroom types given	_____ _____ _____ _____

Table 4.5 provides some questions that might be asked about the outcomes of a homework study. Note that the first question requires, yet again, that a unique number be given to each outcome. So, we now have a four-tiered system that, when the ID numbers are strung together, uniquely identifies each outcome within each sample, each sample within each study, and each study within each report. In some studies outcomes will be reported for, say, more than one grade level or more than one measure of achievement. When such a study is uncovered, the coder would fill out separate sheets for each two-group combination. For example, a study with both a standardized test and a class grade measure of achievement reported separately for students in fifth grade and sixth grade would have four outcome coding sheets associated with it, two each for fifth and sixth graders.

Table 4.4 Example Coding Sheet for Participant and Sample
Characteristics in a Homework Synthesis

Participant and Sample Characteristics	
(Complete these questions separately for each sample within a homework intervention comparison for which there is a separate outcome.)	
P1. What is this sample's ID number?	_____
P2. Which of the following labels were applied to students in this sample? (Place a 1 in each column that applies, 0 if not, ? if not reported.)	
1 = High achieving	_____
2 = Average	_____
3 = "At risk"	_____
4 = Underachieving/below grade level	_____
5 = Possessing a learning deficit	_____
6 = Other (specify) _____	_____
P3. What was the SES of students in the sample? (Place a 1 in each column that applies, 0 if not, ? if not reported.)	
1 = Low SES	_____
2 = Low-middle SES	_____
3 = Middle SES	_____
4 = Middle-upper SES	_____
5 = Upper SES	_____
6 = Only labeled as *mixed*	_____
P5. What were the grade levels of the students in the sample? (Place a 1 in each column that applies, 0 if not. Use options 13 through 16 only if no specific grade information was reported.)	
0 = K	_____
1 = 1	_____
2 = 2	_____
3 = 3	_____

Table 4.4 (Continued)

4 = 4	_____
5 = 5	_____
6 = 6	_____
7 = 7	_____
8 = 8	_____
9 = 9	_____
10 = 10	_____
11 = 11	_____
12 = 12	_____
13 = Labeled as _elementary school_	_____
14 = Labeled as _middle school_	_____
15 = Labeled as _junior high school_	_____
16 = Labeled as _high school_	_____
17 = No grade level information given	_____
P6. What sexes were represented in the sample? (Place a 1 in each column that applies, 0 if not.)	
1 = Males	_____
2 = Females	_____
3 = No sex information given	_____
P6a. If reported, what was the percentage of females in the sample? (Use ? if not reported.)	_____

Table 4.5 Example Coding Sheet for Outcomes in a Homework Synthesis

Outcome Measure (Complete these questions separately for each relevant outcome within each sample.)	
O1. What is this outcome's ID number?	_____
O2. What subject matter did this outcome measure? (Place a 1 in each column that applies, 0 if not.) 1 = Reading 2 = Other language arts 3 = Math 4 = Science 5 = Social studies 6 = Foreign language 7 = Other (specify) _____ 8 = Not a subject matter test	 _____ _____ _____ _____ _____ _____ _____ _____
O3. What type of outcome measure is this? 1 = Standardized achievement test (specify) _____ 2 = Another test measuring achievement (e.g., teacher-developed, textbook chapter tests) 3 = Class grades after homework 4 = Multiple types of student achievement measures combined into one measure 5 = Student study habits and skills ? = Can't tell	_____
O4. Was evidence presented regarding whether the validity/reliability of this outcome measure reached an acceptable criterion? (Note: Place a 1 in each column if acceptable, 0 if not, ? if not reported. A statement indicating that internal consistency was	

Table 4.5 (Continued)

"acceptable" is sufficient, even if the specific value was not reported. A citation to an external source is sufficient.) 1 = Internal consistency 2 = Test-retest correlation 3 = Other (specify) _____	 _____ _____ _____
O5. How many days after the homework intervention was the outcome measure administered? (Enter 0 if outcome measure was given on the last day of the homework study. Enter ? if unable to determine.)	_____

Note as well that it is not just different measures of the same construct that can create multiple measures associated with the same sample (within the same study within the same report). It is also possible for researchers to collect the same measure two or more times. That is one reason why Question O5 is included on the outcome code sheet. Also, researchers might collect data on more than one construct. For example, the homework synthesis might not have focused exclusively on achievement but might also have collected outcomes related to study skills and/or attitudes toward school. If this were the case, the outcomes coding sheets would be expanded to include questions and responses related to measures of

these constructs.

The fourth question (O4) on the outcome code sheet relates to the validity and reliability of the measure. These questions can be phrased in lots of different ways, depending on the level of detail you wish to gather. The example requests information that is not very specific, asking the coders only whether the measure reached an "acceptable" level of reliability.

Coder and coding characteristics.

The coding guide should contain a section for the coders to enter their names or ID number and the date on which they coded the study (see Table 4.6). You might also ask coders to state the amount of time it took for them to code the study, for accounting purposes. In some instances, this information might be formally incorporated into your data files. This section can also provide coders with space to make any narrative comments about the coding process they want to share with you.

Table 4.6   Example Coding Sheet for Coder and Coding Information

Coder and Coding Characteristics	
C1. What is your coder ID number?	_____
C2. On what date did you complete coding this study?	___/___/___
C3. In minutes, how long did it take you to code this study?	___ ___ ___
Notes (provide below any notes about the study or concerns you had regarding your codes):	

Low- and High-Inference Codes

Most of the information requested in the example coding guides might be thought of as low-inference codes. That is, they require the coder to locate the needed information in the research report and transfer it to the coding sheet. In some circumstances, coders might be asked to make some high-inference codes about the studies. It might have occurred to you that there were some inferences that coders were asked to make on the homework coding sheets. For example, I noted previously that coders using the example guide for outcomes (Table 4.5) would be asked to code whether the estimates attained for the internal consistency, test-retest reliability, and other validity/reliability estimates for measures were "adequate" (Question O4). If left to their own devices, the judgment of adequacy would indeed be a somewhat subjective judgment, one that might vary from coder to coder. However, if you gave coders a threshold that defined "adequate," the need for judgment would have been removed from these questions. So, the question might have been rephrased to ask, "Was an estimate of internal consistency present? If yes, was it above .8?" Or the coders might have been asked to gather the exact values of the internal consistency estimates. The exact values then could be used to test whether this measure of the validity/reliability of the measures was related to study outcomes.

Other high-inference codes involve attempting to infer how an intervention or experimental manipulation might have been experienced by the individuals presented with it. A synthesis by Carlson and Miller (1987) provides a good example. They summarized the literature on why negative mood states seem to enhance the likelihood that people will lend a helping hand. In order to test different interpretations of this research, they needed to estimate how sad, guilty, angry, or frustrated different experimental procedures might have made participants feel. To do this, a group of judges were asked to read excerpts from the methods sections of relevant articles. The judges then used a 1 to 9 scale to rate the "extent to which subjects feel specifically downcast, sad, or depressed as a result of the negative-mood induction"

(p. 96). These judgments were then added to the coding sheets for each study.

These high-inference codes create a special set of problems for research synthesists. First, careful attention must be paid to the reliability of high-inference judgments. Also, judges are being asked to play the role of a research participant, and the validity of role-playing methodologies has been the source of much controversy (Greenberg & Folger, 1988). However, Miller, Lee, and Carlson (1991) empirically demonstrated that high-inference codes can lead to valid judgments and can add a new dimension to synthesists' ability to interpret literatures and resolve controversies. This technique deserves a try if you believe you can validly extract high-inference information from articles and persuasively explain your rationale for doing so (i.e., how it will increase the value of your synthesis).

Selecting and Training Coders

The coding of studies for a research synthesis is not a one-person job. Even if a single person eventually does gather information from all the studies, the research synthesists must demonstrate that this person did a good job of data extraction. There is simply too much room for bias (conscious or unconscious), for idiosyncratic interpretation of coding questions and responses, and for simple mechanical error for the unverified codes of a single person to be considered part of a scientific synthesis of research. For example, Rosenthal (1978) looked at 21 studies that examined the frequency and distribution of recording errors. These studies uncovered error rates ranging from 0% to 4.2% of all the data recorded; 64% of the errors in recording were in a direction that tended to confirm the study's initial hypothesis (see also Leong & Austin, 2006).

Recording errors are not the only source of unreliability in study coding. Sometimes, codes cannot be reliably applied because the reports of studies are not clear. Other times, ambiguous definitions provided by the research synthesists lead to disagreement about the proper code for a study characteristic. Finally, as I noted earlier, the predispositions of coders can lead them to favor one interpretation of an ambiguous code over another.

Stock and colleagues (Stock, Okun, Haring, Miller, & Kinney, 1982) empirically examined the number of unreliable codings made in a research synthesis. They had three coders (one statistician, and two post-Ph.D. education researchers) record data from 30 documents into 27 different coding categories. Stock and colleagues found that some variables, such as the means and standard deviations of the ages of participants (a low-inference code), were coded with perfect or near-perfect agreement. Only one judgment, concerning the type of sampling procedure used by the researchers, did not reach an average coder agreement of 80%.

Demonstrating that the coding definitions are clear enough to generate consistent data across coders and that the coders have extracted information from the reports accurately—that is, gave responses to the coding questions that were little different from those that would have been given by any other coder—will involve training at least two coders. Doing so is especially important if the number of studies to be coded is large or if persons with limited research training are called on to do the coding. It is rare today to find a research synthesis in which a single coder gathered information from all studies—and any such syntheses are looked on skeptically. Most syntheses involve at least two coders gathering information from at least a portion of the studies. Some syntheses involve teams of three or more coders. In any case, *it is good practice to treat the coding of studies as if it were a standard exercise in data gathering.*

Some synthesists will have every study coded independently by more than one coder, called *double coding*. The codes for every study are then compared, and discrepancies are resolved in a meeting of the coders or by a third party. This procedure can greatly reduce potential bias, make evident different interpretations of questions and responses, and catch mechanical errors.

While all synthesists must demonstrate the reliability of their codes, how far they can go to ensure reliability will be a function of the number of studies to be coded, the length and complexity of the coding guide, and the resources available to accomplish the task. Clearly, syntheses involving larger numbers of studies with complex coding sheets will require more coding time. Unless lots of time is available, more studies to code will make it more difficult to have every study coded twice. In some cases, if there is complex information to be coded, synthesists can decide to double code some of the information on the coding sheet but not other information. The synthesists must determine how to get the most trustworthy codes possible given their limited resources.

Double coding is not the only way you can enhance the reliability of codes. First, you can pick coders who have the background and interest needed to do a good job. People with lots of experience reading and conducting research make better coders than novices. Training can overcome some limitations of inexperience, but not all.

Second, coding sheets can be accompanied by coding guides that define and explain distinctions in each study characteristic. In the examples given in Tables 4.1 through 4.6, some of these definitions appear directly on the coding sheet. A coding guide with other definitions and conventions for coding particular

questions could accompany the coding sheets. The more, the better.

Third, prior to actual coding, discussions and practice examples should be worked through with coders. *It is important to pilot test your coding guide using the individuals who will actually do the coding.* Use a few research reports, preferably chosen to represent what you know are diverse types of research contained in the literature, and talk through how the coding would proceed. The coders will raise concerns you had not thought of, which will lead to greater clarity in questions, responses, and conventions to use when reports are unclear.

Fourth, the coders should gather information for the same few studies independently and share their responses in a group. You should discuss mistakes with them. Even-greater clarity in the coding guide will result. At this stage and during subsequent coding, some synthesists will attempt to keep the coders unaware of certain aspects of the studies. Some will remove information about the study's authors and affiliation from the report so that coders will not be influenced by any knowledge they may have about the researchers. Some synthesists have the different sections of the report coded by different coders so that, for example, the results of a study do not influence the ratings it might get on the quality of the study design. These procedures are more important to follow when (a) coding decisions might involve high-inference judgments, (b) the research area is distinguished by polarized opinions and findings, and/or (c) the coders are themselves very knowledgeable about the area and might have their own opinions about what the results of studies "should" be.

Estimating reliability.

Once these steps have been completed, you are ready to assess reliability. This should happen before coders are given lots of studies to code and again periodically during coding. It is usually important to obtain numerical estimates of coder reliability. There are many ways to quantify coder reliability and it appears that none is without problems (see Orwin & Vevea, 2009, for a general review of evaluating coding decisions). Two methods appear most often in research syntheses. *Most simply, research synthesists will report the agreement rate between pairs of coders.* The agreement rate is the number of agreed-on codes divided by the total number of coding opportunities. Typically, the percentage of agreement will be broken out by each coding question. If the number of codes is large, the synthesists may provide only the range of agreement percentages and then discuss any that might seem problematically low. For example, in the synthesis of studies relating choice to intrinsic motivation, we found that out of a total of 8,895 codes, there were 256 disagreements; that is, coders disagreed 2.88% of the time. The question that gave coders the most trouble involved the description of the control group, with disagreements occurring 9.4% of the time for this variable.

Also useful is Cohen's kappa, a measure of reliability that adjusts for the chance rate of agreement. The value of *kappa* is defined as the improvement over chance reached by the coders. Often kappa is presented along with the percent agreement.

As mentioned previously, some synthesists will have each study examined by two coders, will compare codes, and then will have discrepancies resolved through discussion or by consulting a third coder. This procedure leads to very high reliability; if it is used, it often is not accompanied by a quantitative estimate of reliability. In order to get an effective reliability for double coding, you would have to form two teams of two coders and an arbiter and compare the results of the two teams' deliberations. You can see that this process is unlikely to result in many differences between the teams, as long as the coding definitions are clear.

Other synthesists have individual coders mark the codes they are least confident about and discuss these codes in group meetings. This procedure also leads to highly trustworthy codes. Regardless of what techniques are used, the question to ask when evaluating the methods of data collection used to carry out research syntheses is,

> Were procedures used to ensure the unbiased and reliable (a) application of criteria to determine the substantive relevance of studies, and (b) retrieval of information from study reports?

Transferring Information to the Data File

In the foregoing paragraphs, I describe techniques for ensuring that information about each study was correctly recorded into coding sheets. I suggest that the best way to do this is to have each study coded by more than one researcher and then compare their codes to one another. Even if the coders agree on the coding sheets, *it is good practice to have two people transfer the results from the coding sheets into separate data files—the files that will be used by the computer when the data are analyzed.* Then, these files can be compared to one another to determine if any errors have been made when data were transferred from the coding sheets or placed directly into the computer. If only one coder is used, this person can be asked to do the data entry twice. Although this task may seem simple, errors in data transcription are to be expected, especially when the task data are complex. Of course, if a computer

program such as Access is used that transfers codes directly into a data set ready for the computer, this type of check is unnecessary. However, the entries into Access still need to be checked.

Problems in Gathering Data From Study Reports

In Chapter 3 I discussed some deficiencies in study retrieval that will frustrate synthesists regardless of how thorough and careful they try to be. Among these, some potentially relevant studies do not become public and defy the grasp of even the most conscientious search procedures. Other studies you will learn about but will not be able to obtain.

Perhaps the most frustrating occurrence in collecting the evidence is when synthesists obtain primary research reports but the reports do not contain the needed information. Reports could be missing information on study characteristics, preventing the determination of whether study outcomes were related to how the study was conducted, or even whether the study was relevant at all. Or information could be missing on statistical outcomes, preventing synthesists from estimating the magnitude of the difference between two groups or the relationship between two variables.

Imprecise Research Reports

Incomplete reporting will be of most concern to research synthesists who intend to perform meta-analyses. What should the meta-analyst do about missing data? Several conventions can be suggested to handle the most common problems.

Incomplete reporting of statistical outcomes.

Research reports sometimes lack important information about the results of statistical procedures carried out by the primary researchers. Statistical data are often omitted when the researcher was testing to reject the null hypothesis and it is not rejected. Instead of giving the exact results of the statistical test, the researchers simply say it did not reach statistical significance. In these cases, the researchers are also less likely to provide the correlation or means and standard deviation associated with the finding. Sometimes they do not even tell which direction the correlation or comparison of group means was in.

You have limited options when you know a relationship or comparison has been tested but the primary researchers do not provide the associated means and standard deviations, sample size, inference test value, p-level, or effect size. One option is to contact the researchers and request the information. As I noted in Chapter 3, the success of this tactic will depend partly on whether the researchers can be located as well as on the status of the requester. The likelihood of compliance with the request will also depend on how easy it is for the researchers to retrieve the information. There is less chance a request will be fulfilled if the study is old, if the desired analyses are different from those originally conducted, or if the requester asks for a lot of data.

The chance of getting a response from researchers will increase if you can make the request as easy to fulfill as possible. This might include providing the researchers with a table in which they simply need to plug in the values you need. Never ask for more information or more detailed information than you need. The more information you ask for, the more authors may worry that you think they did something wrong and suspect that you are interested in more than just including their study in a meta-analysis. (Of course, it is also important to follow up with authors if you think you have uncovered an erroneous result.)

Another approach to finding missing data is to examine other documents that describe the study being reported. For example, if you have found a journal article that reports some but not all the results you need, but the accompanying Author Note says the study was conducted as a doctoral dissertation, it might be that the dissertation itself contains the information. Often, dissertations have appendixes that include thorough descriptions of results. Or some research reports prepared by government agencies and contract research firms might be written with audiences in mind who will not be interested in the details. These organizations also might have available more technical reports with lots more information in them.

If you cannot retrieve the needed data, another option is to treat the outcome as having uncovered an exact null result. That is, for any statistical analysis involving the missing data, a correlation of 0 is assumed, or the means being compared are assumed to be exactly equal. It is reasonable to expect that this convention has a conservative impact on the results of the meta-analysis. In general, when this convention is used, the cumulative average relationship strength will be closer to zero than if the exact results of nonsignificant relationships were known. However, adding zeros to your data set for missing values will change the characteristics of your distribution of findings. For these reasons, it is rare for meta-analysts to use this procedure anymore.

A fourth option is simply to leave the comparison out of your meta-analysis. This strategy will likely lead to a higher average cumulative relationship than if the missing value was known. All else being equal, nonsignificant findings will be associated with the smaller relationship estimates in a distribution of

sampled estimates. However, most meta-analysts choose this fourth option, especially if the number of missing values is small relative to the number of known values. Also, if meta-analysts can classify missing value outcomes according to the direction of their findings—that is, if they know which group had the higher mean or whether the correlation was positive or negative—these outcomes can be included in vote count procedures (discussed in Chapter 6). It is possible to estimate the strength of a relationship using vote counts (see Bushman & Wang, 2009). Also, in Chapter 7 I will discuss ways to test meta-analytic results to see whether the conclusions would be different using different methods to handle missing data. When statisticians analyze the same data using different statistical assumptions, it is called *sensitivity analysis* (see Chapter 7).

Incomplete reporting of other study characteristics.

Research reports also can be missing information concerning the details of study characteristics other than their outcomes. For example, reports might be missing information on the composition of samples (e.g., in a homework study the students' economic background), the setting (e.g., whether the school was in an urban, suburban, or rural community), or treatment characteristics (e.g., the number of homework assignments each week and their length). Meta-analysts want this information so they can examine whether treatment effects or relationship magnitudes are associated with the conditions under which the study was conducted.

You have several options when study information of this sort is missing. First, you can ask yourself whether the information might be available in sources other than the research report. For example, the homework coding guide contains a question about whether the school was in an urban, suburban, or rural community, and a question on the students' economic status. If you know the school district in which this study took place, this information might be available on the district or state website. If information on the psychometric characteristics of measures is not reported, these might be found in reports on the instruments themselves.

Most simply, you can leave the study with missing information out of the analysis, although it may be included in other analyses for which the needed information is available. For example, homework studies missing information on the students' economic background (a frequent occurrence) simply cannot be used in the analyses testing whether this characteristic influences the effect of homework, but they can be used in analyses looking at grade level, a characteristic rarely missing from reports.

Alternatively, it is sometimes appropriate to assume that a missing value suggests what the value is. This will happen because the researchers have assumed readers will take the information for granted. For example, homework researchers are likely in nearly all instances to mention if a study was conducted in an all-boys or all-girls school. So, when the sex composition of classes is not mentioned, it is probably safe to assume that both boys and girls were present, and perhaps in roughly equal numbers. You might have coders use "?" for this code but then have the computer consider this code to mean "both boys and girls." If you do this, you should mention the convention in your methods section when you write up your synthesis. Also, if possible, you might run this analysis twice, once with the studies coded "?" included and once without.

The amount of concern a meta-analyst should have over missing study characteristics will depend partly on why the data are missing.

Some data will be completely missing at random. That is, there will be no systematic reason why some reports include information on the characteristic while others do not. If this is the case, then the outcome of an analysis examining the relationship between study outcomes and study characteristics will be unaffected by the missing data except, of course, for a loss of statistical power.

If the reason data are missing relates systematically to study outcomes, or to the values of the missing data themselves, then the problem is more serious. In this case, the missing data might be affecting the results of the analysis. For example, suppose health researchers are more likely to report that the participants in their study were all females or all males if the result indicates a significant effect of an activity intervention. Nonsignificant effects are more often associated with mixed-sex samples, but this is unknown to the meta-analyst because researchers who find nonsignificant results are less inclined to report the sample's composition. In such a case, the meta-analyst would have a hard time discovering the relationship between the sex composition of the intervention study and the magnitude of the intervention's effect (e.g., exercise is more or less effective when groups are composed of the same sex).

Pigott (2009) suggests several other strategies for dealing with missing study characteristics. First, missing values can be filled in with the mean of all known values on the characteristic of interest. This strategy does not affect the mean outcome of the cumulative analysis, except to raise its power. It is most appropriate when the meta-analyst is examining several study characteristics together in one analysis. In such a case, a single missing value may delete the entire study, which might not be desirable. Second, the missing value can be predicted using regression analysis. In essence, this strategy uses known values of the missing variable found in other studies to predict the most likely value for the missing data point.

Pigott (2009) describes several more-complicated ways to estimate missing data.

In most instances, I would advise meta-analysts to stick with the simpler techniques for handling missing data. As techniques become more complex, more assumptions are needed to justify them. Also, when more-complicated techniques are used, it becomes more important to conduct sensitivity analysis. It is always good to compare results using filled-in missing values with results obtained when missing values are simply omitted from the analysis.

Identifying Independent Comparisons

Another important decision that must be made when data are being gathered involves how to identify independent estimates of relationship strength or group differences. Sometimes a single study may contain multiple tests of the same comparison or relation. This can happen for several reasons. First, more than one measure of the same construct might be used by the researchers with measures analyzed separately. For example, a researcher of choice effects might measure intrinsic motivation using both participants' self-reports and observations of their activities during a free-play period. Second, measures of different constructs might be taken, such as several different personality variables all related to attitudes toward rape. Third, the same measure might be taken at two or more different times. And finally, people in the same study might be broken out into different samples and their data analyzed separately. This would occur, for instance, if a rape-attitude researcher gave the same measures to all participants but then separately examined results for males and females. In all these cases, the separate estimates in the same study are not completely independent—they share methodological and situational influences. In the case of the same measure taken at different times, the study results even share influences contributed by having been collected on the same people with the same measures.

The problem of nonindependence of study results can be taken even farther. Sometimes a single research report can describe more than one study conducted sequentially by the same research team in the same location. So, the two studies likely were conducted in the same context (e.g., the same laboratory), perhaps with the same research assistants, and with participants drawn from the same participant pool. Also, multiple research reports in the same synthesis often describe studies conducted by the same principal investigators. The synthesists might conclude that studies conducted by the same researchers at the same site, even if they appear in separate reports over a number of years, nevertheless contain certain constancies that imply the results are not completely independent. The same primary researcher with the same predispositions may have used the same laboratory rooms while drawing participants from the same population.

Synthesists must decide when statistical results will be considered independent tests of the problem under investigation. Several alternatives can be suggested regarding the proper unit of analysis in research syntheses.

Research Teams as Units

The most conservative way to identify independent results uses the laboratory or researcher as the smallest unit of analysis. Advocates of this approach would argue that the information value of repeated studies by the same research team is not as great as an equal number of studies reported from separate teams. This approach requires the synthesists to gather all studies done by the same research team and to come to some overall conclusion concerning the results for that particular group of researchers. Therefore, one drawback is that this approach requires the synthesists to conduct syntheses within syntheses, since decisions about how to cumulate results first must be made within research teams and then again between teams.

The research-team-as-unit approach is rarely used in practice. It is generally considered too conservative and too wasteful of information that can be obtained by examining the variations in results from study to study, even within the same laboratory. Also, it is possible to ascertain whether research teams are associated with systematic differences in study outcomes by using the researchers as a study characteristic in the search for outcome moderators.

Studies as Units

Using the study as the unit of analysis requires the synthesists to make an overall decision about the results reported in an individual study. If a single study contains information on more than one test of the same group comparison or relation, the synthesists can calculate the average of these results and have that represent the study. Alternatively, the median result can be used. Or if there is a preferred type of measurement—for example, a particular rape-attitude scale with good measurement characteristics—this result can represent the study.

Using the study as the unit of analysis ensures that each study contributes equally to the overall synthesis result. For example, a study estimating the relationship between rape attitudes and need for power using

two different attitude scales and reporting separately for men and women would report four nonindependent correlations. Cumulating these correlations (using one of the techniques suggested previously) so that a single correlation represents this study ensures equal consideration will be given to another study with one sex group and one attitude measure.

Samples as Units

Using independent samples as units permits a single study to contribute more than one result if the tests are carried out on separate samples of people. For example, synthesists could consider statistical tests on males and females within the same study of rape attitudes as independent but not consider as independent two tests that used different measures of the same attitude construct given to the same people.

Using samples as independent units assumes that the largest portion of the variance shared by results in the same study comes from data collected on the same participants. This shared variance is removed (by combining results from different measures within samples) but other sources of dependency (e.g., researchers, settings) that exist at the study level are ignored. If you expect that the study context may have a large effect on study outcomes it is best to average sample sizes within studies before combining them (Borenstein, Hedges, Higgins, & Rothstein, 2009). This is because the contribution of the study to estimates of the variance in effect sizes will differ depending on whether samples or studies are used as the unit of analysis. In Chapter 6 you will learn about fixed-effect models of error (these do not vary regardless of the unit of analysis) and random-effects models for error (these do).

It is also the case that combining results based on subsamples in one study but whole samples in another can be problematic. For example, if a study of homework provides separate results for fourth and fifth grades, the average effect of homework across the two subsamples might be different from the single effect you might have obtained if the study presented one overall result. The overall effect in the study can be obtained if the group means, standard deviations, and sample sizes are available (Borenstein et al., 2009). If you have these, you can calculate them using the *Practical Meta-Analysis Effect Size Calculator* (Wilson, 2015).

When meta-analysts calculate an average comparison or relationship across units, *it is good practice to weight each independent unit—be it a sample within a study or the entire study—by its sample size.* Then, weightings are functionally equivalent whether independent samples within studies or entire studies are used as units of analysis. For example, a study with 100 participants would be weighted by 100 if the study is used as the unit, or its two samples would each be weighted by 50 if the sample is used as the unit (more will be said about this procedure in Chapter 6).

Comparisons or Estimates as Units

The least conservative approach to identifying independent units of analysis is to use each individual group comparison or estimate of relationship strength as if it were independent. That is, each separate comparison or estimate calculated by primary researchers is regarded as an independent estimate by the research synthesist. This technique's strength is that it does not lose any of the within-study information regarding potential moderators of the studies' outcomes. Its weakness is that it is likely to violate the assumption made in the meta-analytic statistical procedures that the estimates are independent. Also, the results of studies will not be weighted equally in any overall conclusion about results. Instead, studies will contribute to the overall finding in relation to the number of statistical tests contained in them, regardless of their sample size. In the example concerning rape attitudes and the need for power, the study with four related comparisons (for two sexes on two measures) will have four times the influence on the overall results as a second study with one comparison (but an equal total sample size). This is generally not a good weighting criterion.

Shifting Unit of Analysis

A compromise approach to identifying comparisons is to employ a shifting unit of analysis. Here, each outcome is initially coded as if it were an independent event. Thus, the single study that contained four estimates of the relationship between attitudes toward rape and the need for power would have four outcome coding sheets filled out for its four results. Two of these outcome code sheets (the two measures) would be linked to two different sample code sheets (the two sexes) associated with this study. Then, when an overall cumulative result for the synthesis is calculated—that is, when the question, "What is the overall relationship between attitude about rape and the need for power?" is answered—the outcome results would first be combined so that each study (requiring that all four results be combined) or each sample (combining the two outcomes for each sample) contributed equally to the overall finding. Of course, each result should still be weighted by its sample size. These combinations would then be added into the analysis across all studies.

However, the shifting unit approach allows that when examining potential moderators of the overall

outcome, a study's or sample's results would be aggregated only *within* the separate categories of the moderator variable. An example should make this clearer. Suppose you have chosen to use studies as the basic unit of analysis. If a rape-attitude and need-for-power study presented correlations for males and females separately, this study would contribute only one correlation to the overall analysis—the average of the male and female correlations—but two correlations to the analysis of the impact of sex on the size of the correlation—one for the female group and one for the male group. To take the process one step farther, assume this study reported different correlations between rape attitudes and need for power within each sex for two different attitude measures—that is, four correlations in all. Then, the two correlations for different attitude scales would be averaged for each sex when the analysis examining the moderating influence of sex was conducted. Likewise, the two sex-related correlations would be averaged for each scale when the type of attitude measure was examined as a moderator.

In effect, the shifting-unit technique ensures that for analyses of influences on study estimates of relationship strength, a single study can contribute one data point to each of the categories distinguished by the moderating variable. This strategy is a good compromise that allows studies to retain their maximum information value while keeping to a minimum any violation of the assumption of independence of statistical tests. However, the approach is not without problems. First, creating and recreating average effect sizes for analysis of each different moderator can be time consuming and difficult in some statistical packages. Also, when the meta-analysts wish to study multiple influences on study outcomes in a single analysis, rather than one influence at a time, the unit of analysis can quickly decompose into individual comparisons.

The synthesis of studies examining correlates of rape attitudes included 65 research reports containing 72 studies with data on 103 independent samples. Primary researchers calculated a total of 479 correlations. Clearly, using the individual correlations as if they were independent results would grossly exaggerate their cumulative information value. For the overall analysis, then, the 103 independent samples were used as the unit and all correlations were averaged within samples. However, an analysis of differences in average correlations for different rape attitude scales was based on 108 correlations, because five primary researchers had given two scales to the same sample of participants.

Statistical Adjustment

Gleser and Olkin (2009) discuss statistical solutions to the problem of nonindependent tests. They propose several procedures that statistically adjust for interdependence among multiple outcomes within studies and for different numbers of outcomes across studies. The key to successfully using these techniques lies in the synthesists having credible estimates of the interdependence of the statistical tests. For instance, assume a study of correlates of rape attitudes includes both a measure of myth acceptance and victim blame. In order to use the statistical techniques, the synthesists must estimate the correlation between the two scales for the sample in this study. Data of this sort often are not provided by primary researchers. When not given, it might be estimated from other studies or the analysis could be run with low and high estimates to generate a range of values.

The Effects of Data Gathering on Synthesis Outcomes

Variation in the procedures used by research synthesists to gather information from studies can lead to systematic differences in how studies are represented in the research synthesis data set. This in turn can lead to differences in what the synthesists conclude about the literature. Variation can happen in at least three ways.

First, if the synthesists only cursorily detail study operations, their conclusions may overlook important distinctions in results. A conclusion that the synthesis results indicate no important influences on study outcomes can occur either because no such influences truly exist or because the synthesists missed representing important influences in their data set. A lack of overlap in the study details considered relevant by different synthesists studying the same problem will create variation in their conclusions. However, the notion that a synthesis leads to more trustworthy results if it includes more tests of potential influences on the overall synthesis result must be tempered by the fact that the more influences tested, the more likely it is that chance alone will lead to significant findings. So, *best practice suggests that you be judicious in your choices of what influences to test*. Still, as noted before, the coding guide should be constructed to be exhaustive; not everything coded needs to be tested.

Second, synthesists can come to different conclusions about a research literature because they code studies with different accuracy. If two syntheses vary in how carefully variables are defined and coders are trained, they likely will also vary in the number of errors in their data sets, and possibly in their conclusions because of these errors. Clearly, all else being equal, the synthesis with the more rigorous coding procedures is the one with more credibility.

And finally, the conclusions of syntheses can vary because the synthesists have used different rules for judging study results as independent tests of the problem. Here, some synthesists may place greater

importance on ensuring independence while others consider it more valuable to extract as much information as possible from their data.

Exercises

For studies on a topic of interest to you:

1. Draw up a preliminary coding guide.
2. Find several reports that describe research that is relevant to the topic.
3. Apply the coding guide to several studies, some of which you have not read before.

5 Step 4 Evaluating the Quality of Studies

What research should be included or excluded from the synthesis based on (a) the suitability of the methods for studying the synthesis question and/or (b) problems in research implementation?

Primary Function in Research Synthesis

To identify and apply criteria that separate studies conducted in ways that correspond with the research question from studies that do not

Procedural Variation That Might Produce Differences in Conclusions

Variation in criteria for decisions about study methods to include might lead to systematic differences in which studies remain in the synthesis.

Questions to Ask When Evaluating the Correspondence Between the Methods and Implementation of Individual Studies and the Desired Inferences of the Synthesis

1. If studies were excluded from the synthesis because of design and implementation considerations, were these considerations (a) explicitly and operationally defined and (b) consistently applied to all studies?
2. Were studies categorized so that important distinctions could be made among them regarding their research design and implementation?

This chapter describes

- Problems with judging the methodological adequacy of primary research
- Different approaches to describing the design and implementation of studies
- How to identify studies that report results so extreme that their exclusion from a synthesis may be warranted

---

The data evaluation stage of a scientific investigation involves making judgments about whether individual data points are trustworthy enough to be included as part of the findings. The researcher asks, "Is this data point (i.e., the study) a legitimate test of the hypothesis under consideration? Or did something happen while the study was being conducted that compromised its ability to speak to the hypothesis?" Data evaluation first requires you to establish criteria for judging the adequacy of the procedures used to gather the data for testing the relationship of interest. Next, you must examine each data point to determine if any irrelevancies or errors might have affected it. Then, you must determine whether these influences are substantial enough to dictate that the data point either should be dropped from your study or interpreted with caution.

The evaluation of individual data points must be carried out whether the data are the scores of individual participants in primary research or the outcomes of studies in a research synthesis. Both primary researchers and research synthesists examine their data looking for contaminants or other indicators that suggest the outcome for the individual participants (in primary research) or individual studies (in research synthesis) may not be trustworthy. Also, they look to see if data points are statistical outliers. This occurs when the value of a data point is so extreme relative to other scores in the data set that it is unlikely to be a member of the same population of values.

The techniques for identifying data that might be contaminated by irrelevancies are different for the two types of research. In primary research, an individual participant's responses are sometimes discarded because the researcher has evidence that the participant did not attend to the appropriate stimuli or that the response instructions were misunderstood. If deception or some other form of misdirection was used in the research, a participant's data may be discarded because the participant did not believe the cover story or deduced the hidden hypothesis.

In research synthesis, there is one important criterion, beyond the conceptual relevance of the study, for questioning the trustworthiness of data: the degree to which the study's design and implementation permit you to draw the inferences that guide your work. If a study's methods are not fully commensurate with your intended inferences, you can make either a discrete decision—whether to include the study at all—or a continuous decision—whether to weight the study's findings less than other studies. A large part of this chapter will be devoted to criteria for judging this match between the design and implementation of studies and the inferences that can be drawn from the study.

You may have noticed that I have couched my discussion of judging a study's research methods in terms of the correspondence between what inferences the methods can support and what inferences the synthesists want to make. You may have wondered, "Why not simply talk about study quality? Aren't some

studies high quality and others low quality?" The answer is that there are certainly some criteria that can be offered as indicators that one study is of "higher quality," or "better," than another. I would suggest that the trustworthiness of measurements may be such a universal criterion, though what makes a measure trustworthy may differ greatly depending on the question it is meant to answer. So, studies with more valid measures can be viewed as of higher quality regardless of whether the variable being measured is achievement, cognitive functioning, or attitude toward rape.

However, other criteria are more context dependent: they depend on the type of relationship that is under consideration. For example, a study of the effects of interventions to get adults to exercise more often that randomly assigns participants to intervention and control conditions would be a "better" study (all else being equal) of the intervention's causal effects than one that allowed participants to choose whether to engage in the intervention. Likewise, a study that begins with older adults who chose to participate in the intervention but then matched participants so that the comparison and intervention groups were roughly equivalent on important third variables is "better" for drawing causal inferences than one that used no equating procedure. On the other hand, random assignment of participants is irrelevant to a study of individual differences in rape attitudes. While we might want to know whether these individual differences cause differences in rape attitudes, no one has yet devised a method for randomly assigning people to different sexes, ages, or personalities. So, correlational studies that are of minimal value for uncovering causal relationships may be of high value for studying naturally occurring associations. The "quality" of the study depends on the question it is being used to answer.

While it is true that using the term *quality* to discuss differences between studies' methods may be good shorthand, it is not good practice if it creates the impression that one set of quality criteria can be applied to all studies regardless of the nature of the inferences called for by the problem under consideration. So, I will use the term *quality* for expository purposes, but you should keep in mind that I am using it in the sense that *high quality means high correspondence between methods and desired inferences*.

Problems in Judging Research Quality

Predispositions of the Judge

Most social scientists agree that the correspondence between methods and inferences should be the primary criterion, if not the only criterion besides substantive relevance, for decisions about how to treat a study in a research synthesis. However, the predispositions of researchers about what the outcome of studies *should* be can have a strong impact on how studies are evaluated. So, it is important to examine the sources and effects of synthesists' prior beliefs about a research area.

Almost every primary researcher and research synthesist begins an inquiry with expectations about its outcome. In primary research, methodologists have constructed elaborate controls to eliminate or minimize the role of artifacts in producing results. Most notable among these are controls for what are termed *experimenter expectancy effects*—that is, techniques to ensure that the experimenter does not treat participants in different conditions in such a manner as to increase the likelihood that the hypothesis under consideration is confirmed.

In research syntheses, protections against expectancy effects are fewer and less foolproof. As the research is being collected, coded, and evaluated, synthesists are most often aware of the outcomes of the studies they are considering. This leads to the possibility that the evaluation of a research project's methodology will be influenced by the evaluator's predisposition toward its outcomes. The impact of predispositions on syntheses was so great in the past that it is worth again quoting Glass (1976) on the old process:

> 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 or friends—and then advance the one or two "acceptable" studies as the truth of the matter. (p. 4)

Mahoney (1977) performed an experiment that directly tested the impact of predispositions on the evaluation of research. He sampled guest editors for the *Journal of Applied Behavior Analysis* and asked them to rate several aspects of a controlled manuscript. Mahoney found that the methods, discussion, and contribution of the manuscript were evaluated more favorably if the study confirmed the raters' predisposition about the results. In a related study, Lord, Ross, and Lepper (1979) found that readers rated studies that supported their attitudes as more methodologically sound than studies with counterattitudinal results. More strikingly, the undergraduates who participated in the Lord and colleagues study showed polarization in attitudes despite the fact that they all read the same research abstracts. That is, even though all participants read one study that supported their prior belief and one that refuted it, after reading the two studies participants saw more support for their initial positions. Nickerson (1998) reviews the empirical literature on confirmatory bias.

Thus, it appears that predispositions favoring a result can influence synthesists' judgments about whether a piece of research is a good test of the hypothesis under consideration. If a study disconfirms the synthesists' predispositions, they may be more likely to find some aspect of the study that renders it irrelevant or methodologically unsound. On the other hand, studies that confirm predispositions may be included although their relevance is questionable or their methods are a bad match for the hypothesis.

One way to minimize the impact of predispositions on the evaluation of research would be to have information gathered from studies by coders who are unaware of the studies' outcomes. This can be done by having separate coders unfamiliar with the topic area code different parts of the research article. For example, one coder might code the method section of a report while another coder codes the results section. However, Schram (1989) evaluated this "differential photocopying" procedure. She found it created new problems and did not lead to much higher interrater reliability.

The potential for coding study methods in a way that favors studies with results consistent with the coder's predispositions provides another reason why *it is good practice to (a) make the criteria for coding decisions explicit before coding begins, and (b) have each study coded by at least two researchers working independently* (see [Chapter 4](#)). The first of these procedures serves to minimize the chance that coders will unconsciously shift their evaluative criteria to favor studies that favor their expectations about results. The second increases the chance that if one coder's codes do reflect their predispositions, this can be caught and fixed before the studies' information is entered into the database.

Judges' Disagreement About What Constitutes Research Quality

Another problem with making quality judgments is that even disinterested judges of research can disagree on what is and is not a high-quality study. For example, numerous demonstrations have examined the reliability of evaluations made about manuscripts submitted to journals in the fields of psychology (Fiske & Fogg, 1990; Scarr & Weber, 1978), education (Marsh & Ball, 1989), and medicine (Justice, Berlin, Fletcher, & Fletcher, 1994). These studies typically calculate some measure of agreement between the recommendations made by manuscript readers concerning whether a manuscript should be accepted for publication. Typically, the level of agreement is surprisingly low.

In an interesting demonstration, Peters and Ceci (1982) resubmitted 12 published articles to the journals in which they initially appeared. The manuscripts were identical to the originals except that the names of the submitters were changed and their affiliations were changed from high-status to low-status institutions. Only three of the twelve articles were detected as being resubmissions. Of the nine articles that completed the re-review process, eight were not accepted for publication.

In many respects, the judgments of manuscript evaluators are more complex than those of research synthesists. The manuscript evaluator must consider several dimensions that do not interest the research synthesist, including the interests of the journal's readership. Also, a journal editor will sometimes deliberately choose evaluators who represent different perspectives. However, the editor still hopes that the evaluators will agree on the disposition of the manuscript. And, of course, if perfectly objective criteria were available (and were employed), the evaluators would come to concurring decisions.

Some of the differences between judgments by manuscript evaluators and research synthesists were controlled in a study conducted by Gottfredson (1978). He removed much of the variability in judges' ratings that might be due to differing initial biases by asking authors to nominate experts competent to evaluate their work. Gottfredson was able to obtain at least two expert evaluations for each of 121 articles. The experts evaluated the quality of the articles on a three-question scale that left the meaning of the term *quality* ambiguous. An interjudge agreement coefficient of $r = .41$ was obtained. On a 36-item evaluation scale, which tapped many explicit facets of research quality, an interjudge agreement coefficient of $r = .46$ was obtained. These levels of agreement are probably lower than we would hope.

Why do overall judgments of quality differ? In addition to differences in the judges' predispositions, it is possible to locate two other sources of variance in quality judgments: (a) the relative importance judges assign to different research design characteristics and (b) how well judges think a particular study met a particular criterion. To demonstrate the first source of variance, I conducted a study in which six experts in school desegregation research were asked to rank the importance of six design characteristics for establishing the utility or information value of a school desegregation study (Cooper, 1986). The six characteristics were (a) the experimental manipulation (or in this case, the definition of desegregation), (b) the adequacy of the control group, (c) the validity of the outcome measure, (d) the representativeness of the sample, (e) the representativeness of the environmental conditions surrounding the study, and (f) the appropriateness of the statistical analyses. The intercorrelations of the rankings among pairs of experts varied from $r = .77$ to $r = -.29$, with the average correlation being $r = .47$. Thus, it was clear that judges differed in how important they thought different evaluative criteria were, even before applying them to particular studies.

In sum, the studies of judges' assessments of methodological quality indicate evaluator agreement is less than we would like. *One way to improve the reliability of quality judgments would be to add more judges*

to the evaluation of any given study. So, for example, a rating of a study (or the decision to include or exclude a study from a research synthesis) based on five judges' average ratings will correspond better with the average rating of five other judges (drawn from the same population of judges) than will the ratings of two judges with any other two judges. However, it is rare that such large pools of judges can be used to make quality judgments about studies in a research synthesis. Two or three is often the practical limit.

Differences Among Quality Scales

I mentioned earlier that two sources of variance in judges' ratings were (a) the relative importance they assign to different research design characteristics and (b) how well they think a particular study met a particular criterion. A technique many research synthesists have used in an attempt to address the first of these involves the use of quality scales. Here, the synthesists use a predeveloped scheme that tells judges what evaluative dimensions are important. Also, the scales typically use a prearranged weighting scheme so that the same weight is placed on a dimension of quality when it is applied to each study. The synthesists hope that by asking coders to apply the same set of explicit criteria, it will lead different coders to more transparent and consistent ratings. The goal of quality scales is to take some of the subjectivity out of the rating process.

While certainly an improvement over allowing each judge to determine his or her own quality criteria, quality scales have met their goals with limited success. In medical research, Jüni, Witshci, Bloch, and Egger (1999) demonstrated that quality scales may create consistency among those using the same scale but it does not mean that different scales will come to the same judgments. Jüni and colleagues applied 25 different scales (constructed by other researchers) to the same set of studies and then conducted 25 meta-analyses, one using each scale. They found that the conclusions of the meta-analyses differed dramatically depending on the scale that was used. For six of the scales, the high-quality studies suggested no difference between a new and old treatment, whereas the low-quality studies suggested a significant positive effect for the new treatment. The pattern was reversed for seven other quality scales. The remaining 12 quality scales resulted in conclusions that showed no difference between the results of high- and low-quality studies. Thus, even though the quality scales may improve somewhat the reliability of judges using the same scale, the validity of the conclusions they lead to is still suspect.

Jeffrey Valentine and I (Valentine & Cooper, 2008) suggested several reasons why the quality scales seem to lead to such poor agreement. First, just as individual judges can disagree about what characteristics of research are important for quality judgments, quality scales do not necessarily agree about this either. For example, in the Jüni and colleagues study, some of the scales focused almost exclusively on the studies' ability to permit causal inferences; other scales addressed multiple characteristics, such as the representativeness of the sample and statistical power.

Second, most quality scales still leave room for judges' subjective assessments to enter the evaluation process. The scales use terms such as *adequate, appropriate,* and *sufficient* to describe design features (e.g., "Was the internal consistency of measures adequate?"), without providing operational definitions for these adjectives. What may be adequate for one judge may be inadequate for another. This means that even though the important characteristics are identified, the reliability of codes for any single dimension will still be less than perfect. It also suggests that while identifying the characteristics makes judgments more transparent, it is still not perfectly clear what the criteria are for applying each of the evaluative labels.

Third, similar to individual researchers, most scales apply different schemes to weight the importance of different methodological characteristics. Typically, quality scales assign a certain portion of the scale's points to each of the characteristics. So, even when scales use the same characteristics, there can be variation among scales about the importance that should be assigned to each characteristic. For example, Jüni and colleagues found that some scales assigned 16 times more weight than other scales to the same design feature. Part of this difference was explainable by the fact that the scales used different numbers of design features and part to the fact that the scale developers might have valued the same design feature differently.

Reliance on a single score to express quality.

Typically, the scores from the various items on a quality scale are summed to a single score. These were the scores that Jüni and colleagues (1999) used to categorize studies into high- and low-quality groupings when they did their 25 meta-analyses. Valentine and Cooper (2008) questioned whether it makes sense to reduce the evaluation of a study to a single dichotomous judgment (Is this study good or bad?) or even to a single continuous judgment (What is this study's quality score?). The single-score approach results in one number that is summed across very different aspects of study design and implementation, many of which are not necessarily related to one another. For example, there is no necessary relationship between the process used to assign participants to experimental conditions and the quality of outcome measures used in a study. So, one study of homework might randomly assign participants to conditions but use a self-report of grades as the measure of achievement. A second study might use matching of students who

did and did not do homework on their own but use grades taken from student records. In such a case, the first study has a stronger design for making causal inferences but the second study has a more valid outcome measure. When a scale combines these two elements of study design into a single score, it may obscure these important differences between them; these two studies might get identical or similar scores. If the two studies produce different results, how should we interpret this difference?

A Priori Exclusion of Research Versus A Posteriori Examination of Research Differences

The role of predispositions and the disagreement about what characteristics of research design define quality demonstrate how subjectivity intrudes on our attempts to be scientifically objective. The point is important because research synthesists often debate whether or not a priori judgments of research quality should be used to exclude studies from their work. This debate was first captured in an exchange of views between Hans Eysenck (1978) and Gene Glass and Mary Smith (1978) concerning Smith and Glass's (1977) early meta-analysis of research on psychotherapy. Smith and Glass synthesized over 300 studies examining the effectiveness of psychotherapy with no a priori exclusion of studies due to methodology. Eysenck felt this strategy represented an abandonment of scholarship and critical judgment:

> A mass of reports—good, bad, and indifferent—are fed into the computer in the hope that people will cease caring about the quality of the material on which the conclusions are based. . . . "Garbage in—garbage out" is a well-known axiom of computer specialists; it applies here with equal force. (p. 517)

Eysenck (1978) concluded that "only better designed experiments than those in the literature can bring us a better understanding of the points raised" (p. 517).

Glass and Smith (1978) made several points in rebuttal. First, they argued that the poor design characteristics of different studies can cancel one another out, if the results of different studies are consistent. Second, as noted earlier, they stated that the a priori quality judgments required to exclude studies are likely to vary from judge to judge and be influenced by personal biases. Finally, Glass and Smith claimed they did not advocate the abandonment of quality standards. Instead, they regarded the impact of design quality on study results as "an empirical *a posteriori* question, not an *a priori* matter of opinion" (Glass et al., 1981, p. 222). They suggested that synthesists thoroughly code the design aspects, good and bad, of each study and then determine empirically (through meta-analysis) if, in fact, the outcomes of studies are related to how the studies were conducted.

I suspect that the best approach to the debate about when, if ever, to exclude studies from a research synthesis is best resolved by a combination of approaches. Generally, the decision to include or exclude studies on an a priori basis requires you to make an overall judgment of study quality that is often subjective and that others may find unconvincing. But there could be instances in which so many high-quality studies exist that low-quality studies can be dismissed without concern. This was done, for example, in our synthesis of aerobic exercise effects on cognitive functioning. Enough studies that employed random assignment existed that we could focus on these studies alone. Our homework synthesis found very few random assignment studies so we needed to include studies with less ability to draw strong causal inferences. In this instance we looked at whether students being randomly assigned to homework conditions was associated with the study's results. Thus, it is important that you ask this question about how a research synthesis was conducted:

> If studies were excluded from the synthesis because of design and implementation considerations, were these considerations (a) explicitly and operationally defined and (b) consistently applied to all studies?

Generally, however, it is good practice to carefully enumerate study characteristics and to compare the results of studies that used different methods to determine if studies' methods and results co-vary with one another. If it is empirically demonstrated that "good" studies (i.e., studies that permit inferences most correspondent with the inferences you wish to make) produce results different from "bad" studies (i.e., studies with methods inconsistent with the intended inferences), the results of the good studies ought to be believed. In this case, no harm is done to the validity of your inferences by looking at the "bad" studies and perhaps something is learned about how to conduct future research. When no difference in results is found, it seems sensible to retain (some or all of) the "bad" studies because they contain other variations in methods (such as different samples and locations) that, by their inclusion, may help you answer many other questions surrounding the problem area. In most cases, letting the data speak—that is, including nearly all studies and examining empirically the differences in results associated with methods—substitutes a discovery process for the predispositions of the synthesist. I will return to this issue again after I suggest a scheme for coding the methodological characteristics of studies.

Approaches to Categorizing Research Methods

The decision to examine empirically the impact of methodology on research results does not relieve you of all evaluation responsibilities. You must still decide what methodological characteristics of studies need to be coded. As I pointed out previously, these decisions will depend on the nature of the question under scrutiny and the types of associated research. If a problem has been addressed mainly through experimental manipulations in laboratory settings—for example, the effects of choice on intrinsic motivation—a different set of methodological characteristics of studies will be important than if correlational designs are at issue, as in studies of the relationship between individual differences and attitudes about rape. Two broad approaches to coding research methods can be identified, though they are rarely used in their pure form. The first approach requires the synthesist to make judgments about the threats to validity that exist in a study. The second approach requires the detailing of the objective design characteristics and other methods of a study, as described by the primary researchers.

Threats-to-Validity Approach

When Campbell and Stanley (1963) introduced the notion of threats to validity, they literally transformed the social sciences. They suggested that a set of extraneous influences associated with each research design could be identified that "might produce effects confounded with the experimental stimulus" (p. 5). Different research designs had different validity threats associated with them. Designs could be compared according to their inferential capabilities. More importantly, less-than-optimal designs could be triangulated so that strong inferences could result from multiple studies when the single "perfect" study could not be performed.

Campbell and Stanley's (1963) notion held the promise of increased sensitivity and objectivity in discussions of research quality. However, it was not long before some problems in the application of their scheme became apparent. The problems related to creating an exhaustive list of threats to validity and identifying what the implication of each threat might be.

Initially, Campbell and Stanley (1963) proposed two broad classes of validity threats: internal validity and external validity. First, threats to internal validity related to the causal correspondence between the experimental treatment and the experimental effect. To the extent that this correspondence was compromised by deficiencies in research design, the ability to interpret a study's results as evidence of a causal relationship would be called into question. Campbell and Stanley listed eight threats to internal validity. The second broad class, threats to external validity, related to the generalizability of research results. Evaluating external validity required assessing the representativeness of a study's participants, settings, treatments, and measurements. While the external validity of a study could never be assessed definitively, Campbell and Stanley suggested four classes of threats to representativeness.

Bracht and Glass (1968) offered an expanded list of threats to external validity. They believed that "external validity was not treated as comprehensively as internal validity in the Campbell-Stanley chapter" (p. 437). To rectify this omission, Bracht and Glass identified two broad classes of external validity: (a) population validity, referring to generalization to persons not included in a study, and (b) ecological validity, referring to settings not sampled. They described two specific threats to population validity, along with 10 threats to ecological validity.

Campbell (1969) added an additional threat to internal validity, called *instability,* defined as "unreliability of measures, fluctuations in sampling persons or components, autonomous instability of repeated or equivalent measures" (p. 411).

Next, Cook and Campbell (1979) offered a list of 33 specific threats to validity grouped into four broad classifications: to the notions of internal and external validity they added the notions of construct validity and statistical conclusion validity. The term *construct validity* referred to "the possibility that the operations which are meant to represent a particular cause or effect construct can be construed in terms of more than one construct" (p. 59). The term *statistical conclusion validity* referred to the power and appropriateness of the data analysis technique. And finally, Shadish et al. (2002) updated the list of threats categorized into the four broad classifications.

From this brief history, the problems in using a strict threats-to-validity approach to assess the quality of empirical studies should be clear. First, different researchers may use different lists of threats. For instance, should the threat of instability offered by Campbell (1969) constitute one threat, as originally proposed, or several threats (e.g., low statistical power, unreliability of measures), as redefined by Shadish et al. (2002)? Or should ecological validity constitute one threat or up to 10 different threats? A second problem is the relative weighting of threats: Is the threat involving historical confounds (other societal events that happened concurrent with the experimental manipulation) weighted equally with the threat involving restricted generalizability across constructs? Expert methodologists may even disagree on how a particular threat should be classified. For instance, Bracht and Glass (1968) listed experimenter expectancy effects as a threat to external validity while Shadish et al. (2002) listed it as a threat to the construct validity of causes.

All these problems aside, the threats-to-validity approach to the evaluation of research still represents an improvement in rigor, and is certainly preferable to the a priori single judgment of quality it replaces. Each successive list of threats represents an increase in precision and a deepening understanding of the relationship between research design and inferences. Also, the list of validity threats gives the synthesist an explicit set of criteria to apply or modify. In that sense, synthesists who use the threats-to-validity approach make their rules of judgment open to criticism and debate. This is a crucial step in making the research evaluation process more objective.

Methods-Description Approach

In the second approach to evaluating study design and implementation, the synthesist codes the objective characteristics of each study's methods as they are described by the primary researchers. For example, experimental designs—how comparisons between groups treated differently are constructed—relate mainly to eliminating threats to internal validity. Campbell and Stanley (1963) described three preexperimental designs, three true experimental designs, and ten quasi-experimental designs, and the list of design variations has been expanded several times since (see Shadish et al., 2002; or May, 2012). So, in this approach, rather than evaluate the internal validity of a study's design—an abstract assessment that could lead to disagreement—the coder simply retrieves the design type by matching the design used in the study to a design on a list of possibilities. This is a low-inference code that should be fairly consistent across coders; when inconsistencies appear, it should be fairly easy to resolve disagreements. In most areas of research, considerably fewer than all the available designs will be needed to describe exhaustively how comparisons were constructed in the relevant research.

One problem with the methods-description approach to evaluating studies is shared with the threats-to-validity approach (and was evident in the use of quality scales): different synthesists may choose to list different methodological characteristics. So, while the methods-description approach may lead to more reliable coding, it does not solve the problem of what characteristics to code in the first place.

Another problem with the methods-description approach is that the list of methodological characteristics that might need to be coded can become extremely long. Remember, there are four classes of threats to validity—internal, external, construct, and statistical validity—and each may require the coding of numerous design and implementation characteristics to capture every aspect of methodology that might influence whether the threat is a concern in a given study. And it may not be advisable to test every one of these characteristics as a moderator of study results: because the number of tests would be so large, some will be significant by chance alone (i.e., the Type I error rate will be inflated). So, there may be a trade-off between the threats-to-validity and methods-description approaches involving parsimony versus reliability.

A judgment about the threat to validity called *low statistical power* (related to statistical conclusion validity) provides a good example. A coder making an overall judgment about whether a study has a good chance to reject a false null hypothesis must do so by combining several explicit study characteristics: size of the sample, whether a between- or within-subjects design was used, the inherent power of the statistical test (e.g., parametric versus nonparametric), the number of other sources of variance extracted in the analysis, and the expected size of the relationship under study. Using the threats-to-validity approach, two coders of the same study might disagree on whether a study has low statistical power because they weighted these factors differently or perhaps failed to consider the same factors. However, they might agree perfectly on their codes of these separate components. This speaks in favor of using the methods-description approach. But using the methods-description approach still leaves room for subjectivity: When is sample size too small for adequate statistical power? And, if the number of codes required becomes too large (I listed five for just one of dozens of threats to validity), relating them all to study outcomes jeopardizes the validity of the results of the research synthesis; with so many tests, a few will be significant by chance. If chance is playing a role in generating significant findings in a meta-analysis, the pattern of results will be difficult to interpret. This speaks in favor of using the threats-to-validity approach.

A Mixed-Criteria Approach: The Study DIAD

The pros and cons of the two approaches makes one wonder if there is a way to combine them that maximizes the strengths and minimizes the weaknesses of each. In such a strategy, you might code many of the potentially relevant aspects of research methodology and perhaps build a scheme for explicitly combining them into judgments about the different validity threats, not unlike my previous example regarding statistical power. Some threats to validity might have to be coded directly. For instance, the threats to internal validity involving aspects of the control group—diffusion of treatments, compensatory rivalry, or resentful demoralization of participants receiving the less desirable treatment—are probably best coded directly as threats to validity, though deciding whether they are present or absent still relies heavily on the description of the study presented by the primary researcher (also, I did this in the example coding sheet Table 4.2 Question I9). While this mixed-criteria approach does not remove all problems from evaluating studies (I will describe several in the following paragraphs), it would be a step

toward the use of explicit and objective quality criteria that also takes into account the utility of the resulting descriptions of studies.

Jeffrey Valentine and I (Valentine & Cooper, 2008) attempted to create an instrument for use in research synthesis that evaluated studies using this mixed-criteria approach. The result was called the Study Design and Implementation Assessment Device, or the Study DIAD. The Study DIAD provides a framework for synthesists to build an evaluative scale unique to their topic area and allows them to choose from several different levels of abstractness for describing the correspondence between a study's methods and desired inferences. However, it requires the user (a) to be detailed and explicit about the chosen criteria, (b) to define these criteria prior to beginning the evaluation of studies, and (c) to apply the criteria consistently across all studies. The Study DIAD is based on the assumption that the user wants to draw causal inferences about the effectiveness of an intervention: For example, do interventions meant to promote aerobic exercise for adults cause improvements in participants' cognitive functioning? However, it is divided into sections corresponding to the four classes of validity, so it can be used for other types of research as well. A full exposition of the Study DIAD is available elsewhere (see Valentine & Cooper, 2008), but a brief introduction should give you an idea of how it combines the threats-to-validity and methods-description approaches, how it works, and whether using it might be appropriate for your synthesis.

At the most abstract level, the Study DIAD provides the user with answers to four global questions relating to the construct, internal, external, and statistical conclusion validity of a study:

1. Fit Between Concepts and Operations: Were the participants in the study treated and the outcomes measured in a way that is consistent with the definition of the intervention and its proposed effects?
2. Clarity of Causal Inference: Did the research design permit an unambiguous conclusion about the intervention's effectiveness?
3. Generality of Findings: Was the intervention tested on participants, settings, outcomes, and occasions representative of its intended beneficiaries?
4. Precision of Outcome Estimation: Could accurate estimates of the intervention's impact be derived from the study report?

The term *intervention* is used in the Study DIAD to stand for any treatment or experimental manipulation. So, all four of these questions would be relevant to our examples involving research on the effectiveness of homework, programs to increase aerobic exercise among adults, and the effects of choice on intrinsic motivation; they all seek to uncover causal relationships. However, because the studies of choice and intrinsic motivation were conducted in laboratories (with great experimental control), all such studies should have good internal validity, so Question 2, "Clarity of Causal Inference," likely could be omitted from the Study DIAD for this synthesis. Our fourth example, individual differences in attitudes toward rape, is not concerned with causal relationships, so Question 2 also would be irrelevant to assessing the correspondence between study methods and inferences in that research synthesis. The other global questions on the Study DIAD are relevant to all the examples.

At a slightly more specific level, the Study DIAD decomposes the four global questions into eight composite questions. These are presented in Figure 5.1. Here, the four global questions are each divided into two more-specific questions. It might have occurred to you that the four global or eight composite questions could be used to form a quality scale by themselves. In other words, judges might simply be asked to answer each of these questions for each study (or to give the study a score on a continuous measure). This would be an example of a pure threats-to-validity approach to quality assessment and would exhibit the strengths and weaknesses associated with such an approach.

But the Study DIAD goes a step farther by attempting to operationally define the methodological characteristics of studies that go into answering each of the eight composite and four global questions. Accomplishing this task requires that the instrument (a) identify the particular design and implementation features that must be considered when answering each of the eight composite questions and (b) provide a way (an algorithm) to sum up the chosen positive and negative methodological features to get to the answers to the eight (and then the four) questions. To do this, the Study DIAD requires coders to answer about 30 questions regarding a study's design and implementation. These are presented in Table 5.1 with the question number associated with each indicating which global and composite question that particular methodological feature is related to.

Looking over the questions in Table 5.1, you might wonder how we decided which aspects of design and implementation should be represented in the Study DIAD. Here, we faced the same problem encountered by everyone who attempts to develop a quality scale. To make these decisions, we first considered the content of other scales and many methods textbooks and articles. We then shared early drafts of the Study DIAD with highly regarded research methodologists and sought input on the instrument at a public meeting and on a website. So, the consensus surrounding the 30-odd questions about study design and implementation on the Study DIAD probably is higher than most sets of such questions used in most quality scales.

Figure 5.1 Eight Composite Questions About Study Quality Taken From the Study DIAD

1. Fit Between Concepts and Operations		2. Clarity of Causal Inference	
1.1 Were the participants treated in a way that is consistent with the definition of the intervention?	1.2 Were the outcomes measured in a way that is consistent with the proposed effects of the intervention?	2.1 Were the participants in the group receiving the intervention comparable to the participants in the comparison group?	2.2 Was the study free of events that happened at the same time as the intervention that confused its effect?

3. Generality of Findings		4. Precision of Outcome Estimation	
3.1 Did the study include variation on participants, settings, and outcomes representative of the intended beneficiaries?	3.2 Was the intervention tested for its effect within important subgroups of participants, settings, and outcomes?	4.1 Were effect sizes and their standard errors accurately estimated?	4.2 Were statistical tests adequately reported?

Table 5.1 Questions About Study Design and Implementation From the Study DIAD

1.1	Were the participants treated in a way that is consistent with the definition of the intervention?
1.1.1	To what extent does the intervention reflect commonly held or theoretically derived characteristics about what it should contain?
1.1.2	Was the intervention described at a level of detail that would allow its replication by other implementers?
1.1.3	Was there evidence that the group receiving the intervention might also have experienced a changed expectancy, novelty, and/or disruption effect not also experienced by the comparison group (or vice versa)?
1.1.4	Was there evidence that the intervention was implemented in a manner similar to the way it was defined?
1.2	Were the outcomes measured in a way that is consistent with the proposed effects of the intervention?
1.2.1	Do items on the outcome measure appear to represent the content of interest to this synthesis (i.e., have face validity)?
1.2.2	Were the scores on the outcome measure acceptably reliable?
1.2.3	Was the outcome measure properly aligned to the intervention condition?
2.1	Were the participants in the group receiving the intervention comparable to the participants in the comparison group?
2.1.1	Was random assignment used to place participants into conditions? (If no, answer the next questions [Question 2.1.1a].)
2.1.1a	For quasi-experiments: Were adequate equating procedures used to recreate the selection model?
2.1.2	Was there differential attrition between intervention and comparison groups after equating occurred?

2.1.3	Was there severe overall attrition after equating occurred?
2.2	Was the study free of events that happened at the same time as the intervention that confused its effect?
2.2.1	Was there evidence of a local history event?
2.2.2	Were the intervention and comparison groups drawn from the same local pool? (If yes, answer the next question [Question 2.2.2a].)
2.2.2a	If yes, were intervention conditions known to study participants, providers, data collectors, and/or other authorities?
2.2.3	Did the description of the study give any other indication of the strong plausibility of other intervention contaminants?
3.1	Did the study include variation on participants, settings, and outcomes representative of the intended beneficiaries?
3.1.1	Did the sample contain participants with the necessary characteristics to be considered part of the target population?
3.1.2	To what extent did the sample capture variation among participants on important characteristics of the target population?
3.1.3	To what extent did the study include variation on important characteristics of the target setting?
3.1.4	To what extent were important classes of outcome measures included in the study?
3.1.5	Did the study measure the outcome at a time appropriate for capturing the intervention's effect?
3.1.6	Was the study conducted during a time frame appropriate for extrapolating to current conditions?
3.2	Was the intervention tested for its effect within important subgroups of participants, settings, and outcomes?
3.2.1	To what extent was the intervention tested for effectiveness within important subgroups of participants?

Table 5.1 (Continued)

3.2.2	To what extent was the intervention tested for effectiveness within important subgroups of settings?
3.2.3	Was the intervention tested for its effectiveness across important classes of outcomes?
4.1	Were effect sizes and their standard errors accurately estimated?
4.1.1	Was the assumption of independence met, or could dependence (including dependence arising from clustering) be accounted for in estimates of effect sizes and their standard errors?
4.1.2	Did the statistical properties of the data (e.g., distributional and variance assumptions, presence of outliers) allow for valid estimates of the effect sizes?
4.1.3	Were the sample sizes adequate to provide sufficiently precise estimates of effect sizes?
4.1.4	Were the outcome measures sufficiently reliable to allow adequately precise estimates of the effect sizes?
4.2	Were statistical tests adequately reported?
4.2.1	To what extent were sample sizes reported (or estimable) from statistical information presented?
4.2.2	To what extent could directions of effects be identified for important measured outcomes?
4.2.3a	To what extent could effect sizes be estimated for important measured outcomes?
4.2.3b	Could estimates of effect sizes be computed using a standard formula (or algebraic equivalent)?

That said, you next might also recognize that the questions in Table 5.1 still involve some degree of judgment on the part of the coder; the 30 questions still include terms like *adequate* and *fully*. So, the Study DIAD goes yet a step farther by requiring the users, before the instrument is applied, to more precisely define the terms listed in Table 5.1 that otherwise would be open to varying interpretations. The mechanism for doing this is presented in Table 5.2, which contains a document that needs to be completed by the research synthesists *before* the Study DIAD can be applied. The first column contains the request for the definitions of terms. Note that these definitions are specific to the research area under consideration. Some are highly related to the specific content area, such as the important characteristics of the intervention (Question 1 in Table 5.2). Others might be a bit more general but still might vary as functions of the topic. For example, the criteria for the minimum acceptable attrition (i.e., loss of participants from the study, i.e., Questions 12 and 13 on attrition) probably have some general boundaries but also might be different for studies on, say, the effectiveness of homework and studies on the effectiveness of an aerobic exercise intervention.

Table 5.2 also shows in the second column the composite question that each answer applies to. In the third column, the table shows you how the questions might be answered by synthesists who were going to apply the Study DIAD to studies about the effectiveness of homework. So, for example, the coders applying the Study DIAD to each study of homework are not asked to make a decision about what a minimally acceptable internal consistency is for outcome measures (Question 4 in Table 5.2). Instead, the

coders are told that the principal investigators have chosen .60 to be the minimal acceptable level of this type of reliability. In this way, all of the judgments in the eight composite questions are given operational definitions.

In the final steps for using the Study DIAD, a set of algorithms are applied to the answers to the design and implementation questions (Table 5.1) so that they are combined to answer the eight composite questions in Figure 5.1. Table 5.3 presents one of these eight algorithms. Algorithms also exist for combining the eight composite questions into the four global questions. As an example, Table 5.4 shows the results of applying the Study DIAD to a study by McGrath (1993) on the effects of homework on academic achievement.

Table 5.2 Contextual Questions That Need Answers From the Users in Order to Apply the Study DIAD

Contextual Question	Study DIAD Composite Question	Example of Answer to Questions for Evaluating Research on the Effects of Homework on Academic Achievement
1. What commonly shared and/or theoretically derived characteristics of the intervention should be present in its definition and implementation?	Fit Between Concepts–Intervention	• Focuses on academic work • Assigned by classroom teacher (or researcher via teacher) • Meant to be done during nonschool hours or during study time at school
i. Which of these characteristics are necessary to define interventions that "fully," "largely," and "somewhat" reflect commonly shared and/or theoretically derived characteristics?		• For "fully," all three must be present • There will be no "largely" or "somewhat" studies
ii. What variations in the intervention are important to examine as potential moderators of effect size?		• Frequency of assignments • Expected amount of time spent on each assignment • Subject matter(s) covered • Degree of individualization • Compulsory versus voluntary nature of assignments • Purpose

Contextual Question	Study DIAD Composite Question	Example of Answer to Questions for Evaluating Research on the Effects of Homework on Academic Achievement
		○ Practice (reinforcement, rehearsal) ○ Preparation (introduce a new skill) ○ Integration (combine two skills) ○ Extension (apply to a new content area) ○ Enrichment • Amount of time to complete • Individual or group assignment
2. What important characteristics of the intervention would we need to know in order to reliably replicate it with different participants, in other settings, at other times?	External Validity	• Frequency of assignments • Expected amount of time spent on each assignment • Actual amount of time spent on each assignment • Subject matter(s) covered • Grade level(s) of students • Degree of individualization • Compulsory versus voluntary nature of assignments • Individual or group assignment

Table 5.2 (Continued)

Contextual Question	Study DIAD Composite Question	Example of Answer to Questions for Evaluating Research on the Effects of Homework on Academic Achievement
3. What are the important classes of outcomes?	Fit Between Concepts–Outcomes	- Achievement tests - Standardized tests - Other test - Class grades - Study habits and skills - Student attitudes toward - School - Subject matter - Student self-beliefs - Parent attitudes toward school
i. What classes of outcomes are needed to conclude that a reasonable range of operations and/or methods have been included and tested?		Any two classes of outcomes is a reasonable range
4. Does the synthesist have a minimum level of score reliability for outcomes to be considered in the review?	Fit Between Concepts–Outcomes	Yes

Contextual Question	Study DIAD Composite Question	Example of Answer to Questions for Evaluating Research on the Effects of Homework on Academic Achievement
If so, what are the specific minimum reliability coefficients for internal consistency, temporal stability, and/or interrater reliability (as appropriate)?		Internal consistency estimate > .60
5. Considering the context of this study, during what interval of time should it have been conducted to be relevant to current conditions?	External Validity–Sampling	1982–2007
6. Considering the context of this study, what are the characteristics of the intended beneficiaries of the intervention?	External Validity–Sampling	- K–12 students - Students in the United States, Canada, United Kingdom, Australia
7. What are the important characteristics of participants that might be related to the intervention's effect and must be equated if a study does not use random assignment?	Causal Inference–Selection	- Pretest of outcome OR prior achievement - Grade level OR age - Socioeconomic status
8. What characteristics of subgroups of participants are important (a) to have variation on and (b) to test within a study to determine whether an intervention is effective within these groups? What levels or labels capture this variation?	External Validity–Effects Tested Within Subgroups	- Achievement labels applied to students - Gifted, average, "at-risk," learning disabled, underachieving/below grade level, possessing a learning deficit

Table 5.2 (Continued)

Contextual Question	Study DIAD Composite Question	Example of Answer to Questions for Evaluating Research on the Effects of Homework on Academic Achievement

		• Grade levels
		  ○ K–12
		• Socioeconomic levels
		  ○ Low
		  ○ Low-middle
		  ○ Middle
		  ○ Middle-upper middle
		  ○ Upper
		• Student sex

i. Which of these characteristics of subgroups of participants are needed to conclude that a "limited" or "reasonable range" of characteristics have been included and tested?		• Any one characteristic for "limited" • Any three characteristics for "reasonable"
9. What characteristics of settings are important to test within a study to determine whether an intervention is effective within these groups?	External Validity–Effects Tested Within Subgroups	In School • Class size • Special versus regular education classrooms

Contextual Question	Study DIAD Composite Question	Example of Answer to Questions for Evaluating Research on the Effects of Homework on Academic Achievement

		• Classroom preparation
		• Provision of materials
		• Teacher-suggested approaches to work
		• Teacher-provided links to the curriculum
		• Start in class and finish at home
		• Feedback
		
		  ○ Written comments
		  ○ Grading
		  ○ Incentives
		  ○ Used as part of class grade
		
		• Alignment with academic content
		• Use in classroom discussion
		
		At Home
		
		• Socioeconomic status of family
		• Number and type of siblings
		• Number of adults in the home

i. Which of these characteristics and settings are needed to conclude that a "full," "reasonable," or "limited" range of variations have been tested?		• All for "full" • Two from "school" and one from "home" for "reasonable" • Any one for "limited"

Table 5.2 (Continued)

Contextual Question	Study DIAD Composite Question	Example of Answer to Questions for Evaluating Research on the Effects of Homework on Academic Achievement
10. What is the appropriate interval for measuring the intervention's effect relative to the end of the intervention?	External Validity–Inclusive Sampling	• Any interval is appropriate
11. Considering the context of this study, for purposes of sampling, what constitutes the local pool of participants?	Internal Validity–Lack of Contamination	• Students attending the same school building at the same grade level
a. If participants are drawn from the same local pool, which groups of individuals (e.g., students, teachers, parents, administrators, caseworkers) might have been able to interfere with the fidelity of the comparison group if they knew who was in the intervention and comparison groups?		• Students in either group • Parents of students in either group • Teachers
12. For research on this topic, how would you define differential attrition from the intervention and control groups?	Internal Validity–Selection	• More than a 10% difference in attrition percentages between groups
13. For research on this topic, how would you define severe overall attrition?	Internal Validity–Selection	• More than 20% loss of the original sample

Contextual Question	Study DIAD Composite Question	Example of Answer to Questions for Evaluating Research on the Effects of Homework on Academic Achievement
14. For research on this topic, what constitutes a minimal sample size that would permit a sufficiently precise estimate of the effect size?	Statistical Validity–Effect Size Estimation	• 50 students in each group
15. What percentage of important statistical information (i.e., sample size, direction of effect, effect size) is needed for the results of this study to be "fully," "largely," and "rarely" reported?	Statistical Validity–Reporting	• If full statistical results are known for all measured outcomes, results are "fully reported." • If full statistical results are known for 75–99% of measured outcomes, results are "largely reported." • If full statistical results are known for less than 75% of measured outcomes, results are "rarely reported."
16. Considering the outcome measure and the context of this research question, what constitutes "overalignment" and "underalignment" of the intervention and outcome?	Fit Between Concepts–Outcome	• No outcomes are overaligned to the intervention for this research question. • If the homework assignments cover distinctly different subject matter than the assessment, then the outcomes are underaligned to the intervention.

For any study, the Study DIAD results in three sets of answers to questions about its methods, or its quality for testing the hypothesis of interest: (a) about 30 design and implementation questions, (b) eight composite questions, and (c) four global questions. Really, any of these characterizations of the study can be used to judge the study or to see if features of a study's methods are related to the study's outcome. In fact, if you wanted to exclude studies a priori based on Study DIAD results, you could set a minimum profile that the study had to meet or exceed in order to be included in the synthesis. For example, the McGrath study might be excluded because it got only a "Maybe yes" on Global Question 2 about internal validity (Table 5.4).

Table 5.3 Algorithm for Combining Design and Implementation Questions to Arrive at an Answer to Composite Question 1.2: "Fit Between Concepts and Operations: Outcome Measure: Were the outcomes measured in a way that is consistent with the proposed effects of the intervention?"

	Response Pattern (Read down columns to determine the answer to the question)		
1.2.1 Do items on the outcome measure appear to represent the content of interest (i.e., have face validity)?	Yes	Yes	Yes or No
1.2.2 Were the scores on the outcome measure acceptably reliable?	Yes	No	Yes or No
1.2.3 Was the outcome measure properly aligned to the intervention condition?	Yes	Yes	No
Answer to Composite Question 1.2 associated with this response pattern:	**Yes**	**Maybe No**	**No**

Table 5.4 Example of Global and Composite Ratings for a Study of the Effects of Homework on Academic Achievement (McGrath, 1993)

Global Ratings		Composite Ratings	
Question	Rating	Question	Rating
1. Were the participants treated and the outcomes measured in a way that is consistent with the definition of the intervention and its proposed effects?	Yes	1.1 Were the participants treated in a way that is consistent with the definition of the intervention?	Yes
		1.2 Were the outcomes measured in a way that is consistent with the proposed effects of the intervention?	Yes
2. Did the research design permit an unambiguous conclusion about the intervention's effectiveness?	Maybe yes	2.1 Were the participants in the group receiving the intervention comparable to the participants in the comparison group?	Yes
		2.2 Was the study free of events that happened at the same time as the intervention that confused its effect?	Maybe yes
3. Was the intervention tested on participants, settings, and outcomes representative of its intended beneficiaries?	No	3.1 Did the study include variation on participants, settings, and outcomes representative of the intended beneficiaries?	No
		3.2 Was the intervention tested for its effect within important subgroupings of participants, settings, and outcomes?	No
4. Could accurate estimates of the intervention's impact be derived from the study report?	No	4.1 Were effect sizes and their standard errors accurately estimated?	No
		4.2 Were statistical tests adequately reported?	Yes

The Study DIAD is a complex and time-intensive instrument to apply, one that takes careful thought and sufficient training to be used properly. But this complexity is a reflection of the fact that coming to careful and transparent decisions about research quality is not a simple task. If we acknowledge this fact in our work, then the Study DIAD has many characteristics to its credit. First, the 30-odd characteristics of a study's design and implementation that form the core of the instrument (presented in Table 5.1) were arrived at using input from a broad sampling of social science researchers. So, there is greater consensus that these are the critical methodological features of studies to consider when judging a study's quality than is the case in other quality scales. Second, the fact that the Study DIAD requires that its users make explicit their definitions of important terms before it is applied (defined in Table 5.2) means that the meaning of these terms will be clear to people who read the synthesis. If there is disagreement about these definitions, then fruitful discussions can ensue about where the disagreement lies. Third, the algorithms are made explicit for combining the study design and implementation features into answers to the more abstract questions about quality (presented in Figure 5.1).

You can use the Study DIAD in several ways. Certainly, it isbest to apply it in its full realization. But, as noted previously, you could also use the global and/or composite questions to guide you if you want to use a threats-to-validity approach. Or you could use the 30-odd design and implementation questions to guide your use of a methods-description approach. It is a simple task to transfer the 30-odd questions on to a coding sheet similar to those presented in Chapter 4. The definitions presented in Table 5.2 can be incorporated directly into the coding definitions.

What is most important is that when you are examining how the issue of evaluating the design and implementation of the constituent studies was handled in a research synthesis you ask this question:

> Were studies categorized so that important distinctions could be made among them regarding their research design and implementation?

Identifying Statistical Outliers

Another aspect of evaluating studies cannot occur until after all of your data have been coded and entered into the computer and a first analysis of study results is ready to be run. At this point, the most extreme outcomes of the individual studies in your data set need to be examined to see if they are statistical outliers. You will want to discover if the most extreme outcomes are so discrepant from the other results that it is unlikely they are actually members of the same distribution of findings. For example, suppose you have a set of 60 correlations between the age of the respondent to his or her attitude about rape. Suppose as well that 59 of the correlations range in value from –.05 to +.45, with positive values indicating that older respondents are less accepting of rape than younger ones. However, the 60th correlation is –.65. You can use statistical procedures and conventions to compare this most extreme data point to the overall sample distribution. You try to determine if this most extreme study outcome is too different from the overall distribution of outcomes to be considered a part of it.

Statistical outliers sometimes occur because of errors committed on your coding sheets or during data transfer. These you can correct. Or outliers can appear because these same types of errors were made by primary researchers. These you cannot correct except by asking the primary researchers to confirm the result for you. Sometimes the cause of a data point being a statistical outlier is unknown. Still, research synthesists agree that something needs to be done when a data point is so extreme that it is unlikely to be a member of the distribution of findings it is being compared to. One approach is simply to remove the study from your database. Another strategy is to reset the value of the outcome to three standard deviations above the mean or to its next nearest neighbor.

As an example, in the meta-analysis of the effects of choice on intrinsic motivation, we applied Grubbs' (1950) test to identify outliers separately for each outcome measure. These analyses identified none, one, or two outliers depending on the outcome. We could not discern the cause of the outliers but we set them to their nearest neighbor and retained the study for further analyses.

Barnett and Lewis (1984) provide a thorough examination of ways to identify statistical outliers and ways to treat them when found. *Whatever method you choose, it is good practice to look for statistical outliers and, when found, to address them in some manner.* This is the final step in evaluating the question of whether a particular study helps you get the best answer to the question that has motivated your research synthesis.

Exercises

1. Complete Table 5.2 for the topic of the synthesis you have chosen.
2. Pick a study that is relevant to your topic and answer the questions in Table 5.1.
3. Construct a method section of a coding guide for your research topic using Table 5.1 and your responses to the questions in Table 5.2. Pair up with a classmate and apply the coding frame to a study for each other's topic. What problems did you encounter? How would the problems change the way you would answer the questions in Table 5.2?

6 Step 5 Analyzing and Integrating the Outcomes of Studies

What procedures should be used to condense and combine the research results?

Primary Function in Research Synthesis

To identify and apply procedures for (a) combining results across studies and (b) testing for differences in results between studies

Procedural Variation That Might Produce Differences in Conclusions

Variation in procedures used to summarize and compare results of included studies (e.g., narrative, vote count, averaged effect sizes) can lead to differences in cumulative results.

Questions to Ask When Analyzing and Integrating the Results of Studies

1. Was an appropriate method used to combine and compare results across studies?
2. If a meta-analysis was performed, was an appropriate effect size metric used?
3. 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?
4. If a meta-analysis was performed, was the homogeneity of effect sizes tested?
5. Were (a) study design and implementation features along with (b) other critical features of studies, including historical, theoretical, and practical variables, tested as potential moderators of study outcomes?

This chapter describes

- A rationale for the use of meta-analyses
- Statistical methods used to summarize research results including
  - Counting study outcomes
  - Averaging effect sizes
  - Examining the variability in effect sizes across studies
- Some practical issues in the application of meta-analytic procedures
- Some advanced meta-analytic procedures

Data analysis involves reducing the separate data points collected by the inquirer into a unified statement about the research problem. It involves ordering, categorizing, and summarizing the data, as well as performing inference tests that attempt to relate data samples to the populations they arise from. Inferences made from data analysis require that decision rules be used to distinguish systematic data patterns from noise (or chance fluctuation). Although different decision rules can be used, the rules involve assumptions about what the target population looks like (e.g., it is normally distributed) and what criteria (e.g., the threshold probability for declaring a finding statistically significant) must be met before an existing pattern in the data is said to be reliable. The purpose of data analysis is to summarize and describe the data in a form that permits valid interpretation.

Data Analysis in Primary Research and Research Synthesis

Just as any scientific inquiry requires the leap from concrete operations to abstract concepts, both primary researchers and research synthesists must leap from patterns found in samples of data to more-general conclusions about whether these patterns also exist in the target populations. However, until the mid-1970s, there had been almost no similarity in the analysis techniques used by primary researchers and research synthesists. Primary researchers were obligated to present sample statistics and to substantiate any inferences drawn from their data by providing the results of statistical tests. Most frequently, primary researchers (a) compared sampled means to one another or calculated other measures of relationship, (b) made the assumptions needed for conducting inference tests relating the sample results to populations, and (c) reported the probabilities associated with whether systematic differences in the sample could be inferred to hold in the target population as well.

Traditional statistical aids to primary data interpretation have not gone uncriticized. Some have argued that significance tests are not very informative since they tell only what the likelihood is of obtaining the observed results when the null hypothesis is true (e.g., Cohen, 1994; Cumming, 2012). These critics argue that in a population of people, the null hypothesis is rarely if ever true and therefore the significance of a given test is mainly influenced by how many participants have been sampled. Also, critics who are skeptical about the value of null hypothesis significance testing point to limitations in the generalization of these findings to the target population. No matter how statistically significant a relation may be, the

results of a study are generalizable only to people like those who participated in that particular research effort.

Skepticism about the value of statistics helps those who use them refine their procedures and keep their output in proper perspective. Nonetheless, most primary researchers use statistics and most would feel extremely uncomfortable about summarizing the results of their studies without some assistance (or credibility) supplied by statistical procedures. Saying, "I looked at the group means and they looked different to me" is simply not acceptable in primary research.

In contrast to primary researchers, until recently research synthesists were not obligated to apply any statistical techniques in the interpretation of cumulative results. Traditionally, synthesists interpreted data using intuitive rules of inference unknown even to themselves. Analysis methods were idiosyncratic to the perspective of that particular synthesist. Therefore, a description of the common rules of inference used in research syntheses was not possible.

The subjectivity in analysis of research literatures led to skepticism about the conclusions of many syntheses. To address the problem, methodologists introduced quantitative methods into the synthesis process. The methods use the statistics contained in the individual studies as the primary data for the research synthesis.

Meta-Analysis

I suggested in [Chapter 1](#) that the two events that had the greatest influence on state-of-the-art research synthesis are the growth in the amount of research and the rapid advances in computerized research retrieval systems. A third major influence is the introduction of quantitative procedures, called *meta-analysis,* into the research synthesis process.

The explosion in social science research focused considerable attention on the lack of standardization in how synthesists arrived at general conclusions from series of related studies. For many topic areas, a separate verbal description of each relevant study was no longer possible. One traditional strategy was to focus on one or two studies chosen from dozens or hundreds. This strategy failed to portray accurately the accumulated state of knowledge. Certainly, in areas where dozens or hundreds of studies exist, synthesists must describe prototype studies so that readers understand the methods used by primary researchers.

However, relying on the results of prototype studies to represent the results of all studies may be seriously misleading. First, as we have seen, this type of selective attention is open to confirmatory bias: synthesists may highlight only studies that support their initial position. Second, selective attention to only a portion of all studies places little or imprecise weight on the volume of available tests. Presenting one or two studies without a cumulative analysis of the entire set of results gives the reader no estimate of the confidence that should be placed in a conclusion. Finally, selectively attending to evidence cannot give a good estimate of the strength of a relationship. As evidence on a topic accumulates, researchers become more interested in *how much* of a relationship exists between variables rather than simply *whether a relationship exists at all.*

Synthesists not employing meta-analysis also face problems when they consider the variation between the results of different studies. They will find distributions of results for studies sharing a particular procedural characteristic but varying on many other characteristics. Without meta-analysis, it is difficult to conclude accurately whether a procedural variation affected study outcomes; the variability in results obtained by any single method likely will overlap with the distributions of results of studies using a different method.

It seems, then, that there are many situations in which synthesists need to turn to meta-analytic techniques. The application of quantitative inference procedures to research synthesis was a necessary response to the expanding literature. If statistics are applied appropriately, they should enhance the validity of a synthesis' conclusions. Quantitative research synthesis is an extension of the same rules of inference required for rigorous data analysis in primary research. If primary researchers must specify quantitatively the relation of the data to their conclusions, the next users of the data should be required to do the same. The inference procedure that sounded so ludicrous in the context of a single study ("The means looked different to me") is no less so in the context of research synthesis.

Meta-Analysis Comes of Age

Early on, meta-analysis was not without its critics, and some criticisms persist. Initially, the value of quantitative synthesis was questioned along lines similar to criticisms of primary data analysis (e.g., Barber, 1978; Mansfield & Bussey, 1977). However, much of the criticism stemmed less from issues in meta-analysis than from inappropriate aggregation procedures that are more general, such as a lack of attention to moderating variables, that were incorrectly thought to be caused by the use of quantitative combining procedures when they were really independent (and poor) decisions on the part of the

research synthesists. I will return to criticism of meta-analysis, and rigorous research synthesis in general, in the final chapter.

Meta-analysis is now an accepted procedure and its application within the social and medical sciences is on the ascent. Today, literally thousands of meta-analyses have been published, and the number published each year continues to grow larger. Figure 6.1 presents some evidence of this increasing impact in the sciences and social sciences. The figure is based on entries in the Web of Science Core Collection (retrieved April 3, 2015). It charts the growth in the number of documents retrieved by using the topics "research synthesis," "systematic review," "research review," "literature review," and/or "meta-analysis" for even-numbered years from 1996 to 2014. The figure indicates that the total number of references has risen every year without exception and is accelerating. Clearly, the role that research syntheses and meta-analysis play in our knowledge claims is large and growing larger.

When Not to Do a Meta-Analysis

Much of this chapter will describe some basic meta-analysis procedures and how they are applied. However, it is important to state explicitly some circumstances for which the use of quantitative procedures in research syntheses is *not* appropriate.

First, quantitative procedures are applicable only to research syntheses, not to literature reviews with other foci or goals (see Chapter 1). For instance, if a literature reviewer is interested in tracing the historical development of the concept "intrinsic motivation," it would not be necessary for him or her to do a quantitative synthesis. However, if the synthesist also intended to make inferences about whether different definitions of intrinsic motivation lead to different research results, then a quantitative summary of relevant research would be appropriate. Also, meta-analysis is not called for if the goal of the literature review is to critically or historically appraise the research study by study or to identify particular studies that are central to a field. In such instances, a proper integration likely would treat the results of studies as an emerging series of events—that is, it would use a historical approach to organizing the literature review rather than a statistical aggregation of the cumulative findings. However, if the synthesists are interested in whether the results of studies *change* over time, then meta-analysis would be appropriate.

Figure 6.1 Web of Science Core Collection Frequency of References to "Research Synthesis," or "Systematic Review," or "Research Review," or "Literature Review," or "Meta-Analysis"
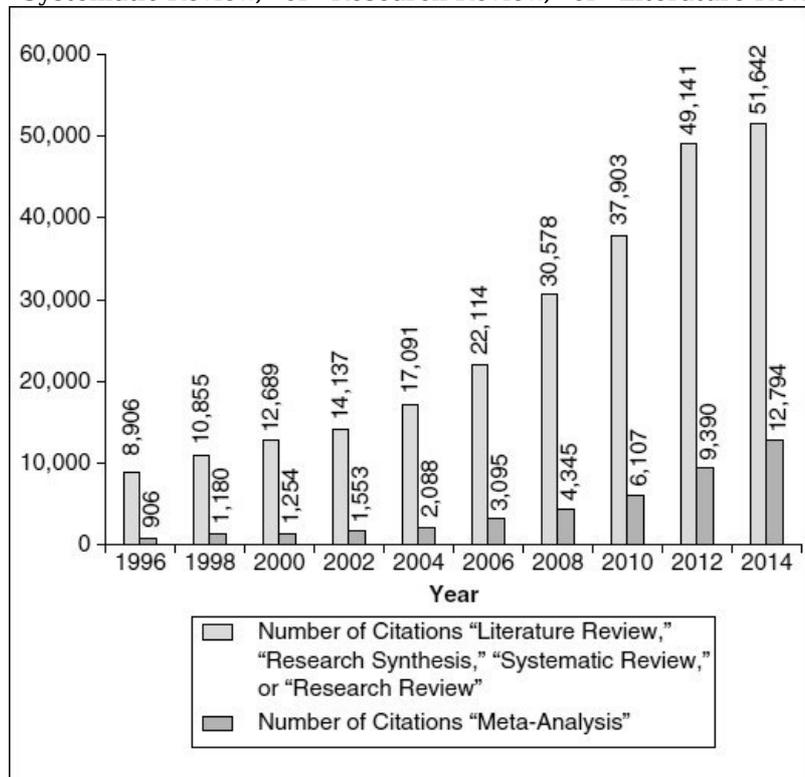
Second, the basic premise behind the use of statistics in research syntheses is that a series of studies address an identical conceptual hypothesis. If the premises of a literature review do not include this assertion, then there is no need for cumulative statistics. Related to this point, a synthesist should not quantitatively combine studies at a broader conceptual level than readers would find useful. At an extreme, most social science research could be categorized as examining a single conceptual hypothesis —social stimuli affect human behavior. Indeed, for some purposes such a hypothesis test might be very enlightening. However, the fact that "it can be done" should not be used as an excuse to quantitatively lump together concepts and hypotheses simply because methods are available to do so (see Kazdin,

Durac, & Agteros, 1979, for a humorous treatment of this issue). Synthesists must pay attention to those distinctions in the literature that will be meaningful to the users of the synthesis. For example, in the meta-analysis of the effects of choice on intrinsic motivation, we did not combine study results across the nine different outcome measures. Doing so would have obscured important distinctions among the outcomes and might have been misleading. Instead, the highest level of data aggregation was within outcome types.

Another instance of too much aggregation occurs when a hypothesis has been tested using different types of controls. For example, one study examining the effect of daily aerobic exercise on adults' levels of cognitive functioning might compare this treatment to a no-treatment control while another study compares it to a treatment in which participants receive written information about the importance of exercise. It might not be informative to statistically combine the results of these two studies. To what comparison does the combined effect relate? Synthesists might find that a distinction in the type of control group is important enough not to be obscured in a quantitative analysis (but an analysis of the moderating effects of different types of control groups might be appropriate here).

Third, under certain conditions meta-analysis might not lead to the kinds of generalizations the synthesists wish to make. For example, cognitive psychologists or cognitive neuroscientists might argue that their methodologies typically afford good controls and reasonably secure findings because the things they study are not strongly affected by the context in which the study is conducted. Thus, the debate about effects in these areas of research usually occurs with reference to the choice of variables and their theoretical, or interpretive, significance. Under these circumstances, a synthesist might convincingly establish generalization using conceptual and theoretical bridges rather than statistical ones.

Finally, even if synthesists wish to summate statistical results across studies on the same topic, they may discover that only a few studies have been conducted and that these use methodologies, participants, and outcome measures that are decidedly different from one another. In circumstances where multiple methodological distinctions are confounded with one another (e.g., a particular research design occurs very frequently with a particular type of subject), the statistical combination of studies might mask important differences in research that make interpretation of the synthesis findings difficult. In these instances, it may make the most sense not to use meta-analysis, or to conduct several discrete meta-analyses within the same synthesis by combining only those studies that share similar clusters of features.

It is also important to point out that *the use of meta-analysis is no guarantee that the synthesist will be immune from all inferential errors*. The possibility always exists that the meta-analyst has incorrectly inferred a characteristic of the target population. As in the use of statistics in primary research, this can occur because the target population does not conform to the assumptions underlying the analysis techniques or because of the probabilistic nature of statistical findings. If you think that the population statistics do not conform to the assumptions of the statistical test you have chosen, find a more appropriate test or eschew the use of meta-analysis altogether. In sum, then, an important question to ask when evaluating a research synthesis is,

Was an appropriate method used to combine and compare results across studies?

The Impact of Integrating Techniques on Synthesis Outcomes

In Chapter 1 I described a study I conducted with Robert Rosenthal (Cooper & Rosenthal, 1980) in which we demonstrated some of the differences in conclusions that might be drawn by nonquantitative synthesists and meta-analysts. In that study, graduate students and university faculty members were asked to evaluate the same set of studies, but half used quantitative procedures and half used whatever criteria appealed to them. We found that the meta-analysts thought there was more support for the hypothesis and a larger relationship between variables than did the non-meta-analysts. Meta-analysts also tended to view future replications as less necessary than did non-meta-analysts, although this finding did not reach statistical significance.

It is also likely that the different statistical procedures used by meta-analysts will create variance in synthesis conclusions. Several different paradigms have emerged for quantitatively integrating research with a traditional inference testing model (Hedges & Olkin, 1985; Rosenthal, 1984; Schmidt & Hunter, 2015), while others use a Bayesian perspective (Sutton, Abrams, Jones, Sheldon, & Song, 2000; United States Department of Health and Human Services Agency for Healthcare Research and Quality, 2013). The different techniques generate different output. Thus, the rules adopted to carry out quantitative analysis can differ from synthesist to synthesist, which might create differences in how synthesis results are interpreted. We can assume as well that the rules used by nonquantitative synthesists also vary, but that because of their inexplicit nature it is difficult to compare them formally.

Main Effects and Interactions in Meta-Analysis

Before examining several of the quantitative techniques available to synthesists, it is important to take a closer look at some of the unique features of accumulated research results. In Chapter 2 on problem formulation, I pointed out that most research syntheses first focus on tests of main effects that were carried out in the primary studies. This is largely because conceptually related replications of main effects occur more frequently than tests of three or more interacting variables. So, for example, you are likely to find in primary studies many more main-effect tests of whether choice influences intrinsic motivation than you are to find tests of interactions of whether this relationship is influenced by the number of choices given. Keep in mind that I am referring here to interaction tests within a single study, not your ability to test for the influence of number-of-choices at the synthesis level because different studies have varied in the number of choices they provide in their test of the main effect.

It is not that interactions tested in primary studies cannot be combined. However, such replications are fewer and, we shall see in the next chapter, their interpretation can be a bit more complex. There are two different ways that interactions tested in primary research could be statistically combined across studies. First, the relationship strengths associated with each study's interaction test could be aggregated. An alternative strategy would be to aggregate separately the relationship of two of the interacting variables at each level of the third variable. For instance, assume there exists a set of studies in which the primary researchers tested whether the effect of choice in intrinsic motivation differed depending on the number of choices given to participants. The synthesists could generate an estimate of the difference in intrinsic motivation depending on the number of choices given. They could aggregate all motivation measures taken under conditions where a choice between two alternatives was compared to no choice. They could do the same for measures taken after, say, two or three choices. Then, the different effect sizes could be compared. This would probably be more useful and easily interpretable than a direct estimate of the magnitude of the interaction effect. However, in order to do this, the primary research reports must contain the information needed to isolate the different simple main effects. The synthesist might also have to group numbers of choices (e.g., three to five choices and six or more choices) in order to have enough tests to generate a good estimate.

Because main effects are most often the focus of meta-analysts and in many instances meta-analysts interested in interactions reduce them to simple effects, my discussion of the quantitative combining techniques will refer to main effects only. The generalization to meta-analyzing interactions is mathematically straightforward.

Meta-Analysis and the Variation Among Study Results

In research syntheses, the most obvious feature of both main effects and interactions is that the results of the separate tests of the same relationship will vary from one study to the next. This variability is sometimes dramatic and requires us to ask where the variability comes from.

Sources of Variability in Research Findings

Differences in the outcomes of studies can be caused by two types of influences. The simplest cause is the one that is most often overlooked by nonquantitative synthesists—sampling variability. Even before the current interest in quantitative synthesis, Taveggia (1974) recognized this important influence:

> A methodological principle overlooked by writers of . . . reviews 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. (pp. 397–398, emphasis in original)

Taveggia highlights one of the implications of using probability theory and sampling techniques to make inferences about populations.

As an example, suppose it was possible to measure the academic achievement of every American student as well as whether each student did homework. Also, suppose that if such a task were undertaken, it would be found that achievement was exactly equal for students who do and do not do homework—that is, exactly equal achievement test mean scores existed for the two subpopulations. Still, if 1,000 samples of 50 homeworkers and 50 no-homeworkers were drawn from this population, very few comparisons between samples would reveal exactly equal group means. About half would show homeworkers achieving better and half would show no-homeworkers achieving better. Furthermore, if the sample means were compared statistically using a t-test and the $p < .05$ significance level (two-tailed), about 25 comparisons would show a significant difference favoring homeworkers while about 25 would favor no-homeworkers. This variation in results is an unavoidable consequence of the fact that the means estimated by sampling will vary somewhat from the true population values. And, just by chance alone,

some comparisons will pair sample estimates that vary from their true population values by large amounts and in opposite directions.

In the example given, it is unlikely that you would be fooled into thinking anything but chance caused the result—after all, 950 comparisons would reveal nonsignificant differences and significant results would be distributed equally for both significant positive and negative outcomes. However, in practice the pattern of results is rarely this clear. As we discovered in the chapter on literature searching, you might not be aware of all null results because they are hard to find. Complicating matters further, even if an overall relation does exist between two variables (i.e., the null hypothesis is false), some studies can still show significant results in a direction opposite to the relation in the population. To continue the example, if the average achievement of homeworkers is better than no-homeworkers, some comparisons of samples randomly drawn from the two subpopulations will still favor no-homeworkers, the number depending on the size of the relation, the size of the samples, and how many comparisons have been performed. In sum, then, one source of variance in the results of studies can be chance fluctuations due to the inexactness of estimates based on samples drawn from populations.

A second source of variance in study outcomes is of more interest to synthesists. This variance in results is created by differences in how studies are conducted. This variance is added to the variance due to sampling participants. Just as people are sampled, you can think of a set of studies as a sample of studies drawn from a population of all possible studies. And, because studies can be conducted in different ways (just as people can differ in personal attributes) that affect the studies outcomes, a sample of studies also will exhibit chance variation from other possible samples of studies. For instance, the homework synthesists might find that studies comparing achievement among students who do and do not do homework have been conducted with students at different grade levels; with unit tests, class grades, or standardized tests as measures of achievement; and with an assortment of classes with different subject matters. Each of these differences in the studies' methods or contexts could create variation in study results and therefore could create results that differ randomly from another sample of studies drawn from the same population of studies. This variation will be added to the variation caused by the sampling of study participants from the population of participants.

It is also possible that this variation associated with study-level differences is systematically related to the variation in study results. For example, homework studies conducted with elementary school students might produce results that differ systematically from studies conducted with high school students. In Chapter 2, the notion of synthesis-generated evidence was introduced to describe what we learn when we find associations between study characteristics and study outcomes.

The existence of the two sources of variance in research results—the one generated by sampling participants and the other by sampling studies—raises an interesting dilemma. When discrepant findings occur within a set of studies, should you seek an explanation for them by attempting to identify systematic differences in results associated with differences in the methods used in studies? Or should you simply assume the discrepant findings were produced by variations due to sampling (of participants and/or study procedures)? Some tests have been devised to help you answer this question. In effect, these tests use sampling error (associated with participants or both participants and studies) as the null hypothesis. They estimate the amount of variance in findings that would be expected if sampling error alone were making the study findings different.[1] If the observed variation in results across studies is too great to be explained by sampling error alone, then the null hypothesis is rejected. It suggests that the notion that all the results were drawn from the same population of results can be rejected.

In the sections that follow, I will introduce some of the quantitative synthesis techniques that are available to you. I have chosen the techniques because they are relatively simple and broadly applicable. The treatment of each technique will be conceptual and introductory but detailed enough to permit you to perform a sound, if basic, meta-analysis. You can consult the primary sources cited in the text if (a) you want a more detailed description of these techniques and their variations, including how they are derived, and/or (b) your meta-analysis has some unique possibilities for exploring data in ways not covered here. For the discussion that follows, I have assumed you have a working knowledge of the basic inferential statistics employed in the social sciences.

Before I begin, though, there are three assumptions crucial to the validity of a conclusion based on an integration of statistical findings from individual studies. First and most obviously, *the individual findings that go into a cumulative analysis should all test the same comparison or estimate the same relationship.* Regardless of how conceptually broad or narrow your ideas might be, you should be comfortable with the assertion that the included statistical tests from the primary studies address the same question. Second, *the separate tests that go into the cumulative analysis must be independent of one another.* Identifying independent comparisons was discussed in Chapter 4, on gathering information from studies. You must take care to identify comparisons so that each one contains unique information about the hypothesis. Finally, you must believe that *the primary researchers made valid assumptions when they computed the results of their tests.* Thus, for example, if you want to combine the effect sizes resulting from comparisons between two means, you must assume that the observations in the two groups in the primary studies are independent and normally distributed, and that their variances are roughly equal to

one another.

Vote Counting

The simplest methods for combining independent statistical tests are the vote counting methods. Vote counts can take into account the statistical significance of findings or focus only on the direction of the findings.

For the first method, the meta-analysts would take each finding[2] and place it into one of three categories: statistically significant findings in the expected direction (I will refer to these as positive findings), statistically significant findings in the unexpected (negative) direction, and nonsignificant findings—that is, findings that did not permit rejection of the null hypothesis. The meta-analysts then might establish the rule that the category with the largest number of findings tells what the direction of the relationship is in the target population.

This vote count of significant findings has much intuitive appeal and has been used quite often. However, the strategy is unacceptably conservative and often can lead to erroneous conclusions (Hedges & Olkin, 1980). The problem is that using the traditional definition of statistical significance, chance alone should produce only about 5% of all findings falsely indicating a significant effect. Therefore, much fewer than one-third positive and statistically significant findings might indicate a real difference exists in the target population. This vote-counting strategy requires that at least 34% of findings be positive and statistically significant before a result is declared a winner.

Let me illustrate just how conservative this approach is. Assume that a correlation of $r = .30$ exists between two variables in a population and 20 studies have been conducted with 40 people in each sample (this would not be an uncommon scenario in the social sciences). The probability that the vote count associated with this series of studies will conclude a positive relation exists—if the plurality decision rule described in the preceding paragraph is used—is less than 6 in 100. Thus, the vote count of significant findings could, and often does, lead vote counters to suggest accepting the null hypothesis, and perhaps abandoning fruitful theories or effective interventions when, in fact, no such conclusion is warranted.

Adjusting the frequencies of the three types of findings (positive, negative, and null) so that the true expected percentage of each finding (95% null and 2.5% significant in each direction) is taken into account solves one problem but it highlights another one. We have seen that null results are less likely to be reported by researchers and are less likely to be retrieved by synthesists. Therefore, if the appropriate expected values are used in a vote-count analysis, it could often occur that *both* positive and negative significant findings appear more frequently than would be expected by chance alone. Thus, it seems that using the frequency of nonsignificant findings in a vote count procedure is of dubious value.

An alternative vote-counting method is to compare the frequency of statistically significant positive findings against the frequency of significant negative ones. This procedure assumes that if the null hypothesis prevails in the population, then the frequency of significant positive and negative findings is expected to be equal. If the frequency of findings is found not to be equal, then the null hypothesis can be rejected in favor of the prevailing direction. A problem with this vote-count approach is that the expected number of nonsignificant findings, even when the null hypothesis is not true, can still be much greater than the expected number of either positive or negative significant findings. Therefore, this approach will ignore many findings (all nonsignificant ones) and will be very low in statistical power.

A final way to perform vote counts in research synthesis involves tallying the number of positive and negative findings regardless of their statistical significance. In this approach, the meta-analyst categorizes findings based solely on the direction of their outcome, ignoring their statistical significance. Again, if the null hypothesis is true—that is, if no relationship exists between the variables in the sampled population—we would expect the number of findings in each direction to be equal.

Once the number of results in each direction is counted, the meta-analyst can perform a simple sign test to discover if the cumulative result suggests that one direction occurs more frequently than would be expected by chance. The formula for computing the sign test is as follows:

(1)
$$Z_{vc} = \frac{(N_p) - (\frac{1}{2}N)}{\frac{1}{2}\sqrt{N}} \tag{1}$$

where

> Z_{vc} = the standard normal deviate, or Z-score, for the overall series of findings;
> N_p = the number of positive findings; and

N = the total number of findings (positive plus negative findings).

The Z_{vc} can be referred to a table of standard normal deviates to discover the probability (one-tailed) associated with the cumulative set of directional findings. If a two-tailed p-level is desired, the tabled p-value should be doubled. The values of Z associated with different p-levels are presented in Table 6.1. This sign test can be used in a vote count of either the simple direction of all findings or the direction of only significant findings, though using the direction of findings is recommended.

Suppose 25 of 36 comparisons find that adults given an intervention to increase aerobic activity exhibited better neurocognitive functioning than those in a no-intervention group. The probability that this many findings would be in one direction, given that in the target population (of all intervention tests) there is equal neurocognitive functioning exhibited by people in the two conditions, is $p < .02$ (two-tailed) associated with a Z_{vc} of 2.33. This result would lead the meta-analyst to conclude a positive intervention effect was supported by the series of findings.

The vote-count method that uses the direction of findings regardless of significance has the advantage of using information from all statistical findings. Still, it has some drawbacks. Similar to the other vote-count methods, it does not weight a finding's contribution to the overall result by its sample size. Thus, a finding based on 100 participants is given weight equal to one with 1,000 participants. Furthermore, the revealed magnitude of the relationship (e.g., the impact of the treatment) in each finding is not considered—a finding showing a large increase in cognitive functioning due to the intervention is given equal weight to one showing a small decrease in functioning. Finally, a practical problem with the directional vote count is that primary researchers frequently do not report the direction of findings if a comparison proved statistically nonsignificant.

Table 6.1 Standard Normal Deviation Distribution

z-score	Area z to −z	p-level 2-tailed	p-level 1-tailed
2.807	.995	.005	.0025
2.576	.99	.01	.005
2.432	.985	.015	.0075
2.326	.98	.02	.01
2.241	.975	.025	.0125
2.170	.97	.03	.015
2.108	.965	.035	.0175
2.054	.96	.04	.02
2.000	.954	.046	.023
1.960	.95	.05	.025
1.881	.94	.06	.03
1.751	.92	.08	.04
1.645	.9	.1	.05
1.440	.85	.15	.075
1.282	.8	.2	.10
1.150	.75	.25	.125
1.036	.7	.3	.150
0.842	.6	.4	.20
0.674	.5	.5	.25
0.524	.4	.6	.30
0.385	.3	.7	.35
0.253	.2	.8	.40
0.126	.1	.9	.45

SOURCE: Adapted from: Wikipedia (2015), http://en.wikipedia.org/wiki/Standard_normal_table
SOURCE: Adapted from: Wikipedia (2015), http://en.wikipedia.org/wiki/Standard_normal_table

Still, the vote count of directional findings can be an informative complement to other meta-analytic procedures, and can even be used to generate an estimate of the strength of a relationship. Bushman and Wang (2009) provide formulas and tables that can be used to estimate the size of a population relationship given that the meta-analysts know (a) the number of findings, (b) the direction of each finding, and (c) the sample size of each finding. For example, let's assume that each one of the 36 comparisons between an activity intervention and no-intervention group was based on a sample size of 50 participants. Using Bushman and Wang's technique, I find that when 25 of the 36 (69%) comparisons revealed better cognitive functioning in the intervention group, the most likely population value for a correlation between group membership and activity is $r = .07$. Of course, this example is artificial because I assumed all the sample sizes were equal. The calculations are more complex in many circumstances, not only because sample sizes vary but also because you will have comparisons (votes) for which you have no direction. This complicates the estimating technique greatly. In the past, when we have used this technique (see Cooper, Charlton, Valentine, & Muhlenbruck, 2000), we conducted the analyses several times, using different sets of assumptions. In general, this technique should be used with caution and only in conjunction with other meta-analytic techniques that produce conclusions that are less tentative.

In sum, then, meta-analysts can perform vote counts to aggregate results across individual studies by comparing the number of directional findings and/or the number of significant directional findings. Both of these procedures will be very imprecise and conservative—that is, they will accept the null hypothesis when more-precise methods suggest it should be rejected. The simple direction of results will not appear

in many research reports in the first case, and nonsignificant findings cannot contribute to the analysis in the second case. Vote counts can be described in meta-analyses but should be used to draw inferences only in combination with more sensitive meta-analysis procedures.

Combining Significance Levels

One way to address the shortcomings of vote counts is to consider combining the exact probabilities associated with the results of each comparison. Rosenthal (1984) cataloged 16 methods for combining the results of inference tests so that an overall test of the null hypothesis can be obtained. By using the exact probabilities, the results of the combined analysis take into account the different sample sizes and relationship strengths found in each comparison. Thus, the combining-significance-levels procedure overcomes the improper weighting problems of the vote count. However, it has severe limitations of its own. First, as with vote counts, the combining-probability procedures answer the "yes or no?" questions but not the "how much?" question. Second, whereas the vote-count procedure is overly conservative, the combining-significance-levels procedure is extremely powerful. In fact, it is so powerful that for hypotheses or relationships that have generated a large number of findings, rejecting the null hypothesis is so likely, because even very small relationships can produce significant combined probabilities, that it becomes a rather uninformative exercise. For this reason, these procedures have largely fallen out of use.

Measuring Relationship Strength

The primary function of the procedures described so far is to help meta-analysts accept or reject the null hypothesis. Until recently, most researchers interested in social theory and the impact of social interventions have been content to simply identify relations that have some explanatory value. The prevalence of this "yes or no" question was partly due to the relatively imprecise nature of social science theories and hypotheses. Social hypotheses typically were crudely stated first approximations to the truth. Social researchers rarely asked how potent theories or interventions were for explaining human behavior or how competing explanations compare with regard to their relative explanatory value. Today, as their theories and interventions are becoming more sophisticated, social scientists are more often making inquiries about the size of relationships.

Giving further impetus to the "how much?" question is a growing disenchantment with the null hypothesis significance test itself. As I noted earlier, whether a null hypothesis can be rejected is tied closely to the particular research project under scrutiny. If an ample number of participants are available or if a sensitive research design is employed, a rejection of the null hypothesis often is not surprising. This state of affairs becomes even more apparent in meta-analyses that include a combined significance level, where the power is great to detect even very small relations. A null hypothesis rejection, then, does not guarantee that an important social insight has been achieved.

Finally, when used in applied social research, the vote-count and combined-significance-level techniques give no information on whether the effect of a treatment or the relationship between variables is large or small, important or trivial. For example, if we find that the relationship between whether a participant (a) is an adolescent or adult and (b) believes that women share some culpability when a rape occurs is statistically significant and the correlation is $r = .01$, is this a strong enough relationship that it should influence how interventions are delivered? What if the result is statistically significant and the correlation is $r = .30$? This example suggests that the "yes or no?" question is often not the question of greatest importance. Instead, the important question is, "How much does the age of the participant influence beliefs about rape?" The answer might be zero or it might suggest a small or large relationship. The answer to this question could help meta-analysts (and others) make recommendations about how best to construct rape-attitude interventions so they are most effective. Given these questions, meta-analysts would turn to the calculation of average effect sizes. Also, as we shall see shortly, the null hypothesis question, "Is the relationship different from zero?" can be answered by placing a confidence interval around the "how much?" estimate, removing the need for separate null hypothesis significance tests.

Definition of Effect Size

In order to answer meaningfully the "how much?" question, we must agree on definitions for the terms *magnitude of difference, relationship strength,* or what generally is called the *effect size.* Also, we need methods for quantitatively expressing these ideas once we have defined them. Jacob Cohen's (1988) book *Statistical Power Analysis for the Behavioral Sciences* presented what is now the standard definition of effect sizes. He defined an effect size as follows:

> Without intending any necessary implication of causality, it is convenient to use the phrase "effect size" to mean "*the degree* to which the phenomenon is present in the population," or "the degree to which the null hypothesis is false." By the above route it can now readily be clear that when the null hypothesis is false, it is false to some specific degree, i.e., *the effect size (ES) is some specific non-zero value in the population*. The larger this value, the greater the *degree* to which the phenomenon

under study is manifested. (pp. 9–10, emphasis in original)

Figure 6.2 presents three hypothetical relationships that illustrate Cohen's definition. Suppose the results come from three experiments comparing the effects of an aerobic exercise intervention versus a no-treatment control on adults' cognitive functioning. The top graph presents a null relationship. That is, the participants given the intervention have a mean and distribution of cognitive functioning scores identical to the no-intervention participants. In the middle graph, the intervention group has a mean cognitive functioning score slightly higher than that of the no-intervention group, and in the bottom graph the difference between intervention and no-intervention is even greater. A measure of effect size must express the three results so that greater departures from the null are associated with larger effect size values.

Cohen's (1988) book contains many different metrics for describing the strength of a relationship. Each effect size index is associated with a particular research design in a manner similar to t-tests being associated with two-group comparisons, F-tests associated with multiple-group designs, and chi-squares associated with frequency tables. Next, I will describe the three primary metrics used by the vast majority of meta-analysts. These metrics are generally useful—almost any research outcome can be expressed using one of them. For more-detailed information on these effect size metrics, as well as many others, the reader should consult Cohen's (1988) book or Cumming's (2012) book. However, Cohen describes several metrics that permit effect size estimates for multiple-degree-of-freedom comparisons (e.g., a comparison involving more than two group means, such as three religious groups' attitudes toward rape), and these typically should not be used, for reasons that will be discussed shortly. Thus, my description of metrics is restricted to those commensurate with single-degree-of-freedom tests.

Figure 6.2 Three Hypothetical Relations Between an Exercise Intervention and a No-Intervention Group
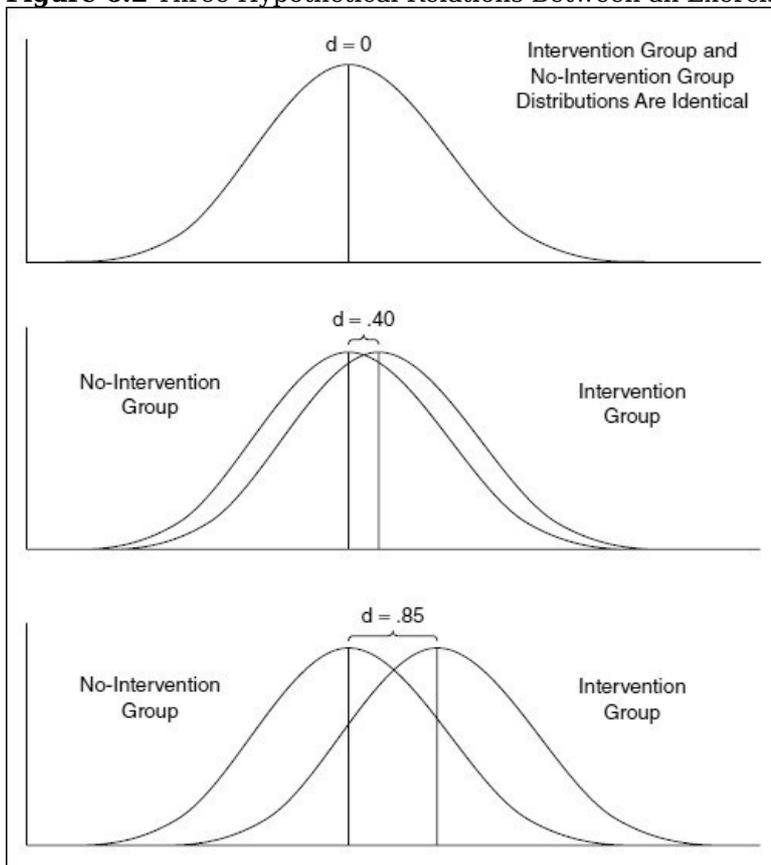
Standardized Mean Difference: The *d*-index or *g*-index

The *d*-index, or standardized mean difference measure, of an effect size is appropriate to use when the difference between two means is being compared. The *d*-index is typically used in association with t-tests or F-tests based on a comparison of two groups or experimental conditions. The *d*-index expresses the distance between the two group means in terms of their common standard deviation. By the term *common standard deviation,* I mean that the assumption is made that if we could measure the standard deviations within the two subpopulations sampled into the two groups, we would find them to be equal.

The hypothetical research results for three studies presented in Figure 6.2 comparing an intervention meant to promote aerobic activity among adults with a no-intervention condition illustrates the *d*-index. The dependent variable is some measure of neurocognitive functioning, maybe short-term memory or

speed of processing. For the top graph, the research result supports the null hypothesis and the d-index equals zero. That is, there is no distance between the means of the exercise intervention and no-intervention group. The middle research result reveals a d-index of .40—that is, the mean of the intervention group lies 4/10ths of a standard deviation to the right of the no-intervention group's mean. In the third example, a d-index of .85 is portrayed. Here, the intervention group mean rests 85/100ths of a standard deviation to the right of the mean of the no-intervention group.

Calculating the d-index is simple. The formula is as follows:

(2)

$$d = \frac{\bar{X_1} - \bar{X_2}}{SD_{within}} \tag{2}$$

where

$\bar{X_1}$ and $\bar{X_2}$ = the two group means; and
SD_{within} = the estimated common standard deviation of the two groups.

To estimate SD_{within}, you can use the formula

(3)

$$SD_{within} = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}} \tag{3}$$

where

SD_1 and SD_2 = the standard deviations of Group X_1 and Group X_2, respectively, and
n_1 and n_2 = the sample sizes in Group X_1 and Group X_2, respectively.

The d-index is not only simple to compute, but is also scale free. That is, the standard deviation adjustment in the denominator of the formula means that studies using different measurement scales can be compared or combined. So, for example, if one study of the exercise intervention's effect used a measure of short-term memory as the outcome measure and another study used a measure of processing speed as the outcome measure, it would make little sense to combine the two raw differences between the intervention and no-intervention group means—that is, combine the numerators of the d-index formula. However, it might make sense to combine the two results if we first convert each to a standardized mean difference. Then, if we assume the two outcomes measure the same underlying conceptual variable (i.e., cognitive functioning), the two outcomes have been transformed to a common metric.

The variance of the d-index can be closely approximated using the following formula:

(4)

$$v_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)} \tag{4}$$

where

all variables are defined as above.
The 95% confidence interval for the d-index is then computed as $d - 1.95 \sqrt{v_d} \le d \ge d + 1.95 \sqrt{v_d}$.

In many instances, meta-analysts will find that primary researchers do not report the means, standard deviations, and sample sizes of the separate groups but do report the t-test or F-test associated with the difference in means, and the direction of their relationship. In such cases, Rosenthal (1984) provided a computation formula that closely approximates the d-index and does not require the meta-analysts to have specific means and standard deviations. This formula is as follows:

(5)

$$d = \frac{2t}{\sqrt{df_{error}}} \tag{5}$$

where

> t = the value of the t-test for the associated comparison, and
> df_{error} = the degrees of freedom associated with the error term of the t-test ($n_1 + n_2 - 2$).

In instances where F-tests with a single degree of freedom in the numerator are reported, the square root of the F-value (i.e., $t = \sqrt{F}$) and its denominator degrees of freedom can be substituted in the above formula. Again, these approximations of the d-index assume the meta-analysts know the direction of the mean difference.

In fact, it is possible to calculate d-indexes from lots of different pieces of data and from numerous different designs. I refer you to the *Practical Meta-Analysis Effect Size Calculator* (Wilson, 2015). This free website will calculate the d-index for you based on 30 variations in the information you have available and for different research designs. Some meta-analysis software programs will also calculate effect sizes for you but you must be sure the available options match the type of data and design you are working with. If not, you can calculate the effect size using an (reliable) Internet calculator and transfer these to the meta-analysis program.

Removing small sample bias from estimates of population values: The g-index. A sample statistic—be it an effect size, a mean, or a standard deviation—typically is based on measurements taken on a small number of people drawn from a larger population. These sample statistics will differ in known ways from the values obtained if we could measure every person in the population. Meta-analysts have devised ways to adjust for the known biases that occur because effect size estimates based on samples are not always unbiased reflections of their underlying population values.

Hedges (1980) showed that the d-index based on small samples may slightly overestimate the size of an effect in the population. However, the bias is minimal if the sample size is more than 20. If meta-analysts are calculating standardized mean differences from primary research based on samples smaller than 20, Hedges' g-index should be used. The difference between the d and g formulas is simply that in the g-index formula the pooled estimate for the population standard deviation is substituted for the pooled sample standard deviation in the denominator of Formula (2). Conveniently, a search of the Internet for "Effect Size Calculators" will locate websites that will simultaneously calculate for you effect size estimates based on several different formulas (e.g., Ellis, 2009).

In addition to the small sample bias in effect size estimates meta-analysts should always be cautious in interpreting any statistics based on a small number of data points. When samples are small, a single extreme value can create an exceptionally large effect size estimate.

Choosing an estimate for the standard deviation of the d-*index*. Clearly, an important influence on the d-index is the size of the standard deviation used to estimate the variance around group means. I mentioned previously that the d-index formula is based on the assumption that the standard deviations would be equal in the two groups if they could be measured precisely. Many times, meta-analysts have no choice but to make this assumption because the d-index must be estimated from an associated t-test or F-test, which also makes this assumption. However, in instances where information about standard deviations is available and they appear to be unequal, the meta-analyst can choose one group's standard deviation to serve as the denominator in the d-index for purposes of standardizing the mean difference. For example, if an intervention and no-intervention group are being compared and the standard deviations appear to be different (perhaps because the intervention shifts the group mean and also creates greater variance in outcomes), then the control group standard deviation should be used.

Effect Sizes Based on Two Continuous Variables: The *r*-Index

A second effect size, the r-index, is simply the Pearson product-moment correlation coefficient. The r-index is the most appropriate metric for expressing an effect size when the researcher is interested in describing the relationship between two continuous variables. So, for example, if we are interested in the relationship between participants' amount of exposure to pornography and their degree of belief that women share culpability for rape, we would use the correlation coefficient to estimate this association.

The r-index is familiar to most social scientists but the formula for it requires both the variances and covariances of the two continuous variables, so it rarely can be computed from information typically presented in primary research reports. Luckily, primary researchers do report their r-indexes in most instances where they are applicable. However, if only the value of the t-test associated with the r-index is given, the r-index can be calculated using the following formula:

(6)
r=t2t2+dferror

$$r = \sqrt{\frac{t^2}{t^2 + df_{\text{error}}}} \qquad (6)$$

where

     all terms are defined as above.

The variance of the *r*-index can be calculated using the following formula:

(7)
vr=(1−r2)2n−1

$$v_r = \frac{(1-r^2)^2}{n-1} \qquad (7)$$

where

     all terms are defined as above.

The formula can be used to calculate the 95% confidence interval as $r - 1.95\ \mathrm{Vr}\sqrt{V_r} \leq r \geq r + 1.95\ \mathrm{Vr}\sqrt{V_r}$.

Normalizing the distribution of r-*indexes.* When *r*-indexes are large—that is, when they estimate population values very different from zero—they will exhibit non-normal sampling distributions. This occurs because *r*-indexes are limited to values between +1.00 and −1.00. Therefore, as a population value approaches either of these limits, the range of possible values for a sample estimate will be restricted on the tail toward the approached limit (see Shadish & Haddock, 2009).

To adjust for this, most meta-analysts convert *r*-indexes to their associated *z*-scores before the effect size estimates are combined or tested for moderators. The *z*-scores have no limiting value and are normally distributed. Conceptually, the transformation "stretches" the restricted tail of the distribution and restores the bell shape of the curve. Once an average *z*-score has been calculated, it can be converted back to an *r*-index. An examination of *r*-to-*z* transformations reveals that the two values are nearly identical until the absolute value of *r* equals about .25. However, when the *r*-index equals .50, the associated *z*-score equals .55, and when the *r*-index equals .8, the associated *z*-score equals 1.1. The *z*-score can also be calculated directly from

(8)
z=.5[ln(1+r)−ln(1−r)]

$$z = .5\left[\ln(1+r) - \ln(1-r)\right] \qquad (8)$$

where

     ln = natural logarithm and
     all other terms are defined as above.

The variance of the *z*-score is

(9)
vz=1(n−3)

$$v_z = \frac{1}{(n-3)} \qquad (9)$$

where

     all terms are defined as above.

For greatest ease, you can find *r*-to-*z* transform calculators on the Internet (e.g., http://vassarstats.net/tabs_rz.html) that will also calculate measures of dispersion. Be sure to remember that once you have calculated the average *z*-score of the transformed correlations, you must transform this back into a correlation coefficient when you present your results. The *z*-score will have little meaning for your audience.

Effect Sizes Based on Two Dichotomous Variables: The Odds and Risk Ratios

A third class of effect size metric is applicable when both variables are dichotomous—for example, when elderly adults either receive or do not receive an aerobic activity treatment and the outcome variable is whether or not they are diagnosed with Alzheimer's disease five years later. In this case, one measure of effect, called an *odds ratio*, is often used in medical research, where researchers are frequently interested in the effect of a treatment on mortality or the appearance or disappearance of disease. It is used also in criminal justice research where the outcome variable might be recidivism (re-arrest after the passage of a certain amount of time) or in education studies—for example, when high school graduation (yes or no) is the outcome of interest.

As its name implies, the odds ratio describes the relationship between two sets of odds. For example, suppose meta-analysts come across a study of the effects of an intervention promoting aerobic exercise among elderly adults. Two hundred randomly assigned participants either received or did not receive the intervention; 5 years later they were assessed for the presence of Alzheimer's disease. The results of the study were as follows:

	Intervention	No Intervention
No Alzheimer's Disease Indicated	75	60
Alzheimer's Indicated	25	40

In order to calculate an odds ratio, the meta-analysts first determine that the odds against a participant in the intervention condition having Alzheimer's disease were 3 to 1 (75 to 25). The odds against having Alzheimer's disease in the no-intervention condition were 1.5 to 1 (60 to 40). In this case, the odds ratio is 2, meaning the odds of finding evidence of the disease in the no-intervention group were twice those in the intervention group. When the odds are the same in both conditions (i.e., when the treatment had no effect or the null hypothesis was true), the odds ratio will be 1. The odds ratio can be calculated directly from the table by dividing the product of the main diagonal elements by the product of the off-diagonal elements, in our example $(75 \times 40)/(60 \times 25)$.

Another measure of effect for two dichotomous variables is the risk ratio. This expresses the relative risk of one condition against the other. So, in the example about the risk of getting Alzheimer's disease among the elderly adults who received the intervention was .25, or 25 chances in 100. For no intervention, the risk was .40, or 40 in 100. The risk ratio is then the ratio of these two numbers: .625 if the treated condition is in the numerator or 1.60 if the untreated condition is in the numerator.

Again, the *Practical Meta-Analysis Effect Size Calculator* (Wilson, 2015) can calculate both odds ratios and risk ratios for you. Similar to the *r*-index, before you calculate an average ratio, the individual ratios should be transformed to their log (also provided by the calculator). Then, the average should be transformed back for purposes of interpretation.

Because the odds ratio is used less often in the social sciences, it will not be treated extensively in the next section. However, most of the techniques discussed in the next section are easily adapted to its use. There are many other metrics that can be used when two dichotomous variables are being related to one another; Fleiss and Berlin (2009) provide an overview of numerous effect size estimates gauging the relationship between two dichotomous variables.

As general rules, I have two suggestions when you use effect size calculators available on the Internet. First, check the formulas used in these programs. They might differ in some ways from my simple formulas given above. As long as the website comes from a reliable source, the calculations should be reliable but it is always good to calculate a few effect sizes by hand. This way you can be more confident you understand how your data are being analyzed by the software program.

Practical Issues in Estimating Effect Sizes

The formulas for calculating effect sizes are straightforward. In practice, however, meta-analysts face many technical issues when they attempt to calculate a magnitude of effect. The most important of these is missing data, which I discussed in Chapter 4 and will return to again in the next chapter. Other issues arise because different studies use somewhat different designs and because of some unique characteristics of the effect size metrics themselves. I will describe a few of these.

Choosing a metric when studies have different designs. Some primary researchers use parametric statistics (those that assume normal distributions) and others use nonparametric statistics (ones that make no assumptions about distributions) to test and express the same relationship. For instance, this would be the case if one researcher measured intrinsic motivation in a choice study by calculating the average time each participant spent on the chosen task during a free-play period (a continuous variable dictating the use of parametric tests), and another simply recorded whether each participant did or did not choose a particular task during a free-play period (a dichotomous variable dictating use of nonparametric tests). Most often, in a research literature statistical techniques based on one set of assumptions will predominate greatly over the other. Then, the statistics from the lesser-used approach can be converted to their dominant-approach equivalents and aggregated as though they shared the dominant approach's assumptions. As long as the number of these conversions is small, there will be no great distortion of results. If there are substantive reasons to distinguish between the outcome variables or if the split between parametric and nonparametric tests is relatively even, the two sets of studies might be meta-analyzed separately.

Related to the issue of studies that use different statistical procedures is that different primary

researchers sometimes convert continuous variables to dichotomous ones. For instance, some primary researchers studying the relation between individual differences and attitudes toward rape might dichotomize personality scores into high and low scoring groups. Then, they might use a t-test to determine if the high and low group means were different on a continuous measure of attitudes toward rape. This suggests that a d-index would be most appropriate to estimate the relation. However, other researchers might leave the same personality scale in its continuous form and report the correlation between them. Conveniently, the different effect size metrics are easily converted from one to the other. The r-index can be transformed into a d-index using the following formula:

(10)
d=2r1−r2

$$d = \frac{2r}{\sqrt{1-r^2}} \qquad (10)$$

or the d-index into the r-index using

(11)
r=dd2+a

$$r = \frac{d}{\sqrt{d^2 + a}} \qquad (11)$$

where

  a = a correction factor to adjust for different sample sizes between the two groups.

This correction factor, a, can be calculated using this formula:

(12)
a=(n1+n2)2n1n2

$$a = \frac{(n_1 + n_2)^2}{n_1 n_2} \qquad (12)$$

where

  all variables are defined as above.

When a chi-square statistic associated with a 2 × 2 contingency table is given, the r-index can be estimated as follows:

(13)
r=χ2n

$$r = \sqrt{\frac{\chi^2}{n}} \qquad (13)$$

where

  χ^2 = the chi-square value associated with the comparison, and
  n = the total number of observations in the comparison.

If you search the Internet using "effect size converter," you will find several websites that will allow you to easily convert between different effect size metrics.

Even though metrics can be converted easily, meta-analysts still must pick a single metric in which to describe their results. The choice of how to express the effect size should be determined by which metric best fits with the measurement and design characteristics of the variables under consideration. So, *the effect size metric used should be based on the characteristics of the conceptual variables.* Therefore, an important question to ask when evaluating a research synthesis is,

  If a meta-analysis was performed, was an appropriate effect size metric used?

When we related individual differences to rape attitudes, the r-index was appropriate most often (e.g., when personality dimensions were of interest) because the two variables were conceptually continuous in nature. If a study created two artificial groups by dichotomizing the continuous individual difference measure into high and low scorers, we would calculate a d-index comparing the group means, then convert it to an r-index using Formula (11).

Estimating effect sizes when studies compare more than two groups. Suppose we find a study of

interventions to promote aerobic exercise that compared three groups—say, an exercise group, an information group, and a no-intervention group. In this instance, we likely would calculate two d-indexes, one comparing exercise to no-intervention and another comparing the exercise intervention to the information intervention (we could also consider comparing the information intervention to no intervention, if this were the focus of our meta-analysis).[3] These two d-indexes are not statistically independent since both rely on the means and standard deviations of the same intervention group. However, this complicating factor is preferable to the alternative strategy of using an effect size metric associated with a multiple-group inference test. Here is why.

One effect size metric that can be used when more than two groups are being compared simultaneously involves calculating the percentage of variance in the dependent variable explained by group membership. This effect size has the initially appealing characteristic that it can be used regardless of the number of groups in the study (indeed, it can be used with two continuous measures as well). So, it is very generally applicable. However, it has the unappealing characteristic that the resulting effect size tells us nothing about which of the multiple conditions has the highest mean, or, more specifically, how the values of the means are ordered and how much each differs from the others. So, identical percentages of variance explained can result from different rank ordering of, and distances between, the group means. It is then impossible for the meta-analysts to draw conclusions about how the different groups stack up relative to one another. In fact, the results might cancel one another out if we looked at single-degree-of-freedom comparisons, suggesting no differences between groups. The percentage of variance explained would not catch this. This is why it is rarely, if ever, used by meta-analysts.

Estimating effect sizes from analyses including multiple predictor variables. Another way that research design influences effect sizes involves the number of factors employed in the primary data analysis procedures. For example, a primary researcher testing the effect of homework versus no-homework on achievement might also include individual difference variables—such as the sex or previous achievement of the students, or even their pretest scores on the outcome measure—in a multi-factored analysis of variance. The primary researcher also might not report the simple means and standard deviations for the homework and no-homework groups. Meta-analysts then are faced with two choices.

First, they can calculate an effect size estimate based on the F-test reported by the researchers. However, this test uses an error term that has been reduced by the inclusion of the individual difference factors. This is equivalent to reducing the size of the estimate of S_{within} in the d-index formula. This approach creates the problem that different effect sizes going into the same quantitative synthesis are likely to be known to differ in a systematic way—that is, in how the within-group standard deviation has been calculated. Likely, if the additional factors in the analysis are associated with variance in the outcome measure (e.g., the scores on a unit test), then this study will produce a larger effect size for homework than a study that did not include these additional factors in the analysis, all else being equal.

A second approach is to attempt to retrieve the standard deviations that would have occurred had all the extraneous factors been ignored (i.e., not been removed from the error term used to calculate the F-test). Whenever possible, this strategy should be used—that is, an attempt should be made to calculate the effect size as though the comparison of interest was the sole comparison in the analysis. The best way to do this is to contact the authors of the primary research and see if they will share the data you need. Perhaps a more realistic approach is to adjust the effect size by estimating the relationships between the additional variables and the outcome measure. Borenstein et al. (2009) present some ways to calculate these estimates. The problem here, of course, is that the resulting estimate of the effect size is only as good as the estimates of the relationships used to make adjustments.

Practically speaking, then, it is often difficult for meta-analysts to retrieve the unadjusted standard deviation estimates for the two groups if they are not given in the primary research report, nor is a simple t-test or one-degree-of-freedom F-test. In such cases, when you look for influences on study outcomes, you should either (a) leave these estimates out, if they are few, or (b) examine whether or not the number of factors included in the analysis is associated with the size of the effect. If a relation is found, you should report separately the results obtained from analyses of studies that used only the single factor of interest. So, for example, in the meta-analysis of homework research, we found one experimental study that reported the effect of homework only in an analysis of covariance with several covariates. This study's results could not be combined with studies that did not adjust for covariates. We also found other studies that presented results regarding the relation between time spent on homework and achievement only in multiple regression analyses. These could not be combined with the studies that presented simple bivariate correlations.

Adjusting for the impact of methodological artifacts. The magnitude of an effect size will also be influenced by the presence of methodological artifacts in the primary data collection procedures. Schmidt and Hunter (2015) describe 10 such artifacts that can make an effect size smaller than it might otherwise be. These include, for example, errors (lack of reliability) in the measurement of the independent and dependent variable, imperfect construct validity of measures, dichotomizing of continuous variables, and restrictions in the range of sampled values.

In the case of less-reliable measures, measures with more error are less sensitive for detecting relationships involving its conceptual variables. For example, assume two personality dimensions have equal true relationships with attitudes toward rape. However, if one personality variable is measured with more error than the other, this less-reliable measure will produce a smaller correlation, all else being equal. So you might estimate the impact of the reliability of measures on effect sizes by obtaining the reliabilities (e.g., internal consistencies) of the various measures. Or, if the reliabilities of some measures were not available you could estimate the distribution (mean and standard deviation) of the reliabilities. Using procedures described by Schmidt and Hunter (2015), you could then estimate what the average effect sizes would be if all measures were perfectly reliable. You could also calculate a credibility interval, the estimated standard deviation of the disattenuated effect sizes.

Whether effect sizes should be corrected for artifacts depends first and foremost on the goal of the primary research and research synthesis. In particular, are you interested in the relationship between the constructs that underlie the measures or in what can be expected in the real world? For example, the amount of homework students do and their subsequent achievement may be imperfectly measured but if the synthesis is meant to describe what effect of homework parents, teachers, and student might expect on test scores, correcting for artifacts is inappropriate.[4] On the other hand, the meta-analysis of studies of the effect of choice on motivation might legitimately correct for unreliability in the motivation measures because they are interested in testing a theoretical notion. Error in the measurements might lead to accepting a null hypothesis when, in fact, it should be rejected.

In addition, you should keep in mind that when you correct for artifacts, your results are only as good as your estimates of the impact of the artifact. If the measures of artifacts are unreliable or you must estimate the distribution of artifact effects based on limited data, it might be good to perform a sensitivity analysis—that is, to conduct your analyses with high and low estimates of the artifact correction to see how your results differ.

Coding Effect Sizes

The statistics you need to calculate effect sizes and all the other statistics described next should be collected as part of your more general coding procedures. For example, Table 6.2 provides a simple example of the information on the statistical results of studies that might be collected by study coders. Here, the example involves experimental studies of the effects of homework on achievement. Most meta-analyses in which two conditions are being compared (having a choice among tasks, participation versus no participation in an exercise intervention) would look very similar. Coding sheets for correlational studies or studies relating two dichotomous variables would also be similar, but these might be even a bit simpler than my example in Table 6.2. Some of the information on the coding sheet may never be used and much of this information will be left blank. For example, when studies give the means and standard deviations, you may never use the information on the *t*-test. However, when means and/or standard deviations are missing, you will need the information on the null hypothesis significance test to calculate the *d*-index. Or if you want to examine whether the standard deviations in the experimental and control group are roughly equal, you will need this regardless of how you calculate the *d*-index. So, you might not know exactly what information is important to you until after you have begun your analysis.

Table 6.2   An Example Coding Sheet for the Statistical Outcomes of Experimental Studies on the Effects of Homework on Achievement

Effect Size Estimate	

E1. What was the direction of the effect of homework on the achievement measure?
    + = positive
    − = negative
	___

E2. Information about each experimental group (Note: Leave blank if not reported. M = Mean. SD = standard deviation.)	
Homework Group	
a. Pretest M on outcome (if any)	__ __ __ . __ __
b. Pretest SD	__ __ __ . __ __
c. Posttest M on outcome	__ __ __ . __ __
d. Posttest SD	__ __ __ . __ __
e. Sample size	__ __ __
No-Homework Group	
f. Pretest M on outcome (if any)	__ __ __ . __ __

Table 6.2 (Continued)

Effect Size Estimate	
g. Pretest *SD*	___ ___ ___ . ___ ___
h. Posttest *M* on outcome	___ ___ ___ . ___ ___
i. Posttest *SD*	___ ___ ___ . ___ ___
j. Sample size	___ ___ ___
k. Total sample size (if not given for each group separately)	___ ___ ___ ___
E3. Information about null hypothesis significance tests	
a. Value of independent *t*-statistic (or square root of *F*-test in one-factor ANOVA)	___ ___ . ___ ___
b. Degrees of freedom for test (in the denominator)	___ ___ ___
c. *p*-value from test	< . ___ ___ ___
d. Dependent *t*-statistic	___ ___ . ___ ___
e. Degrees of freedom for test (in the denominator)	___ ___ ___
f. *p*-value from test	< . ___ ___ ___
g. *F*-statistic (when included in a multifactored ANOVA)	___ ___ . ___ ___
h. Degrees of freedom for denominator of *F*-test	___ ___ ___
i. *p*-value from *F*-test	< . ___ ___ ___
j. # of variables in multifactored ANOVA	___
E4. Effect Size estimate	
a. What is the metric of the effect size (*d*, *r*, OR, RR, other)	___
b. Was an effect size calculator used to calculate this effect? 0 = No 1 = Yes	___
If yes, what calculator was used?	_____

Combining Effect Sizes Across Studies

Once each effect size has been calculated, the meta-analysts next average the effects that estimate the same comparison or relationship. It is generally accepted that these averages should weight the individual effect sizes based on the number of participants in their respective samples. This is because larger samples give more precise population estimates. For example, a *d*-index or *r*-index based on 500 participants will give a more precise estimate of its underlying population effect size than will an estimate based on 50 participants. The average effect size should reflect this fact. So, while unweighted average effect sizes are sometimes presented in meta-analyses, they are typically accompanied by weighted averages.

One way to take the precision of the effect size estimate into account when calculating an average effect size is to multiply each estimate by its sample size and then divide the sum of these products by the sum of the sample sizes. However, there is a more precise procedure, first described in detail by Hedges and Olkin (1985), which has many advantages but also involves more complicated calculations.

The *d*-Index

For the *d*-index, this procedure first requires the meta-analyst to calculate a weighting factor, w_i, which is the inverse of the variance associated with each *d*-index estimate. It can be calculated taking the inverse

of the result of Formula (4), or more directly by using the following formula:

(14)
Wi=2(ni1+ni2)ni1ni22(ni1+ni2)2+ni1ni2di2

$$W_i = \frac{2(n_{i1} + n_{i2})n_{i1}n_{i2}}{2(n_{i1} + n_{i2})^2 + n_{i1}n_{i2}d_i^2} \qquad (14)$$

where

> n_{il} and n_{i2} = the number of data points in Group 1 and Group 2 of Study i; and
> d_i = the d-index of the comparison under consideration.

While the formula for w_i looks imposing, it is really a simple arithmetic manipulation of three numbers available whenever a d-index is calculated. It also is easy to program a statistical software package to perform the necessary calculation. Programs designed to perform meta-analysis (e.g., Comprehensive Meta-Analysis, 2015) will do it for you automatically.

Table 6.3 presents the group sample sizes, d-indexes, and weighting factors (the w_is) associated with the results of seven hypothetical comparisons. Let us assume the seven comparisons come from experiments that compared the effects of homework versus no homework on a measure of academic achievement. All seven of the experiments produced results favoring homework assignments. The results could just as easily have come from seven comparisons of groups doing aerobic exercise or not, and the measure could be cognitive functioning. Or, the participants in one group in Table 6.3 could have been given a choice between two tasks while the other group was given no choice and the outcome could be subsequent interest in the task. It is good to look at the hypothetical data with multiple concrete examples in your head. That way you can see the conceptual similarity between the examples. The key here is that you recognize that the research design in this table compares two group means on a continuous variable. If for some reason the outcome variable was a dichotomy (Did the student pass the course? Did the elderly get Alzheimer's disease? Did the subject choose the task during free time?) but the majority of outcomes were continuous, the odds or risk ratio could have been converted to a d-index and the study included along with the others.

To further demystify the weighting factor, note in Table 6.3 that its values equal approximately half the average sample size in a group (it becomes less similar to half the average sample size as the sample sizes in the two groups become more different). It should not be surprising, then, that the next step in obtaining a weighted average effect size involves multiplying each d-index by its associated w_i and dividing the sum of these products by the sum of the weights. This is done using the following formula:
d.=Σi=1kdiwiΣi=1kwi

$$d. = \frac{\sum_{i=1}^{k} d_i w_i}{\sum_{i=1}^{k} w_i} \qquad (15)$$

where

> k = the total number of comparisons and
> all other terms are defined as above.

Table 6.3 An Example of *d*-Index Estimation and Tests of Homogeneity

Study	n_{i1}	n_{i2}	d_i	w_i	$d_i^2 w_i$	$d_i w_i$	Q_b Grouping
1	259	265	.02	130.98	.052	2.619	A
2	57	62	.07	29.68	.145	2.078	A
3	43	50	.24	22.95	1.322	5.509	A
4	230	228	.11	114.32	1.383	12.576	A
5	296	291	.09	146.59	1.187	13.193	B
6	129	131	.32	64.17	6.571	20.536	B
7	69	74	.17	35.58	1.028	6.048	B
5	1083	1101	1.02	544.27	11.69	62.56	

NOTE: Weighted average $d. = 62.56/544.27 = +.115$;

$$CI_{d.95\%} = .115 \pm 1.96 \sqrt{\frac{1}{544.27}} = .115 \pm .084;$$

$$Q_t = 11.69 - \frac{62.56^2}{544.27} = 4.5;$$

$$Q_w = 1.69 + 2.36 = 4.05;$$

$$Q_b = 4.5 - 4.05 = 0.45$$

Table 6.3 shows the average weighted *d*-index for the seven comparisons is $d. = .115$.

One advantage of using the w_is as weights, rather than sample sizes, is that the w_is can also be used to generate a confidence interval around the average effect size estimate. To do this, an estimated variance for the average effect size must be calculated. First, the inverse of the sum of the w_is is found. Then, the square root of this variance is multiplied by the *z*-score associated with the confidence interval of interest. Thus, the formula for a 95% confidence interval is

(16)
$$CI_{d.95\%} = d. \pm z_i \sqrt{\frac{1}{\sum_{i=1}^{k} w_i}}$$

$$CI_{d.95\%} = d. \pm z_i \sqrt{\frac{1}{\sum_{i=1}^{k} w_i}} \qquad (16)$$

where

> z_i = the *z*-score associated with the confidence interval of interest and
> all terms are defined as above.

Table 6.3 reveals that the 95% confidence interval for the seven homework comparisons encompasses values of the *d*-index .084 above and below the average *d*-index. Thus, we expect 95% of estimates of this effect to fall between $d = .031$ and $d = .199$. Note that this interval does not contain the value $d = 0$. It is this information that can be taken as a test of the null hypothesis that no relation exists in the population, in place of directly combining the significance levels of null hypothesis tests. In this example, we would reject the null hypothesis that there was no difference in achievement between students who did and did not do homework.

The *r*-Index

The procedure for finding the average weighted *r*-index and its associated confidence interval is similar. Here, I will illustrate how to do this when each *r*-index is first transformed to its corresponding *z*-score, z_i. In this case, the following formula is applied:

(17)
$$z. = \frac{\sum_{i=1}^{k}(n_i - 3)z_i}{\sum_{i=1}^{k}(n_i - 3)}$$

$$z. = \frac{\sum_{i=1}^{k}(n_i - 3)z_i}{\sum_{i=1}^{k}(n_i - 3)} \qquad (17)$$

where

n_i = the total sample size for the ith comparison and

all other terms are defined as above.

Notice that formulas for calculating average effect sizes all follow the same form: multiply the effect size by a weight, sum the products, and divide by the sum of the weights. So, to combine the r-indexes directly, multiply each by its weighting factor—in this case, like the d-index, it is the inverse of its variance (Formula [7])—and divide the sum of this product by the sum of the weights, just as was done for the d-index.

To obtain a confidence interval for the average z-score, the formula is

(18)

$$CI_{z.95\%} = z. \pm \frac{1.96}{\sqrt{\sum_{i=1}^{k}(n_i - 3)}} \qquad (18)$$

where

all terms are defined as above.

To obtain a confidence interval for the r-indexes combined directly, simply substitute the sum of the weights in the denominator of Formula (18).

Remember that it is important to transform your r-indexes to z-scores before you begin to combine them, especially if many of the correlations are above .25. Once the confidence interval has been established, meta-analysts convert the z-scores back to the correlations.

Table 6.4 presents an example of how average r-indexes are calculated. For example, the six correlations might come from studies relating participants' individual differences on authoritarianism and their score on a measure of rape myth acceptance. Or, the correlations might be between time spent on homework and a unit test score. Again, the key here is that both measures are continuous. The average z_i was 207 with the 95% confidence interval ranging from .195 to .219. Note that this confidence interval is quite narrow. This is because the effect size estimates are based on large samples. Note also that the r-to-z transformations result in only minor changes in two of the r-index values. This would not be the case had the r-indexes been larger. As with the earlier example, $z_i = 0$ is not contained in the confidence interval. Therefore, we can reject the null hypothesis that there is no relation between participants' individual differences on authoritarianism and their scores on a measure of rape myth acceptance (or, on time spent on homework and a unit test score).

Table 6.4 An Example of r-Index (Transformed to z) Estimation and Tests of Homogeneity

Study	n_i	r_i	z_i	$n_i - 3$	$(n_i - 3)z_i$	$(n_i - 3)$ z_i^2	Q_b Grouping
1	3,505	.06	.06	3,502	210.12	12.61	A
2	3,606	.12	.12	3,603	432.36	51.88	A
3	4,157	.22	.22	4,154	913.88	201.05	A
4	1,021	.08	.08	1,018	81.44	6.52	B
5	1,955	.27	.28	1,952	546.56	153.04	B
6	12,146	.26	.27	12,143	3278.61	885.22	B
5	26,390	1.01	1.03	26,372	5462.97	1310.32	

NOTE: Weighted average $z. = \dfrac{5462.97}{26,372} = .207;$

$CI_{z.95\%} = .207 \pm 1.96\sqrt{26,372} = .207 \pm .012;$

$Q_t = 1310.32 - \dfrac{(5462.97)^2}{26,372} = 178.66;$

$Q_w = 34.95 + 50.40 = 85.35;$

$Q_b = 178.66 - 85.35 = 93.31.$

In sum, each of the effect size metrics can be averaged across studies and confidence intervals can be placed around these mean estimates. Therefore, when evaluating a research synthesis, it is important to ask,

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?

A Note on Combining Slopes From Multiple Regressions

Up to this point, the procedures for combining and comparing study results have assumed that the measure of effect is a difference between means, a correlation, or an odds ratio. However, regression analysis is a commonly used technique in the social sciences, particularly in nonexperimental studies where many variables are used to predict a single criterion. Similar to the standardized mean difference or correlation coefficient, the regression coefficient, *b,* or the standardized regression coefficient, *β,* are also measures of effect size. *β* will typically be of most interest to meta-analysts because, like the *d*-index and *r*-index, it standardizes effect size estimates when different measures of the same conceptual variable are used in different studies. *β* represents the change in a standardized predictor variable, controlling for all other predictors, given one standard unit change in the criterion variable.

Meta-analyses using regression coefficients as effect sizes are difficult to conduct for a variety of reasons. First, with regard to using the unstandardized *b*-weight, this is like using raw score differences as measures of effect—the scales of the predictor and outcome of interest typically vary across studies. Directly combining them can lead to uninterpretable results. This problem can be overcome by using *β,* the fully standardized estimate of the slope for a particular predictor.[5] But still, the other variables included in models using multiple regression generally differ from study to study (note the related earlier discussion about multifactored analyses of variance). Each study may include different predictors in the regression model and, therefore, the slope for the predictor of interest will represent a different partial relationship in each study (Becker & Wu, 2007). For example, in our meta-analysis of homework and achievement, we found numerous studies that performed analyses of the relationship between time spent on homework and achievement that reported *β.* However, each was based on a regression model that included different additional variables. This made it questionable that the *β*s should be directly combined. So, rather than average them, we described these studies' individual *β*s and the range of *β*-values across the studies. These were overwhelmingly positive, were generally based on very large samples, and used a variety of achievement outcome measures. As such, they strengthened our claim about the positive effects of homework on achievement that was based on the few small studies that purposively manipulated homework and tested its effect on a single limited outcome measure, unit test scores.

Regression slopes can be directly combined when (a) the outcome and predictor of interest are measured in a similar fashion across studies, (b) the other predictors in the model are the same across studies, and (c) the predictor and outcome scores are similarly distributed (Becker, 2005). It is rare that all three of these assumptions are met; typically, measures differ across studies and regression models are diverse in terms of which additional variables are included in them.

The Synthesis Examples

Both the standardized mean difference and the correlation coefficient measures of effect size were used in the synthesis examples. In the synthesis of the effects of homework, the *d*-index was used to express the findings from comparisons that purposively manipulated homework and then measured the difference in terms of unit test scores. The weighted average *d*-index across five studies was *d.* = .60, with a 95% confidence interval encompassing values from *d* = .38 to *d* = .82. Clearly, then, the null hypothesis could be rejected. The homework research synthesis also used correlation coefficients to estimate the relationship between student or parent reports of the amount of time spent on homework and a variety of measures of achievement. Of 69 such correlations, 50 were positive and 19 were negative. The weighted average correlation was *r* = .24 with a very narrow 95% confidence interval, encompassing the values between .24 and .25. The confidence interval was so small because of the large number of participants in these studies; the adjusted mean sample size in the studies was 7,742.

The meta-analysis of individual differences and attitudes toward rape also used correlation coefficients as the measure of the strength of relationships. Among the many correlations involving individual differences, we found, for example, that across 15 correlations older participants were more accepting of rape than younger ones, average *r.* = .12 (95% CI = .10−.14).

The meta-analyses on (a) interventions to increase aerobic exercise among adults and (b) the effects of choice on intrinsic motivation used the standardized mean difference to measure effects. The weighted average *g*-index across 29 studies indicated that adults who participated in the interventions revealed improvements in attention and processing speed, *g.* = .158 (95% CI = .055−.260), executive functioning, *g.* = .123 (95% CI = .021−.225), and memory, *g.* = .128 (95% CI = .015−.241). The average weighted effect size for the 47 estimates of the impact of choice on measures of intrinsic motivation was *d.* = .30 (95% CI = .25−.35), indicating choice led to greater intrinsic motivation.

Analyzing Variance in Effect Sizes Across Findings

The analytic procedures described thus far have illustrated how to estimate effect sizes, average them, and use the confidence interval surrounding the average to test the null hypothesis that the difference between two means or the size of a correlation is 0. Another set of statistical techniques helps meta-analysts discover why effect sizes vary from one comparison to another. In these analyses, the effect sizes found in the separate comparisons are the *dependent* or predicted variables and the characteristics of the comparisons are the predictor variables. The meta-analysts ask whether the magnitude of relation between two variables in a comparison is affected by the way the study was designed or carried out.

One obvious feature of the effect sizes in Tables 6.3 and 6.4 is that they vary from comparison to comparison. An explanation for this variability is not only important, but also represents the most unique contribution of research synthesis. By performing an analysis of differences in effect sizes, the meta-analyst can gain insight into the factors that affect the strengths of relationships even though these factors may have never been studied in a single experiment. For instance, assume that the comparisons were looking at the effects of homework and the first four studies listed in Table 6.3 were conducted in elementary schools while the last three studies were conducted in high schools. Is the effect of homework different for students at different grades? This question could be addressed through the use of the analytic techniques described next, even though no single study included both elementary and high school students and tested to see if the grade level of students moderated the effect of homework.

The techniques that follow are a few examples of many procedures for analyzing variance in effect sizes. I do not cover some of the more complex synthesis techniques but will return to them after exposition of the most frequently used meta-analysis techniques.

Traditional Inferential Statistics

One way to analyze the variance in effect sizes is to apply the traditional inference procedures that are used by primary researchers. Meta-analysts interested in whether an exercise intervention's effects on older adults' cognitive functioning were stronger for males than for females might do a *t*-test on the difference between effect sizes found in comparisons exclusively using males versus comparisons exclusively using females. Or, if the meta-analysts were interested in whether the intervention effect size was influenced by the length of the intervention and the measurement of cognitive functioning, the meta-analysts might correlate the length of treatment in each comparison with its effect size. In this instance, the predictor and dependent variables are continuous, so the significance test associated with the correlation coefficient would be the appropriate inferential statistic. For more complex questions, a synthesist might categorize effect sizes into multifactor groupings—for instance, according to the gender and age of participants—and perform an analysis of variance or multiple regression on effect sizes. For Table 6.3, if a one-way analysis of variance were conducted comparing the first four *d*-indexes with the last three *d*-indexes, the result would not be statistically significant.

Standard inference procedures were the techniques initially used by some meta-analysts for examining variance in effects. Glass et al. (1981) detailed how this approach is carried out. However, at least two problems arise with the use of traditional inference procedures in meta-analysis. The first is that traditional inference procedures do not test the hypothesis that the variability in effect sizes is due solely to sampling error (recall the discussion earlier in this chapter). Therefore, the traditional inference procedures can reveal associations between design characteristics and effect sizes without determining first whether the overall variance in effects is greater than that expected by sampling error alone.

Also, because effect sizes can be based on different numbers of data points (sample sizes), they can have different sampling variances associated with them—that is, they are measured with different amounts of error, or differing levels of precision. If this is the case (and it often is), then the effect sizes violate the assumption of homogeneity of variance that underlies traditional inference tests. For these two reasons, traditional inferential statistics are no longer used when performing a meta-analysis.

Comparing Observed to Expected Variance: Fixed-Effect Models

In place of traditional procedures, several approaches have gained acceptance. One approach is called the *fixed-effect model.* I will explain this simplest model first and then explain a second more complex model, called the *random-effects model.* The fixed-effect model compares the variation in the observed effect sizes with the variation expected if only error due to the sampling of participants were causing differences in effect size estimates. In other words, it makes the assumption that there is one value of the effect size underlying all the observations and the only thing making the observations different is differences in the participants sampled into each study. This approach involves calculating (a) the observed variance in the effect sizes from the known findings and (b) the expected variance in these effect sizes given that all are estimating the same underlying population value. Sampling theory allows us to calculate precise estimates of how much sampling variation to expect in a group of effect sizes if only differences between the participants is making the effect sizes different. This expected value is a function of the average effect size estimate, the number of estimates, and their sample sizes.

The meta-analysts then compare the observed with the expected variance. If the variance estimates are deemed not to differ then sampling error of participants is the simplest explanation for the variance in effect sizes. If they are deemed different—that is, if the observed variance is (significantly) greater than that expected due to sampling error of participants, then the meta-analysts begin the search for systematic influences on effect sizes. This is done by grouping the effect sizes and asking whether the group averages are more different than sampling error alone would predict.

Homogeneity Analyses

A homogeneity analysis is a formal way to compare the observed variance to that expected from sampling error. It involves the calculation of how probable it is that the variance exhibited by the effect sizes would be observed if only sampling error was making them different. This is the approach used most often by meta-analysts, so I will provide a few more of its details.

Homogeneity analysis first asks the question, "Is the observed variance in effect sizes statistically significantly different from that expected by sampling error alone?" If the answer is "no," then some statisticians advise that the meta-analysts stop the analysis there. After all, chance or sampling error is the simplest and most parsimonious explanation for why the effect sizes differ. If the answer is yes—that is, if the effect sizes display significantly greater variability than expected by chance, the meta-analysts then begin to examine whether study characteristics are systematically associated with variance in effect sizes. Some meta-analysts believe that the search for moderators should proceed regardless of whether sampling error is rejected as a plausible sole cause of variability in effect sizes *if* there are good theoretical or practical reasons for choosing moderators. This is the approach I usually take. Regardless of the approach you prefer, when evaluating a research synthesis, it is important to ask,

If a meta-analysis was performed, was the homogeneity of effect sizes tested?

Suppose a meta-analysis reveals a homogeneity statistic that has an associated *p*-value of .05. This means that only 5 times in 100 would sampling error create this amount of variance in effect sizes. Thus, the meta-analysts would reject the null hypothesis that sampling error alone explains the variance in effect sizes and they would begin the search for additional influences. They would then test whether study characteristics explain variation in effect sizes. Studies would be grouped by common features, and the average effect sizes for groups would be tested for homogeneity in the same way as the overall average effect size.

An approach to homogeneity analysis will be described that was introduced simultaneously by Rosenthal and Rubin (1982) and Hedges (1982). The formula presented by Hedges and Olkin (1985; also see Hedges, 1994) will be given here and the procedures using *d*-indexes will be described first.

The d-*index.* In order to test whether a set of *d*-indexes is homogeneous, the meta-analysts must calculate a statistic Hedges and Olkin (1985) called Q_t. The formula is as follows:

(19)
$$Q_t = \sum_{i=1}^{k} w_i d_i^2 - \frac{\left(\sum_{i=1}^{k} w_i d_i\right)^2}{\sum_{i=1}^{k} w_i}$$

$$Q_t = \sum_{i=1}^{k} w_i d_i^2 - \frac{\left(\sum_{i=1}^{k} w_i d_i\right)^2}{\sum_{i=1}^{k} w_i} \qquad (19)$$

where all terms are defined as above.

The *Q*-statistic has a chi-square distribution with $k - 1$ degrees of freedom, or, one less than the number of comparisons. The meta-analysts refer the obtained value of the total *Q* statistic, Q_t, to a table of (upper tail) chi-square values. If the obtained value is greater than the critical value for the upper tail of a chi-square at the chosen level of significance, the meta-analysts reject the hypothesis that the variance in effect sizes was produced by sampling error alone. Table 6.5 presents the critical values of chi-square for selected probability levels.

Table 6.5 Critical Values of Chi-Square for Given Probability Levels

DF	\.500	\.250	\.100	\.050	\.025	\.010
	Upper Tail Probabilities					
1	.455	1.32	2.71	3.84	5.02	6.63
2	1.39	2.77	4.61	5.99	7.38	9.21
3	2.37	4.11	6.25	7.81	9.35	11.3
4	3.36	5.39	7.78	9.49	11.1	13.3
5	4.35	6.63	9.24	11.1	12.8	15.1
6	5.35	7.84	10.6	12.6	14.4	16.8
7	6.35	9.04	12.0	14.1	16.0	18.5
8	7.34	10.2	13.4	15.5	17.5	20.1
9	8.34	11.4	14.7	16.9	19.0	21.7

DF	.500	.250	.100	.050	.025	.010
	Upper Tail Probabilities					
10	9.34	12.5	16.0	18.3	20.5	23.2
11	10.3	13.7	17.3	19.7	21.9	24.7
12	11.3	14.8	18.5	21.0	23.3	26.2
13	12.3	16.0	19.8	22.4	24.7	27.7
14	13.3	17.1	21.1	23.7	26.1	29.1
15	14.3	18.2	22.3	25.0	27.5	30.6
16	15.3	19.4	23.5	26.3	28.8	32.0
17	16.3	20.5	24.8	27.6	30.2	33.4
18	17.3	21.6	26.0	28.9	31.5	34.8
19	18.3	22.7	27.2	30.1	32.9	36.2
20	19.3	23.8	28.4	31.4	34.2	37.6
21	20.3	24.9	29.6	32.7	35.5	33.9
22	21.3	26.0	30.8	33.9	36.8	40.3
23	22.3	27.1	32.0	35.2	38.1	41.6
24	23.3	28.2	33.2	36.4	39.4	43.0
25	24.3	29.3	34.4	37.7	40.6	44.3
26	25.3	30.4	35.6	38.9	41.9	45.6
27	26.3	31.5	36.7	40.1	43.2	47.0
28	27.3	32.6	37.9	41.3	44.5	48.3
29	28.3	33.7	39.1	42.6	45.7	49.6
30	29.3	34.8	40.3	43.8	47.0	50.9
40	49.3	45.6	51.8	55.8	59.3	63.7
60	59.3	67.0	74.4	79.1	83.3	88.4
	.500	.750	.900	.950	.975	.990
	Lower Tail Probabilities					

For the set of comparisons given in Table 6.3, the value of Q_t equals 4.5. The critical value for chi-square at $p < .05$ based on 6 degrees of freedom is 12.6. Therefore, the hypothesis that sampling error explains the differences in these d-indexes cannot be rejected.

The procedure to test whether a methodological or conceptual distinction between studies explains variance in effect sizes involves three steps. First, a Q-statistic is calculated separately for each subgroup of comparisons. For instance, to compare the first four d-indexes in Table 6.3 with the last three, a separate Q-statistic is calculated for each grouping. Then, the values of these Q-statistics are summed to form a value called Q_w, or Q-within. This value is then subtracted from Q_t to obtain the Q statistic for the difference between the two group means, Q_b, or Q-between:

(20)
Qb=Qt–Qw

$$Q_b = Q_t - Q_w \qquad (20)$$

where

      all terms are defined as above.

The statistic Q_b is used to test whether the *average* effects from the two groupings are homogenous. It is compared to a table of chi-square values using as degrees of freedom one less than the number of groupings. If the average d-indexes are homogeneous, then the grouping factor does not explain variance in effects beyond that associated with sampling error. If Q_b exceeds the critical value, then the grouping factor is a significant contributor to variance in effect sizes.

In Table 6.3 the Q_b comparing the first four and last three d-indexes is .45. This result is not significant with one degree of freedom. So, if the first four effect sizes were taken from studies of the effect of homework on achievement using elementary school students and the last three using high school students, we could not reject the null hypothesis that effect sizes were equal in the two populations of students.

The r-index. The analogous procedure for performing a homogeneity analysis on r-indexes transformed to z-scores involves the following formula:

(21)
Qt=Σi=1k(ni−3)zi2−[Σi=1k(ni−3)zi]Σi=1k(ni−3)2

$$Q_t = \sum_{i=1}^{k}(n_i-3)z_i^2 - \frac{\left[\sum_{i=1}^{k}(n_i-3)z_i\right]^2}{\sum_{i=1}^{k}(n_i-3)} \qquad (21)$$

where

      all terms are defined as above.

To compare groups of r-indexes, Formula (21) is applied to each grouping separately, and the sum of these results, Q_w, is subtracted from Q_t to obtain Q_b.

The results of a homogeneity analysis using the z-transforms of the r-indexes are presented in Table 6.4. The Q_t value of 178.66 is highly significant, based on a chi-square test with 5 degrees of freedom (the number of correlation minus one). While it seems that a range of r-indexes from .06 to .27 is not terribly large, Q_t tells us that, given the sizes of the samples on which these estimates are based, the variation in effect sizes is too great to be explained by sampling error alone. Something other than sampling of participants likely is contributing to the variance in r-indexes.

Suppose we know that the first three correlations in Table 6.4 are from samples of high school students and the last three are from elementary school students. A homogeneity analysis testing the effect of grade level on the magnitude of r-indexes reveals a Q_b of 93.31. This value is highly statistically significant, based on a chi-square test with one degree of freedom. For high school students the average weighted r-index is .253, whereas for elementary school students it is $r = .136$. Thus, the null hypothesis can be rejected and the grade level of the student is one potential explanation for the variation in r-indexes.

Comparing Observed and Expected Variance: Random-Effects Models

An important decision you will make when conducting a meta-analysis involves whether a fixed-effect or random-effects model should be used to calculate the variability in effect size estimates averaged across studies. As I discussed above, fixed-effect models calculate only error that reflects variation in studies' outcomes due to the sampling of participants. However, other features of studies also can be viewed as

influences on outcomes. For example, the studies in a synthesis of homework may vary by the length of the assignment and/or subject matter. Exercise interventions may vary in their intensity or modality. Choices may vary in number or domain. These variations will cause variation in effect sizes not due to sampling of participants. However, they are not error in the sense of being chance because even though they may at first be unexplained they may also be systematic in ways we are not aware of. For example, more-intense exercise interventions may improve cognitive functioning more than less-intense interventions.

For this reason, in many cases it may be most appropriate to treat studies as randomly sampled from a population of all studies. The variation that might be added to the estimate of error due to variations in study methods is ignored when a fixed-effect model is used. In a random-effects model (Raudenbush, 2009), study-level variance is assumed to be present as an additional source of random influence. The question you must answer, then, is whether you believe the effect sizes in your data set are noticeably affected by study-level influences.

Regrettably, there are no hard-and-fast rules for making this determination. Overton (1998) found that in the search for moderators, fixed-effect models may seriously underestimate error variance and random-effects models may seriously overestimate error variance when their assumptions are violated. Thus, neither can be chosen because it is statistically more justified. In practice, many meta-analysts opt for the fixed-effect assumption because it is analytically easier to manage. But some meta-analysts argue that fixed-effect models are used too often when random-effects models are more realistic, such as when interventions like homework or exercise programs can be expected to have different empirical realizations from one study to another in ways that will influence their effectiveness. Others counter this argument by claiming that a fixed-effect model can be applied if a thorough, appropriate search for moderators of effect sizes is part of the analytic strategy—that is, if the meta-analysts examine the systematic effects of study-level influences—and in this way make moot the issue of random effects at the study level. The fixed-effect model may also be favored if the number of effect sizes is small, making it difficult to achieve a good estimate of variation in the effect sizes at the study level.

What should your decision be based on? One approach is to decide based on the outcome of the test of homogeneity of effects using a fixed-effect model; if the hypothesis of homogeneous effects is rejected under the fixed-effect assumption, then you switch to a random-effects model. However, as Borenstein et al. (2009) argue, this strategy is discouraged; it is based on statistical outcomes, not on the conceptual characteristics of your studies. Many researchers interested in evaluating applied interventions (such as homework) often choose the random-effects model because they believe that random sampling of studies is more descriptive of their real-world circumstances and also will lead to a more conservative conclusion about the range of impacts the intervention might have (because the estimate of the variation around the average estimate is larger using the random-effects model). So, if you suspect a large influence of study-level sources of random error, then a random-effects model is most appropriate in order to take these sources of variance into account.

Other researchers studying basic social processes—processes that likely do not change greatly due to the contexts in which they are being studied (such as, perhaps, tests of reaction times)—tend to favor fixed-effect models. Hedges and Vevea (1998) stated that fixed-effect models are most appropriate when the goal of the research is "to make inferences only about the effect size parameters in the set of studies that are observed (or a set of studies identical to the observed studies except for uncertainty associated with the sampling of subjects)" (p. 3). In studies of basic processes, this type of inference might suffice, because you make the extra-statistical assumption that the relationship you are studying is largely insensitive to its context.

To summarize then, you might consider applying the following rules:

- Do not use the outcome of a fixed-effect homogeneity analysis to decide whether a random-effects analysis is called for. The decision should be based on the nature of the research question.
- In most instances where interventions are being evaluated or the research takes place in real-world contexts that vary from one another in important ways, random-effects models should be favored. However, if the number of studies being combined is small, consider using a fixed-effect model; the estimate of study-level variance will be too rough.
- In most instances where laboratory studies of basic processes are being combined, fixed-effect models should be appropriate. Here, the context of the study (study-level variations) should be less consequential to study findings.

Which model of effects you use and the set of assumptions your choice is based on needs to be incorporated into the interpretation and discussion of your findings. I will return to the issue of interpreting fixed-effect and random-effects models in Chapter 7.

Calculating random-effects estimates of the mean effect size, confidence intervals, homogeneity statistics, and moderator analyses is computationally complex. Because of this complexity, the formulas I have provided in this chapter are for fixed-effect models. I will not go into the calculation of the variance estimate in random-effects models (see Borenstein et al., 2009, if you are interested) but conceptually it

involves calculating the variation in effect sizes (using the effect size as the unit of analysis) and adding this to the variation due to sampling of participants (the fixed-effect). Thankfully, the statistical packages developed specifically for meta-analysis and the program macros associated with more general statistical packages allow you to conduct analyses using both fixed-effect and random-effects assumptions.

I^2: The Study-Level Measure of Effect

It may have occurred to you that meta-analysts point out the shortcomings of null hypothesis significance testing but then use it to test whether groups of studies have significantly different average effect sizes. This is only partially true. Certainly, a good meta-analysis presents the confidence interval around overall estimates of effect and for all subgroups when a moderator of effects is tested. A measure of effect also exists for quantifying the percentage of the variance in a set of studies that is due to the studies themselves and not sampling error. This statistic is called I^2 and is calculated as follows:

(22)
I2=ΣQ–dfQ×100%

$$I^2 = \sum \frac{Q - df}{Q} \times 100\% \qquad (22)$$

where

all quantities are defined as above.

I^2 tells you what portion of the total variance in the effect sizes is due to variance between the studies. The Cochrane Collaboration (Deeks, Higgins, & Altman, 2008) gives a rough guide to when the percentage of study variance may be important. In addition to the significance of the Q-statistic, it suggests that I^2 below 40% might not be important while I^2 above 75% suggests considerable heterogeneity.

Statistical Power in Meta-Analysis

The above discussion leads naturally into a consideration of the power of meta-analyses to detect effects. Meta-analyses have different statistical power for answering its multiple questions. First, meta-analysts ask the question, "What is the average effect size and the precision of this estimate, or, alternatively, with what certainty can we reject the null hypothesis?" The answer to this question will depend on the model, fixed or random, used to estimate the expected variation in effects. When a fixed-effect model is used, we can say with certainty that the power of the meta-analysis to detect an effect and the precision of the estimate will be greater in the meta-analysis than in any one or any subset of the primary studies going into the research synthesis. This is because the meta-analytic estimate will always be based on a larger sample of participants. If the assumption of the fixed-effect model is true (i.e., sampling error alone is making sample estimates different) the meta-analysis estimate will always be more precise.

However, this is not necessarily true when a random-effects model is employed. Here, the variability due to variations in study characteristics must be added to sampling error at the participant level. This source of variance is not present in any one study. So, if study-level variance is large it is possible when we calculate the precision of the average effect size that the precision of the individual studies (or one or some of them) can be greater than the precision of the meta-analytic effect size estimate. You can think of it this way: if the estimate of study-level variation adds nothing to the participant-level variance, then a fixed-effect and random-effects model will provide the same estimates of variability (equal to participant sampling alone) and meta-analytic estimates of effect will always be more precise than any single-study estimate. As the study-level variability moves away from a zero contribution, the precision of the meta-analytic estimate decreases and at some point, depending on the amount of study-level variability and the number and sample size of the primary studies, may become less precise than any single study estimate.

Next, meta-analysts ask whether there is sufficient power to detect a significant Q-statistic, or to reject the null hypothesis that sampling of participants alone is making the effect sizes different. Similar to power analysis with primary data, the power to detect a difference between an observed Q-statistic and an expected one is a function of the number of effect sizes you have, the sample sizes contributing to those effects sizes, the size of the expected study-level variation in effects (the I^2) as well as how well the effects conform to the necessary statistical assumptions (e.g., normal distribution).

Finally, meta-analysts might be interested in the power to detect differences between groups of studies: "Was the average effect in the group of Studies A different from the average effect in the group of Studies B?" This power analysis requires a variation on the analyses described in the last paragraph.

Conducting power analysis in meta-analysis often has a different purpose from that in primary research. After all, meta-analysts do not do power analysis to help decide how many studies to run. Perhaps, if an existing literature contains a very large number of studies, the meta-analyst might conduct an a priori

power analysis to determine how many studies to sample from it. Otherwise meta-analytic power analyses are most informative as guides to interpretation. The power of meta-analytic tests can be very low, especially for tests of moderators of study effects when a random-effects model is used and the number of studies is small. By conducting such an analysis, the interpretation of the results can include the possibility that accepting the null hypothesis might lead to a Type II error.

Meta-Regression: Considering Multiple Moderators Simultaneously or Sequentially

Homogeneity statistics can become unreliable and difficult to interpret when the meta-analysts wish to test more than one moderator of effect sizes at a time. Hedges and Olkin (1985) present one technique for testing multiple moderators. The model uses simultaneous or sequential tests for homogeneity. It removes the variance in effect sizes due to one moderator and then removes from the remaining variance any additional variance due to the next moderator. So, for example, if we were interested in whether the sex of the student influenced the effect of homework on achievement after controlling for the student's grade level, we would first test grade level as a moderator, then test the student's sex as a moderator *within* each grade-level category.

This procedure can be difficult to apply because characteristics of studies are often correlated with one another and the number of effect sizes in categories of interest rapidly becomes small. For example, suppose we wanted to test whether the effect of homework on achievement is influenced by both the grade level of students and the type of achievement measure. We might find that these two study characteristics are often confounded—more studies of high school students used standardized tests while more studies of elementary school students used class grades. Studies of homework with elementary school students using standardized tests may be rare. The problem would get even worse if yet a third variable were added to the mix.

Another statistical approach to testing multiple moderators of effect sizes simultaneously or sequentially is called *meta-regression*. As the name implies, this approach is the meta-analysis analog to multiple regression. In meta-regression, the effect sizes are the criterion variables and the study characteristics are the predictors (Hartung, Knapp, & Sinha, 2008). Meta-regression shares with multiple regression all the problems regarding the interpretation of the analysis' output when the predictors are intercorrelated (a likely characteristic of research synthesis data) and when the number of data points (effect sizes in meta-regression) are small.

Still, meta-regression is becoming more popular, especially now that meta-analysis programs are available to help you do them. One important consideration regarding when to use meta-regression involves the effect sizes that serve as the dependent variables. Remember that the regression analysis makes the assumption that the effect sizes are independent of one another. In Chapter 4 I discussed the units of analysis in research synthesis and some strategies for minimizing multiple outcomes that come from the same sample of participants. In meta-regression it is not unusual for the outcome rather than the sample to be used as the independent unit. This requires adjustments lest the estimates of error appear to be more precise than they actually are (see Hedges, Tipton, & Johnson, 2010).

Another approach to addressing the intercorrelation of study characteristics is to first generate homogeneity statistics for each characteristic separately, by repeating the calculation of *Q*-statistics. Then, when the results concerning moderators of effect sizes are interpreted, the meta-analysts also examine a matrix of intercorrelations among the moderators. This way, the meta-analyst can alert readers to study characteristics that may be confounded and draw inferences with these relations in mind. For example, we followed this procedure in the meta-analysis of the effects of choice on intrinsic motivation. We found that the effect of giving choices influenced children's intrinsic motivation more positively than adults' motivation. But we found also that the age of the participant was associated with the setting in which the choice experiment was conducted; studies with adults were more likely to be conducted in a traditional lab setting than were studies with children. This means that the different effect of choice on motivation for children and adults might not be due to the participants' age, but rather to where the study was conducted.

In sum, then, you need to make many practical decisions when conducting a meta-analysis, and the guidelines for making these are not as clear as we would like. While it is clear that a formal analysis of the variance in effect sizes is an essential part of any research synthesis containing large numbers of comparisons, it is also clear that you must take great care in the application of these statistics and in the description of how they were applied.

Using Computer Statistical Packages

Needless to say, calculating average weighted effect sizes and homogeneity statistics by hand is time-consuming and prone to error. Today, it is unheard of for meta-analysts to compute statistics for themselves, as I have done in the previous examples. Still, it is good for you to examine my examples

carefully and conduct the calculation yourself, so that you understand them. Then, the output of computer packages should be more interpretable by you and you should be more able to notice any errors that might have occurred.

Conveniently, the major computer statistics packages have macros developed that allow their use to conduct meta-analysis. For example, meta-analyses can be run using Excel spreadsheets (Neyeloff, Fuchs, & Moreira, 2012). David Wilson's very helpful website provides free macros for use with the SPSS, STATA, and SAS software packages. These packages are generally familiar to most social scientists. A book is available that shows how to use the statistical package R to conduct meta-analysis (Chen & Peace, 2013; see also http://cran.r-project.org/web/views/MetaAnalysis.html for a useful compendium of R programs). A free (though support comes at a cost) program dedicated to meta-analysis alone is called RevMan (http://tech.cochrane.org/revman/download). There are also stand-alone meta-analysis packages that can be purchased such as Comprehensive Meta-Analysis (2015) that will produce all the results for you, and give you many options for how to carry out your analyses.

Regardless of how the statistics are calculated, when evaluating a research synthesis, you should ask,

> Were (a) study design and implementation features along with (b) other critical features of studies, including historical, theoretical, and practical variables, tested as potential moderators of study outcomes?

Some Advanced Techniques in Meta-Analysis

Several more advanced approaches to meta-analysis have emerged in recent years. These typically require more advanced statistical knowledge and complex calculations than can be covered in an introductory textbook. Below, I will provide a brief conceptual introduction to some of the approaches receiving the most attention. Because the complex meta-analysis techniques require full treatment to be applied and are still used relatively infrequently I will not dwell on them here. If you are interested in more advanced techniques, you should first examine these in more detailed treatments, especially those given in Cooper et al. (2009) and the references provided below.

Hierarchical Linear Modeling

One new approach to meta-analysis involves using hierarchical linear modeling (Raudenbush & Bryk, 2001). This approach treats study outcomes as nested data; for example, students' achievement scores can be viewed as influenced by (nested within) classroom-level variables that are themselves influenced by school characteristics and at a higher level still by the community the school is in. In the case of meta-analysis, a study outcome (an individual effect size) can be viewed as nested within a sample of participants who in turn are nested within a study (and even within a laboratory that has conducted multiple studies). Again, the computations for the analyses are complex but this approach is conceptually appealing and meta-analyses using the hierarchical linear modeling approach are used increasingly frequently.

Model-Based Meta-Analysis

The statistical procedures for meta-analysis described so far apply to synthesizing two-variable relationships from experimental and descriptive research. Meta-analysis methodologists are working to extend statistical synthesis procedures to more-complex ways to express the relations between variables. Previously, I discussed the difficulties in synthesizing the effect sizes associated with a variable that was included in a multiple regression. But what if the question of interest involves integrating the output of *entire* regression equations? For example, suppose we were interested in how five personality variables (perhaps those in the five-factor model) jointly predicted attitudes toward rape? Here, we would want to develop from a meta-analysis a regression equation, or perhaps a structural equation model, based on the results of a set of studies. To do so, we would need to integrate results of studies concerning not one correlation between the variables but rather an entire matrix of correlations relating all the variables in the model of interest to us. It is this correlation matrix that forms the basis of the multiple regression model.

The techniques used to do this are still being explored, as are the problems meta-analysts face in using them. For example, can we simply conduct separate meta-analyses for each correlation coefficient in the matrix and then use the resulting matrix to generate the regression equation? The answer is "probably not." The individual correlations would then be based on different samples of participants and a regression analysis using them can produce nonsensical results, such as prediction equations that explain more than 100% of the variance in the criterion variable. Still, there are circumstances under which these applications of meta-analysis to complex questions can produce highly informative results. Becker (2009) presents an in-depth examination of the promise and problems involved in model-driven meta-analysis.[6]

Bayesian Meta-Analysis

Another approach to meta-analysis involves applying Bayesian statistics rather than the frequentist approach used in the statistics described in this book. In a Bayesian approach (Sutton & Abrams, 2013; Sutton et al., 2000), the researcher must first establish a prior estimation of the parameters of the effect size. These can include both the magnitude and the distribution of effect sizes. The Bayesian priors can be based on past research, and not necessarily on research that used identical conceptual variables or empirical realizations. For example, the prior estimation of the effect of an exercise intervention might be based on other interventions to improve cognitive functioning, such as puzzle solving. Or, the estimation might be based on samples drawn from other populations (e.g., using adult samples to estimate the effect of choice on children's motivation) or even on subjective beliefs and personal experience (e.g., teachers' thoughts on the degree to which homework affects achievement). The meta-analysis then tells the synthesists how these prior beliefs should change in light of the new empirical evidence. The need for prior estimations in Bayesian analyses is seen as both a strength and a weakness of the approach. The computations for Bayesian analyses are also very complex and less intuitively accessible than the traditional meta-analysis methods but can yield trustworthy and interpretable results (Jonas et al., 2013).

Meta-Analysis Using Individual Participant Data

The most desirable technique for combining results of independent studies is to have available and to integrate the raw data from each relevant comparison or estimate of a relationship (Cooper & Patall, 2009). Then, the individual participant data (IPD) can be placed into a new primary data analysis that employs the comparison that generated the data as a blocking variable. When IPD are available, the meta-analysis can perform subgroup analyses that were not conducted by the initial data collectors in order to

- Check data in the original studies,
- Ensure that the original analyses were conducted properly,
- Add new information to the data sets,
- Test with greater power variables that moderate effect sizes, and
- Test for both between-study and within-study moderators.

Obviously, instances in which the integration of IPD can be achieved are rare. IPD are seldom included in research reports, and attempts to obtain raw data from researchers often end in failure. However, the incentives and requirements for sharing data are increasing, as conditions both for receiving research support and for publishing findings. If IPD are retrievable, the meta-analyst still must overcome the use of different metrics in different studies, an important limit to the ability to statistically combine the results. Also, meta-analyses using IPD can be expensive because of the recoding involved in getting the data sets into similar form and content. So it is unlikely that meta-analyses using IPD will be replacing the meta-analysis techniques described previously any time soon. Still, meta-analysis of IPD is an attractive alternative, one that has received considerable attention in the medical literature, and likely will become more attractive as the availability of raw data sets improves. Also, methods are appearing that allow synthesists to use both IPD from some studies and aggregate data from others (Pigott, 2012).

Cumulating Results Across Meta-Analyses

The terms *cumulative* or *prospective meta-analyses* are used to refer to meta-analyses that are updated as new evidence on a topic becomes available. The methods for conducting the new analyses can be the same as those used originally, or can be changed, perhaps to reflect advances in meta-analytic methods or to conduct new analyses that time and experience suggest are warranted; for example, looking at a new moderators variable that recent theorizing suggests might influence results. Many cumulative meta-analyses include the year of the study as a moderating variable to determine whether the evidence suggests the impact of the treatment or intervention is changing over time. Cumulative meta-analyses are much more frequently encountered in the medical than social sciences. In fact, the Cochrane Collaboration (2015) requires that synthesists who submit to its database commit to updating the reports as new information appears.

Overviews of reviews. Overviews of reviews, sometimes also called *reviews of reviews, umbrella reviews,* or *meta-reviews,* compile evidence from multiple research syntheses. Cooper and Koenka (2012) catalogued several reasons why an overview of reviews might be undertaken. These included (a) to summarize evidence from more than one research synthesis focused on the same or overlapping research problems or hypotheses, (b) to compare findings and resolve discrepancies in the conclusions drawn in more than one research synthesis, and (c) to catalog the mediators and moderators tested in research syntheses on the same research problem. Like all research syntheses, there are sound methods for conducting an overview of reviews that are unique to them. For example, overviewers must evaluate the quality of the constituent research syntheses.

Overviews have their limitations as well. For example, the studies included in an overview of reviews can

be quite old, considering not only that the studies must be conducted, but also that the review of studies then must be conducted and this is the evidence in the overview. Still, the same forces that are giving rise to the need for research syntheses, the expanding research literature, will also provide impetus for a growing appearance of overviews of reviews.

Second-order meta-analysis. One type of overview is called a *second-order meta-analyses.* It involves using the outcomes of meta-analyses as the data in yet another meta-analysis (Schmidt & Hunter, 2015). In second-order meta-analyses the average effects found in meta-analyses conducted in the same problem area are themselves combined. Obviously, second-order meta-analysis is used when neither the IPD nor even the study-level results from the constituent meta-analyses can be retrieved.

One problem faced by second-order meta-analysts (as well as any overviewer, only more formally) is how to handle meta-analyses with overlapping evidence—that is, the constituent meta-analyses were conducted on the same set or a substantial subset of the same primary studies. The approaches that have been taken to this nonindependence of evidence include simply ignoring the lack of independence, removing meta-analyses that are highly redundant with others, and conducting sensitivity analyses—that is, doing the second-order meta-analysis with different sets of constituent meta-analyses. Also, the ability of second-order meta-analyses to look at influences on the average effect sizes can be limited because the moderating and mediating variables examined must exist at the level of the meta-analyses that go into the second-order meta-analysis, not the individual studies. Still, second-order meta-analyses can be done (e.g., Tamim, Bernard, Borokhovski, Abrami, & Schmid, 2011). When the more desirable alternatives are not feasible, you should give consideration to doing a second-order meta-analysis.

Finding	n_{i1}	n_{i2}	d_i
1	193	173	−.08
2	54	42	.35
3	120	160	.47
4	62	60	.00
5	70	84	.33
6	60	60	.41
7	72	72	−.28

Exercises

1. For the findings in the table below, what is the average weighted *d*-index?
2. Are the effect sizes of the seven studies homogeneous? Calculate your answer both by hand and by using a computer statistical package.

Notes

1. And they permit you to choose whether you want the test to estimate sampling error based on participant variation alone or both participant and study variation. I will return to this choice later, when I discuss fixed-effect and random-effects models.

2. Throughout this chapter and forward, I will use the terms *findings, studies,* and *comparisons* interchangeably to refer to the discrete, independent hypothesis tests or estimates of relationships that compose the input for a meta-analysis. I do this for exposition purposes, though sometimes these terms can have different meanings; for example, a study could contain more than one comparison between the same conditions.

3. Borenstein et al. (2009) present formulas for how to combine two nonindependent effect sizes. These authors also provide formulas for how to combine effect sizes for different outcome measures taken on the same sample and for the same outcome measure taken on the same sample but at different times. The *d*-index for any two-group comparison can also be calculated if you have the means, samples sizes, and overall multi-degree-of-freedom *F*-test using the *Practical Meta-Analysis Effect Size Calculator* (Wilson, 2015).

4. Remember that measurement reliability can also be used as a moderator of effects, so without adjusting measures you could group them by reliability and ask, "Is the size of the impact of homework related to the reliability of the achievement measure?"

5. Half-standardizing is an alternative way to create similar slopes when only outcomes are dissimilar (see

Greenwald, Hedges, & Laine, 1996).

6. The use of structural equation modeling in meta-analysis is an emerging area that incorporates many of the approaches I have described, not only to exploring multiple relationships in the same analysis, but also different model assumptions and even missing data techniques (Cheung, 2015). Synthesists will need a comfortable knowledge of these methods of structural equation modeling before they can use them successfully, though they can use the available software packages to carry them out.

7 Step 6 Interpreting the Evidence

What conclusions can be drawn about the cumulative state of the research evidence?

Primary Function in Research Synthesis

To summarize the cumulative research evidence with regard to its conclusiveness, generalizability, and limitations

Procedural Variation That Might Produce Differences in Conclusions

Variation in (a) criteria for labeling results as *important* and (b) attention to details of studies might lead to differences in interpretation of findings.

Questions to Ask When Interpreting the Cumulative Evidence in the Research Synthesis

1. Were analyses carried out that tested whether results were sensitive to statistical assumptions and, if so, were these analyses used to help interpret the evidence?
2. Did the research synthesists (a) discuss the extent of missing data in the evidence base and (b) examine its potential impact on the findings of the synthesis?
3. Did the research synthesists discuss the generality and limitations of the synthesis' findings?
4. Did the synthesists make the appropriate distinction between study-generated and synthesis-generated evidence when interpreting the synthesis' results?
5. If a meta-analysis was performed, did the synthesists (a) contrast the magnitude of effects with other related effect sizes and/or (b) present a practical interpretation of the significance of the effects?

This chapter describes

- How to account for missing data
- Statistical sensitivity analysis
- Generalization and specification of findings
- Study-generated and synthesis-generated evidence
- Substantive interpretation of effect sizes

Properly interpreting the results of your research synthesis will require you to carefully (a) state the claims you want to make based on the evidence, (b) specify what results warrant each claim, and (c) make explicit any appropriate qualifications to claims. In this chapter I discuss five important issues related to the interpretation of results in research synthesis:

- The impact of missing data on conclusions
- The sensitivity of your conclusions to changes in assumptions about the statistical characteristics of your data
- Your ability to generalize your conclusions to people and circumstances not included in the constituent studies
- Whether conclusions are based on study-generated or synthesis-generated evidence
- The substantive interpretation of effect sizes

Missing Data

Even after the careful planning, searching, and coding of research reports, missing data can influence the conclusions drawn from research syntheses. When data are systematically missing, not only is the amount of evidence you gathered reduced but the representativeness of your results may be compromised. In Chapter 4 I discussed the issue of missing data and suggested a few ways to address the problem when you code studies. But even these techniques for estimating missing data do not solve the problem entirely. You do not have an equal chance of retrieving every study, so there might be some studies completely missing from your data set (Rosenthal, 1979, called this the *file drawer problem*). Also, in some instances studies you do have may have collected data on some outcomes and tested them but then failed to give any indication of this in the report. These completely missing results cannot be estimated by the procedures described in Chapter 4. Compounding the problem is the fact that in many instances a disproportionate number of completely missing results will be associated with studies that use small sample sizes and have statistically nonsignificant inference tests (Borenstein et al., 2009). As such, they would tend to be the smaller effect sizes in the distribution of estimates. This means the effect sizes you do find may overestimate the true population value.

Yet, there are some things you can do. A number of graphical and statistical techniques can be used to assess the possible existence of completely missing data and its implications for the interpretation of your results. It is good practice to apply at least one of these to your data. These techniques include the Rank

Correlation Test (Begg & Mazumdar, 1994), the Linear Regression Method (Egger, Davey Smith & Minder, 1997), the Funnel Plot Regression Method (Macaskill, Walter, & Irwig, 2001), and the Trim-and-Fill Method (Duval & Tweedie, 2000a, 2000b). All of these methods have been shown to have strengths and weaknesses depending on the characteristics of the literature they are applied to (Kromrey & Rendina-Gobioff, 2006). I refer you to *Publication Bias in Meta-Analysis* (Rothstein et al., 2005) for a thorough but accessible treatment of ways to prevent, assess, and adjust for missing data due to publication bias.

The method we have found especially useful is the Trim-and-Fill Method (Duval & Tweedie, 2000a, 2000b). Though not perfect, the Trim-and-Fill Method makes reasonable assumptions about the missing data, is intuitively appealing, and is easy to understand. The Trim-and-Fill Method tests whether the distribution of effect sizes used in the analyses is consistent with the distribution that would be predicted if the estimates were symmetrically distributed around their mean. If the distribution of observed effect sizes is found to be asymmetric in some way—indicating possible missing effect sizes caused by a search limitation or by data censoring on the part of primary researchers—the Trim-and-Fill Method provides ways to estimate the values for missing studies that would improve the symmetry of the distribution. Then, after imputing these values, it permits you to estimate the impact of data censoring on the observed mean and variance of effect sizes. Duval (2005) gives a good introduction to how to carry out the analysis. It can also be carried out using the Comprehensive Meta-Analysis (2015) software program.

In our homework meta-analysis, we conducted trim-and-fill analyses using the five effect sizes we found for studies that manipulated whether or not students received homework. Figure 7.1 shows the results of the analyses using a funnel plot (to be discussed shortly). You can see that the trim-and-fill analysis suggested that, if the distribution of effect sizes were truly symmetrical, two studies might be missing from the left side of the funnel plot. Methods are provided to calculate what these values might be (under both fixed-effect and random-effects models), and then recalculate average effect sizes and confidence intervals. In this case, recalculating the average homework effect produced a smaller average effect size ($d = .48$) but one whose confidence interval still did not include zero. Thus, this technique for estimating completely missing effect sizes leaves us more confident that our results would not change substantively had the missing data been found.

Figure 7.1 Example of a Trim-and-Fill Funnel Plot
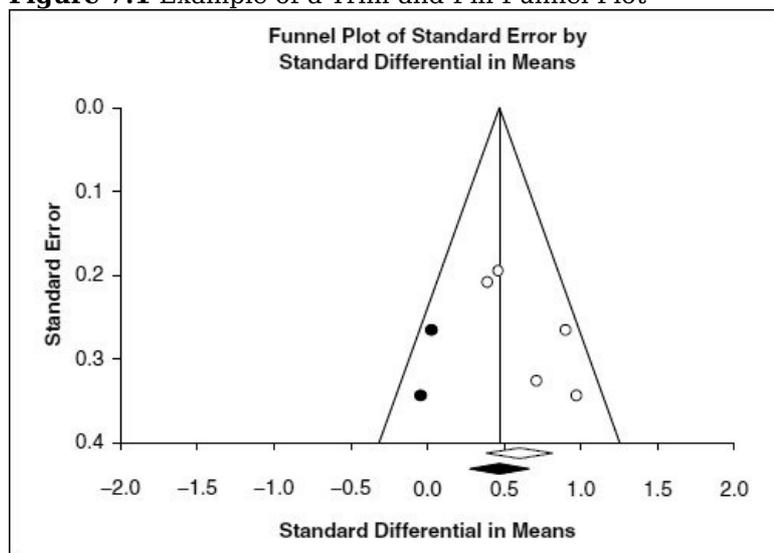
Regardless of the techniques you use, *you are obligated to discuss whether data might have been missing either because of missing reports or analyses not reported in the individual study reports you do have, how you handled it, and why you chose to treat the missing data the way you did.* Thus, an important question to ask about missing data when evaluating research syntheses is,

Did the research synthesists (a) discuss the extent of missing data in the evidence base and (b) examine its potential impact on the synthesis' findings?

Statistical Sensitivity Analyses

The next important step in the interpretation of data from meta-analyses also is undertaken as part of data analysis: the performance of statistical sensitivity analyses. Statistical sensitivity analyses are used to determine whether and how the conclusions of your analyses might differ if they were conducted using different statistical procedures or different assumptions about the data. There are numerous decisions you make about how to analyze your data that are candidates for sensitivity analysis. For example, the

calculation of weighted and unweighted effect sizes can be considered a form of sensitivity analysis. When you present these measures of central tendency and their confidence intervals calculated differently, in essence you are answering the question, "Do I come to a different conclusion about the average effect size when I ignore the precision of the individual effect size estimates than when I take their precision into account?"

You might also consider (as we did in our meta-analysis of the association between the time students spend on homework and achievement) conducting your analyses twice, once using a fixed-effect model and once using a random-effects model. Rather than choosing a single model for error, we chose to apply both models to our data. By employing this sensitivity analysis, differences in results based on which set of assumptions about error was used could then be part of our interpretation of results. If an analysis revealed that a finding was significant under fixed-effect assumptions but not under random-effects assumptions, this result suggests that the significant finding relates only to what past studies have found but not necessarily to the likely results of a broader universe of similar studies. For example, we found that a small negative association between time on homework and achievement for elementary school students was statistically significant using a fixed-effect model, but the association was not significant using a random-effects model. A similar set of results occurred for this association when parents reported time on homework. Also, we found that of four moderator variables that produced significant effects using a fixed-effect model, two were also significant using a random-effects model (the association was stronger for high school than elementary school students and when students reported time on homework than when parents were reporters), while two were not significant using the random-effects model (the type of outcome measure and the subject matter). Sensitivity analyses can also be conducted using different assumptions and techniques to estimate missing data.

Each time you do a sensitivity analysis, you are seeking to determine whether a particular finding is robust across analyses conducted with different sets of statistical assumptions. In the interpretation of evidence, a finding that a conclusion does not change using different statistical tests or assumptions means greater confidence can be placed in the conclusion. *If results are different under different assumptions, this suggests a caution, or different interpretation, is needed when you share your results with the users of your synthesis.* So, another question to ask when you evaluate the interpretation of results in research synthesis is,

> Were analyses carried out that tested whether results were sensitive to statistical assumptions and, if so, were these analyses used to help interpret the evidence?

Specification and Generalization

Research synthesis, like any research, involves specifying the targeted participants, program or intervention types, occasions, settings, and outcomes. *When you interpret your results, you must assess whether and how well each of the target elements is represented in the evidence base.* For example, if you were interested in making claims about the effects of choice on motivation, you would need to note whether important age groups were included or missing from the samples of participants.

The trustworthiness of any claim about the generality of a research finding will be compromised if the elements in the accessed samples are not representative of the target elements, be they targeted people, programs, settings, times, or outcomes. Thus, you may find you need to respecify your covered elements once your data analysis is complete. For example, if only college students were used in studies of beliefs about rape, then either any claims about the relationship must be restricted to this particular type of participant or the rationale for extrapolation beyond the included type of participants must be provided in your interpretation of the data.

Your permissible generalizations in a research synthesis are constrained by the types of elements sampled by primary researchers. Still, generalization in research synthesis injects a note of optimism into the discussion. There is good reason to believe research syntheses will pertain more directly to the target participants, programs, settings, times, and outcomes—or to more subgroups within these targets—than will the separate primary studies. The cumulative literature can contain studies conducted on participants and programs with different characteristics at different times and in different settings using different outcome measures. For certain topics containing numerous replications, participants and circumstances accessible to the synthesists may more closely approximate the targeted elements than does any individual primary study. For example, if some studies of aerobic exercise programs use treadmills in a gym and others use riding a bike outdoors, then you can ask whether program effects were similar or different across the two types of exercises and their context, a question unanswerable by the individual studies.

Integrating Interaction Results Across Studies

A problem in interpreting the results of research syntheses that is related to the issue of generalization

concerns the interpretation of interactions. Often, the integration of interaction results in research synthesis is not as simple as averaging the effect sizes from each study. Figure 7.2 illustrates the problem by presenting the results of two hypothetical studies comparing the effects of homework versus in-school study on students' ability to retrieve the covered material from memory. These two studies tested whether the effect of the two instructional strategies was mediated by the delay between when the instruction occurred and when the measure of retention was taken. In Study 1, retention was compared across students at both one and eight weeks after the intervention. In Study 2, the intervals of delay were one and four weeks. Study 1 might have produced no significant main effect but a significant interaction involving the measurement interval while Study 2 might have reported only a significant main effect.

If you uncovered these two studies and did not examine the precise form of the data, you might be tempted to conclude that they produced inconsistent results. However, an examination of Figures 7.2(A) and 7.2(B) illustrates why this might not be an appropriate interpretation. The results of Study 2 would have closely approximated those of Study 1 had the measurement delay in Study 2 been the same as that in Study 1. Note that the slopes for the lines for the two groups in Study 1 and Study 2 are nearly identical.

Figure 7.2 Results of Two Hypothetical Studies Comparing the Effects of Homework and In-School Study on Retention
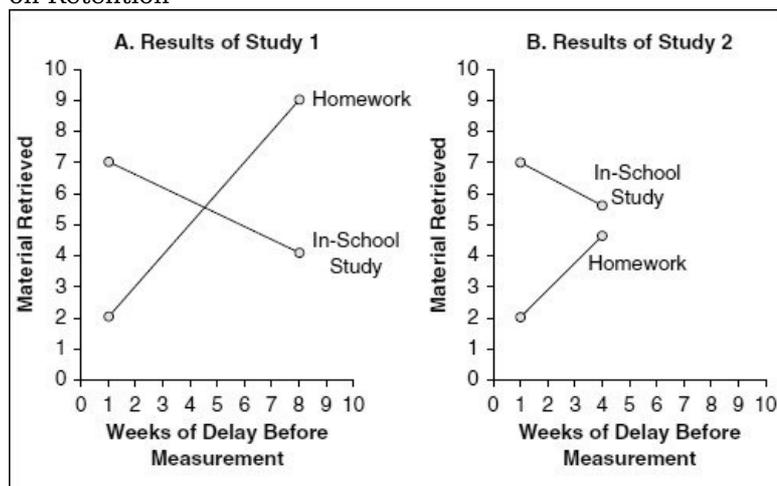
This example demonstrates that synthesists should not assume that different strengths of interaction uncovered by different studies necessarily imply inconsistent results. *Synthesists need to examine the differing ranges of values of the variables employed in different studies, be they measured or manipulated.* If possible, you should chart results taking the different levels into account. In this manner, one of the benefits of research synthesis is realized. While one study's authors might conclude that measurement delay has no effect on the ability to retrieve information from memory when homework and in-school study are compared and a second study's authors might conclude such an interaction exists, a research synthesist could discover that the two results are in fact perfectly consistent.

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 an across-study analysis similar to my example without this information. If the researchers in Study 1 and Study 2 neglected to specify their range of measurement delays—perhaps they simply said they compared "short" delays with "longer" delays—the commensurability of the results would have been impossible to demonstrate.

I should mention that variations in ranges of values for variables also can produce discrepancies in results involving two-variable or main-effect relationships. For example, had Study 1 included a retention measurement at one week, Study 2 at four weeks, and a third study at eight weeks, the three studies would have produced three different results. In such a case, we would hope that the research synthesists would examine measurement delay as a moderator variable and reveal the influence it had on study results. I mention here the impact of ranges in values 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.

In general, then, the next question to ask when evaluating the interpretation of a research synthesis is,

Did the research synthesists discuss the generality and limitations of the synthesis findings?

Study-Generated and Synthesis-Generated Evidence

In I made the important point that research syntheses can contain two different sources of evidence about the research problem or hypothesis. Study-generated evidence is present when a single study contains results that directly test the relation being considered. Synthesis-generated evidence does not come from individual studies but rather from the variations in procedures across studies. I noted that only study-generated evidence based on experimental research allows you to make statements concerning causality. I return to the point here to emphasize that *making the distinction between what evidence in your synthesis supports causal relationships and what evidence does not is an important aspect of proper interpretation of your results.* Therefore, the next important question to ask when evaluating the interpretation of evidence in a research synthesis is,

> Did the synthesists make the appropriate distinction between study-generated and synthesis-generated evidence when interpreting the synthesis' results?

The Substantive Interpretation of Effect Size

In quantitative syntheses, one function of a discussion section is the interpretation of the size of reported effects, be they the magnitudes of group differences, correlations, or odds ratios. Once you have calculated effect sizes, how do you know if they are large or small, meaningful or trivial? Since statistical significance cannot be used as a benchmark—small effects can be statistically significant and large effects nonsignificant—a set of rules must be established for determining the explanatory or practical importance of a given effect magnitude.[1]

The Size of the Relationship

To help interpret effect sizes, social scientists have applied labels that describe the size of the relationship between two variables. Jacob Cohen provided the first guides to interpreting effect sizes in this way in the 1977 edition of his book on statistical power analysis (reprinted as Cohen, 1988). He proposed a set of values to serve as definitions of small, medium, and large effects. Cohen recognized that judgments of largeness and smallness are relative. In order to make them, a comparison between two items is required, and the choice of a contrasting element for an observed effect could be governed by many different rules. Interestingly, however, Cohen did not intend his labels to serve as guides for substantive interpretation by social scientists. Rather, he intended his rules to assist with power analyses in planning future studies, a very different objective. However, since they have been used so often for substantive interpretation, we should look at their characteristics.

In defining the adjectives for magnitude, Cohen compared different average effect sizes he had encountered in the behavioral sciences. He defined a small effect as $d = .2$ or $r = .1$ (equivalent values), which his experience suggested were typical of those found in personality, social, and clinical psychology research. A large effect of $d = .8$ or $r = .5$ was more likely to be found in sociology, economics, and experimental or physiological psychology. Medium effects, $d = .5$ and $r = .3$, fell between these extremes. According to Cohen, then, a social psychologist studying the impact of choice on intrinsic motivation might interpret an effect size of $d = .2$ to be small when compared with all behavioral science effects but about average when compared with all other effects in social psychology.

Cohen (1988) was careful to stress that his conventions were to be used "only when no better basis for estimating the ES [effect size] is available" (p. 25). Today, with so many meta-analytic estimates of effects available, other more closely related contrasting effects often can be found. So, when you interpret the magnitude of effects, it is most informative to use contrasting elements that are more closely related to your topic than to use Cohen's benchmarks.

For example, our meta-analysis of the effect of choice on motivation revealed an effect size of $d = .30$ using a fixed-effect model and .36 using a random-effects model. This indicates that on average participants scored about a third of a standard deviation higher on a motivation measure when given a choice of tasks than when no choice was offered. Using Cohen's guide, we would label this effect *small.* However, other contrasting elements might be available to us. These might come from other meta-analyses that looked at entirely different ways to motivate people to do tasks, such as providing rewards. Thus, one way to interpret the effect of choice would be to ask whether it revealed smaller or larger effects than other contextual manipulations meant to increase motivation.

Alternatively, other meta-analyses might share the same treatment but vary in outcome measure. For example, some studies of choice may have examined participants' self-esteem after being given a choice rather than their subsequent motivation. Then, a good interpretation would consider whether choice-promoting manipulations have a smaller or larger effect on motivation than on self-esteem. Of course, these types of interpretations could occur among results within the same research synthesis as well; for example, the measures of exercise effects on different cognitive functions. Is the effect larger on memory than executive functioning?

When you cannot find meta-analyses this closely aligned with your topic, you might be able to find compendia of meta-analyses on more distant but related topics that are still closer to your topic than Cohen's benchmarks. For example, Lipsey and Wilson (1993) compiled the results of 302 meta-analyses across the fields of education, mental health, and organizational psychology. In education, the authors cataloged 181 average effect sizes. The middle one-fifth of these ranged from $d = .35$ to $d = .49$. The top one-fifth of effect size estimates were above $d = .70$, and the bottom one-fifth were below $d = .20$. So, in our meta-analysis of experimental studies examining the effects of homework on student's grades on unit tests, we reported an average d-index of about .60. Comparing this to Lipsey and Wilson's estimates, we might label the homework effect as *above average* for educational interventions. There are also now compendia of meta-analysis that relate to broad topics but are narrower than the fields covered by Lipsey and Wilson (e.g., Hattie, 2008, synthesizes 800 meta-analyses on visual learning and achievement).

Effect sizes also need to be interpreted in relation to the methodology used in the primary research. Cohen (1988) acknowledged this when he pointed out that field studies should expect smaller effects than laboratory studies (p. 25). In addition, studies that provide an intervention in greater frequency and intensity (e.g., more-frequent or more-intense aerobic exercise) would have greater likelihood of showing a large effect than more conscribed versions of another intervention, even if the interventions were equally effective when tested under comparable conditions. Or studies that remove more error from the measurement of the outcome (perhaps by using a more reliable measure) should produce larger effects than studies that allow more extraneous influences on outcomes. Therefore, research design differences must be considered when drawing a conclusion about the relative size of effects; more-sensitive research designs and measures that therefore have less random error can be expected to reveal larger effect sizes, all else being equal.

In sum, then, the choice of contrasting elements is critical in interpreting the magnitude of an effect. In fact, almost any effect can be deemed large or small depending on the chosen contrast. In addition, contrasting elements often vary on dimensions other than the impact of the manipulations or predictor variables they are sizing up. *It is essential that (a) contrasting elements be chosen that hold constant those aspects of research design that are known to influence effect size estimates but are not inherent to the intervention itself or (b) differences in research design be considered when the contrast between effects is interpreted.*

Using Adjectives to Convey the Practical Significance of Effects

Researchers are well aware that the term *significant effect* has a different meaning for statisticians than it does for the broader public. To those who calculate social science statistics, a significant effect typically means it is one that permits the rejection of the null hypothesis with some minimal potential for error, usually 1 chance in 20. To the general public, however, the meaning of the word *significant* is different. The *Merriam-Webster* (2015a) online dictionary defines significance first as "having meaning" and second as "having or likely to have influence or effect: important." Most researchers recognize this distinction between colloquial and scientific usage, and when they use the term *significant effect* in public conversations, they typically use it to mean important, notable, or consequential, rather than in the statistical sense.

The question becomes, then, "When can you use terms such as *significant, important, notable,* or *consequential* (or their antonyms) to describe effect sizes?" At least two organizations set this bar at $d = .25$ (Promising Practices Network [PPN], 2014; What Works Clearinghouse, 2014). In contrast, after comparing the results of psychoeducational meta-analyses with results from the field of medicine, Lipsey and Wilson (1993) conclude, "We cannot arbitrarily dismiss statistically modest values (even 0.10 or 0.20 SDs) as obviously trivial" (p. 1199).

You also can assess how much *any* relation might be valued by consumers of your synthesis. This assessment involves the difficult task of making practical judgments about significance. So, for example, an effect size of aerobic exercise on improved memory in adults of $d = .128$ may be small when compared to Cohen's benchmarks and other contrasting elements. Still, we might argue that this improvement translates into an equivalent measure suggesting that a practically important number of adults will remain cognitively robust until later in life (see Rosenthal, 1990, for a similar argument). It might then be argued that the cost of the intervention was minimal relative to its change in life satisfaction, or even the cost of health care saved by participants. Levin (2002) has laid out some ground rules for conducting this type of cost-effectiveness analysis for educational programs.

In a similar vein relating to the impact of afterschool programs, Kane (2004) made the case that the interpretation of an effect also needs to be influenced by what a reasonable expectation might be for the impact of the intervention or manipulation. His assessment of valuations of afterschool programs led Kane to use a threshold even lower than that suggested by Lipsey and Wilson. He pointed out that the national samples used to norm the Stanford 9 achievement test suggested that test scores of fifth graders in the spring were one-third a standard deviation higher in reading and one-half a standard deviation higher in math than they were in the spring of fourth grade. This effect is the result of "everything that happens to a student between the end of fourth grade and the end of fifth grade" (p. 3). Given this effect,

Kane argued it would be unreasonable to expect an effect size of even $d = .20$ (Cohen's definition of a small effect) from the added instruction made available through afterschool programs. He went on to argue that a more reasonable expectation for interventions such as afterschool programs could be set by calculating (a) the fraction of time during a school year that students spend in such programs or (b) the gain in earnings in later life needed to offset the cost of the program. Kane suggested that in both cases the more reasonable expectation for a consequential effect of afterschool programs would be between $d = .05$ and .07.

Thus, it appears that criteria for labeling an effect size as *practically significant, important,* or *consequential* also vary, and a case can be made that developing such criteria requires a contextualization of the effect little different from that involved in applying labels on the size of an effect relative to other effects.

Using Adjectives to Convey Proven and Promising Findings

Two other descriptors of research results that relate to the evaluations of social programs have received attention among some social scientists. These are the terms *proven* and *promising.* For example, the PPN (2014) requires that for a program or practice to be labeled *proven,* the associated evidence must meet the following criteria:

- The program must directly affect one of the important outcome measures.
- At least one outcome is changed by 20%, $d = 0.25$, or more.
- At least one outcome with a substantial effect size is statistically significant at the 5% level.
- The study design used a convincing comparison group to identify program impacts, such as random assignment or some quasi-experimental designs.
- The sample size exceeds 30 in both the treatment and comparison groups.
- The report is publicly available.

These criteria appear to refer to as few as one study.

To claim that anything is proven through research is always problematic, as most philosophers of science accept the notion that a rejection of the null hypothesis at any statistical level, no matter how improbable, is not an affirmation of any specific alternative hypothesis (Popper, 2002). Of course, there are different levels of uncertainty about alternative hypotheses depending on the number and nature of other explanations that fit the data.

The PPN goes on to define the term *promising* as follows:

- The program may affect an intermediary outcome associated with one of the indicators of interest.
- There is an associated change in outcome of more than 1%.
- Outcome change is significant at the 10% level.
- The study has a comparison group, but it may have some weaknesses; for example, the groups lack comparability on preexisting variables or the analysis does not employ appropriate statistical controls.
- The sample size exceeds 10 in both the program and comparison groups.
- The report is publicly available.

The Merriam-Webster online dictionary (2015b) defines the word *promising* as "likely to succeed or be good: full of promise." Thus, there is some correspondence between the PPN definition and common parlance, if we assume that having an impact on mediating variables and using less-than-optimal research designs can provide hope that the practice will produce positive results in future studies that test the intervention more directly and rigorously. However, the PPN definitions of both proven and promising also include reference to the magnitude of the program's effect. So, it appears that a program would be labeled *promising,* even if it measured the outcomes of most interest and used more-rigorous designs but revealed a smaller effect size than PPN requires for a program to be considered "proven" (i.e., at least one outcome is changed by 20%, $d = 0.25$, or more). This confounding of the trustworthiness of the evidence and the magnitude of the effect may be a divergence from the everyday understanding of what both proven and promising mean.

Should Researchers Supply Labels at All?

Cohen (1988) cautioned that magnitude labels applied to effect sizes would involve an arbitrary choice of contrasting elements. To serve your audience best, then, *it is best practice to present multiple contrasting elements, perhaps picking some contrasting elements that make the effect of interest look relatively small and other elements that make it look relatively large.* The search for definitions of "significance" has revealed that effects of small magnitude, relatively speaking, may not be inconsequential. And, effects that are even smaller than small may be all we can reasonably expect with some interventions. Trying to provide such benchmarks are valiant and instructive efforts that, at the least, caution you *not to apply labels for effects without providing your audience with much additional context, even multiple contexts.*

Furthermore, I suggest you cast a critical eye on efforts to define terms such as *proven* and *promising* by associating them with different clusters of research characteristics and results, whether these clusters are based on the number of studies and participants, the research designs used, the statistical significance of results, the size of effects, and so on, in various combinations. Efforts to define these labels seem destined to always come up short, lead to a lack of consensus concerning what cluster of evidence justifies what label, and, perhaps of most concern, provide esoteric definitions of commonly used words that simply do not map onto the ways these terms are understood in everyday language.

Metrics That Are Meaningful to General Audiences

One way you can get around providing qualitative labels is to try to express the quantitative results of your meta-analysis in metrics that have meaning for your audience. If this can be accomplished, the audience should be able to apply their own qualitative labels to the quantitative results and perhaps debate the appropriateness of different labels they have applied. Put simply, if you can convey a clear understanding of what constitutes an ounce, then audiences should be able to determine, or debate, whether an eight-ounce glass containing four ounces of liquid is half empty or half full. Next I describe a few ways to translate effect size metrics into specific contexts that have enough intuitive meaning that general audiences can apply, and debate, the appropriateness of different labels.

Raw Scores and Familiar Transformed Scores

Many metrics are familiar enough to be understood implicitly by audiences not steeped in statistical training. These include some metrics in their raw form, such as a person's blood pressure. So, if you tell your audience that an activity-promotion intervention for older adults led to a 10-point reduction in systolic blood pressure and a 5-point reduction in diastolic blood pressure, your audience might be able to interpret this finding as "important" or "trivial" without you supplying a label, though you might still want to present the effects of some other interventions meant to accomplish the same type of result and, perhaps, a relative cost-benefit analysis. Other scores are familiar transformations of raw scores. These would include, for example, IQ and SAT scores. You can report the change in these familiar transformed scores as a function of being exposed to an intervention (e.g., intervention X led to a 50-point improvement in SAT scores) and be fairly confident that the results will be understood by a general audience.

One shortcoming of presenting effect sizes in terms of raw and familiar transformed scores is that the scores cannot be combined across different types of measures. For example, effects expressed as changes in SAT scores cannot be directly combined with changes in ACT scores. Thus, the results associated with each type of measure must be reported separately. This is not necessarily a bad thing if you consider it important to maintain these distinctions among outcomes, even if all measures relate to the same broader construct. If you want to describe an intervention's general effect on standardized achievement test scores, these effects will have to be standardized before you do so.

Translations of the Standardized Mean Difference

The three effect size metrics used most frequently in meta-analysis—the *d*-index, *r*-index, and odds ratio—are examples of standardized measures of effect. However, describing them to general audiences without additional explanation leaves most people scratching their heads.

For standardized mean differences, I have developed two ways to express the *d*-index for general audiences that are helpful in interpreting an intervention's effect on achievement (see Cooper, 2007). Both are based on a metric associated with the *d*-index that Cohen (1988) called U_3. U_3 represents the percentage of the units in the group with the lower mean that is exceeded by 50% of the scores in the higher-meaned group. Table 7.1 presents the equivalent values for the *d*-index and U_3. Thus, U_3 answers the question, "What percentage of the scores in the lower-meaned group was surpassed by the average score in the higher-meaned group?" For example, consider a randomly assigned group of middle-school students, one-half of whom received study skills instruction and the other half of whom received no study skills instruction. The principal outcome measure is an end-of-unit test in algebra. If the study found a *d*-index of .30, it would be associated with a U_3 value of 61.8%. This means that the average student receiving homework (50th percentile) scored higher on the unit test than 61.8% of students who received no homework.

Table 7.1 Equivalents for *d*-Index and U$_3$ Effect Size Metrics

d	U$_3$
0	50.0
.1	54.0
.2	57.9
.3	61.8
.4	65.5
.5	69.1
.6	72.6
.7	75.8
.8	78.8
.9	81.6
1.0	84.1
1.1	86.4
1.2	88.5
1.3	90.3
1.4	91.9
1.5	93.3
1.6	94.5
1.7	95.5
1.8	96.4
1.9	97.1
2.0	97.7
2.2	98.6
2.4	99.2
2.6	99.5
2.8	99.7
3.0	99.9
3.2	99.9
3.4	a
3.6	a
3.8	a
4.0	a

a. >99.9.

>99.9.

But there is no need to stop with U$_3$. It is still quite abstract and not necessarily more intuitive than the *d*-index itself. For example, staying in the educational context, U$_3$ can also be used to express the change in achievement associated with an intervention when achievement is graded on a curve. Here, you must begin by proposing the grade curve. In this case, the researcher reveals the effect by showing how the average student's grade would change if only that student received the intervention. Figure 7.3 presents one such grade curve. It also illustrates the effect that algebra homework would have on the unit test grade received by the average student (had homework not occurred). As shown in Figure 7.3, the average student in a class of students who all received study skills instruction would receive the middle C. If that student were the only student in class to get study skills instruction (and all else was unchanged), the intervention would improve the student's grade to a C+, graded on the proposed curve.

Figure 7.3 "Grading" a Hypothetical Study Skills Intervention on a Curve

Grading on a curve, the student who would have received the middle C grade on the algebra test in a class receiving study skills instruction would move up to a C+ grade had he or she been graded in a class in which no one else received instruction

When only the average student gets study skills instruction, her or his grade moves . . .

Grade	% of Scores
A	4
A–	5
B+	6.5
B	7
B–	8.5
. . . to here > C+	11.5
. . . from here > C	15
C–	11.5
D+	8.5
D	7
D–	6.5
F+	5
F	4

SOURCE: From "The search for meaningful ways to express the effects of interventions," by H. Cooper, Child Development Perspectives, 2(3). Copyright 2008 by Blackwell Publishing. Reprinted with permission.

In my example, it is critical that the researchers point out to the audience that they have supplied the grade curve and that other curves could be more or less sensitive to changes in the outcome measure. Therefore, the grade curve used in Figure 7.3 might be considered very tough by today's standards; the average student gets a C and only 9% of students get an A or A–. Had a more lenient curve been used, the middle grade could be higher than C and the discrimination of scores on the top half of the curve would be diminished. The result would suggest a lesser change in grade as a function of study skills instruction.

Why is offering an arbitrary grade curve better than providing an arbitrary yardstick, such as Cohen's adjectives, for the magnitude and significance of effects? First, the grade curve metric is perfectly transparent. All its assumptions are known and are easily displayed. All its values are familiar to most audiences. Second, because it is familiar audiences can evaluate the appropriateness of the curve and adjust the effect of the intervention on grades for themselves, if they wish. Finally, the audience does not need special expertise—that is, knowledge of which other research outcomes might have been used as yardsticks—to translate findings to other curves they find more legitimate.

My second use of U_3 gets around the problem of choosing one grade curve among the many. It shows how a student's class rank might change as a function of the intervention. For example, assume that an intervention provides a randomly chosen group of ninth graders with a course in general study skills, and the outcome measure is students' cumulative grade point average upon graduation. Assume as well that the effect of the intervention is again $d = .3$ and $U_3 = 61.8\%$. In this scenario, the student who would have placed in the middle of the final class ranking (50%) would surpass 11% more students if he or she were the only student to receive instruction (11% is the rounded difference between the 50th percentile student and the 61.8th percentile student). Figure 7.4 presents this result visually for a graduating class with 100 students.

Figure 7.4 Hypothetical Change in a Student's Class Rank Due to Study Skills Instruction
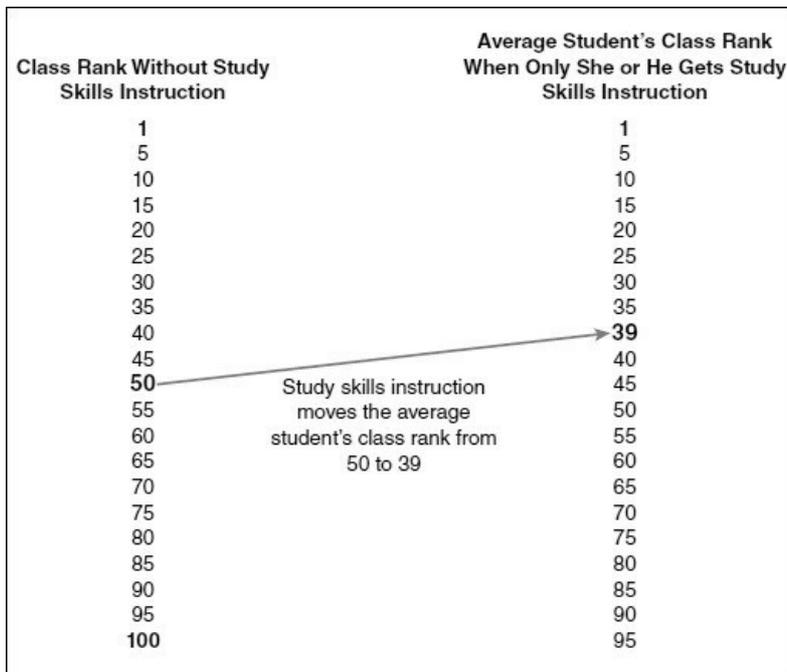
Class Rank Without Study Skills Instruction	Average Student's Class Rank When Only She or He Gets Study Skills Instruction
1	1
5	5
10	10
15	15
20	20
25	25
30	30
35	35
40	**39**
45	40
50	45
55	50
60	55
65	60
70	65
75	70
80	75
85	80
90	85
95	90
100	95

Study skills instruction moves the average student's class rank from 50 to 39

SOURCE: From "The search for meaningful ways to express the effects of interventions," by H. Cooper, Child Development Perspectives, 2(3). Copyright 2008 by Blackwell Publishing. Reprinted with permission.

These are just two examples of how standardized effect sizes can be contextualized to convey greater intuitive meaning to general audiences. The grade curve translation is most meaningful when applied to outcome measures that are natural candidates for grading on a curve, such as class exams. However, the need to provide a grading curve is a drawback to its use. The class rank translation is most meaningful in the context of high school interventions that are meant to have general effects on achievement, as measured by cumulative grade point averages, and class rank has meaning because of its use in college admissions. One of your creative challenges is to think of appropriate metrics for the results of your research synthesis and how these can be conveyed to your audience in a meaningful way.

Translations of Binomial Effect Size Display

Rosenthal and Rubin (1982) provide a translation of the effects of discrete interventions on dichotomous outcomes, called *the binomial effect size display* (BESD). They suggest it could be used for other effect size metrics as well. The BESD transforms a *d*-index and *r*-index into a 2 × 2 table with the marginals assumed to be equal for both rows and columns. In their examples, Rosenthal and Rubin assume 100 participants, with 50 in each of the two conditions and 50 outcomes indicating intervention success and 50 indicating failure. They show that Cohen's relatively small effect of $d = .20$ (equivalent to $r = .1$, explaining 1% of the variance) is associated with an increase in success rates from 45% to 55%. For example, an intervention meant to increase students' reading scores above a proficiency threshold with this effect size would mean that 10 more children in every 100 would meet the minimum requirement. This should be a metric that most general audiences will understand.

The BESD is not without its critics (little is in this area), especially because of its assumptions regarding marginal values (Randolph & Shawn, 2005). Even so, it seems that the BESD is an intuitively appealing expression of effect when the intervention outcome is dichotomous, and even more so when the observed marginals can be retrieved. Indeed, when this information is available, the BESD reduces to a display of raw score results. Its application is more difficult when it requires the audience to mentally convert continuous outcome measures into dichotomous ones.

Translations of Effects Involving Two Continuous Measures

Providing translations for associations between two continuous variables—*r*-indexes and ß-weights—requires knowledge of the raw scales and the standard deviations of the predictor and outcome. With this information, you can describe the change in outcome associated with a specified additional amount of exposure to the intervention. For example, assume a predictor variable is the number of minutes a child with a behavior problem spends in counseling each week, and the standard deviation for this variable is 30 minutes. The outcome variable is the number of absences from school, and its standard deviation is 4. Both are measured across a full school year. In such case, a ß-weight or *r*-index of –.50 would mean that,

on average, students in the sample who spent 30 more minutes in counseling each week also had two fewer absences that year.

Conclusion

To conclude, then, along with analyses that examine the impact of missing data and varying assumptions for the statistical analyses, the next question you should ask when evaluating the interpretation of effect sizes in meta-analysis is,

> Did the synthesists (a) contrast the magnitude of effects with other related effect sizes and/or (b) present a practical interpretation of the significance of the effects?

A complete and careful assessment of the generality of the synthesis' findings and the confidence with which you can draw causal inferences from it are also critical parts of how you will interpret the findings of a research synthesis.

Exercises

Find two primary research reports on the same topic that vary in method. Then

1. Calculate the effect sizes reported in each.
2. Compare the effect sizes to one another, taking into account the influences of their different methods.

   Decide whether you consider the magnitude of the effect sizes to be
      1. Large, medium, or small; and
      2. Important or not important.
4. Justify your decision.

Notes

1. Portions of this discussion include minor modifications of a similar discussion I provided in Cooper (2009).

8 Step 7 Presenting the Results

What information should be included in the report of the synthesis?

<div style="border:1px solid black; background:#d9d9d9; padding:1em;">

Primary Function in Research Synthesis

To identify the aspects of methods and results readers of the report will need to know to evaluate the synthesis

Procedural Variation That Might Produce Differences in Conclusions

Variation in reporting might (a) lead readers to place more or less trust in synthesis outcomes and (b) influence others' ability to replicate results.

Question to Ask When Presenting the Research Synthesis Methods and Results

Were the procedures and results of the research synthesis documented clearly and completely?

This chapter describes

- A format for research syntheses reports
- How to present tabulated data in syntheses

</div>

The transformation of your notes, printouts, and coding forms into a cohesive public document describing your research synthesis is a task with profound implications for the accumulation of knowledge. All your efforts to conduct a trustworthy and convincing integration of the research literature will be for naught unless you pay careful attention to how your synthesis is described in the report.

Report Writing in Social Science

The codified guidelines used by many social science disciplines for reporting primary research are contained in APA's *Publication Manual* (APA, 2010). The *Publication Manual* is quite specific about the style and format of reports, and it even gives some guidance concerning grammar and the clear expression of ideas. It tells researchers how to set up a manuscript page, what the major section headings should be, and what conventions to use when reporting the results of statistical analyses, among many other details of report preparation. Naturally, however, it is much less explicit in guiding judgments about what makes a finding important to readers. It would be impossible to explicate a general set of rules for defining the scientific importance of results. Hopefully, the previous chapter has provided you with some guidance on how to interpret the findings of your research synthesis.

Because the integration of research results has grown in importance, several attempts have been made to develop standards for the reporting of research syntheses, especially those that contain meta-analyses. Several proposals regarding what information should be included in the report of a meta-analysis come from researchers and statisticians in the medical sciences. The Equator Network (2015) keeps track of developments in research synthesis reporting standards as well as other types of research. In the social sciences, a task force of the APA proposed a set of reporting standards for meta-analysis, called MARS (Meta-Analysis Reporting Standards; APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008).[1] The MARS was incorporated into the APA *Publication Manual* (APA, 2010). The MARS was constructed by, first, comparing the content of many of the aforementioned standards and developing a list of elements contained in any of these. Second, the items on this list were rewritten to make the terms used in them more familiar to social science audiences. Third, the members of the working group added some items of their own. Then, this set of items was shared with members of the Society for Research Synthesis Methodology, who were asked to make suggestions about the inclusion of other items or the removal of items that seemed unnecessary. Finally, the Publications and Communications Board of the APA reacted to the items. After receiving these reactions, the working group arrived at the list of recommendations contained in Table 8.1. The emergence of these reporting guidelines is critical to progress in the social sciences because they will promote the complete and transparent reporting of methods and results for meta-analyses. Next, I will provide a bit more context and detail regarding the items in the MARS.

Meta-Analysis Reporting Standards

As Table 8.1 reveals, the format for reporting meta-analyses has evolved to look a lot like that of reports

of primary research, with an introduction, method section, results section, and discussion. If a research synthesis does not include a meta-analysis, there is still much sound advice for preparing a report in Table 8.1, though many of the items listed under the "Method" and "Results" sections would be irrelevant. In the following, I will assume that your report is describing the results of a research synthesis that employed meta-analytic techniques.

Title

It is important that the title of your report include the term *meta-analysis* if one was conducted, or *research synthesis, research review,* or a related term, if a meta-analysis was not performed. These terms are very informative about what is contained in your report. Also, people who are searching the literature for documents on your topic using a computerized reference database or online search may use one of these terms if they are interested in finding only those documents that contain summaries of the literature. If your title does not contain one of these terms and a search is conducted on titles only, your report will not be included in the search results. So, for example, our title "A Meta-Analysis on the Effect of Choice on Intrinsic Motivation and Related Outcomes" includes the three terms most likely to be used in a search by someone interested in finding documents like ours.

Table 8.1   Meta-Analysis Reporting Standards

Paper Section and Topic	Description
Title	• Make it clear that the report describes a research synthesis and include "meta-analysis," if applicable • Footnote funding source(s)
Abstract	• The problem or relation(s) under investigation • Study eligibility criteria • Type(s) of participants included in primary studies • Meta-analysis methods (indicating whether a fixed or random model was used) • Main results (including the more important effect sizes and any important moderators of these effect sizes) • Conclusions (including limitations) • Implications for theory, policy, and/or practice

Introduction	• Clear statement of the question or relation(s) under investigation
	  o Historical background
	  o Theoretical, policy, and/or practical issues related to the question or relation(s) of interest
	  o Rationale for the selection and coding of potential moderators and mediators of results
	  o Types of study designs used in the primary research, their strengths and weaknesses
	  o Types of predictor and outcome measures used, their psychometric characteristics
	  o Populations to which the question or relation is relevant
	  o Hypotheses, if any

Method	
Inclusion and Exclusion Criteria	• Operational characteristics of independent (predictor) and dependent (outcome) variable(s) • Eligible participant populations • Eligible research design features (e.g., random assignment only, minimal sample size) • Time period in which studies needed to be conducted • Geographical and/or cultural restrictions

Table 8.1 (Continued)

Paper Section and Topic	Description
Moderator and Mediator Analyses	• Definition of all coding categories used to test moderators or mediators of the relation(s) of interest

Search Strategies	• Reference and citation databases searched
	• Registries (including prospective registries) searched
	   ○ Keywords used to enter databases and registries
	   ○ Search software used and version
	• Time period in which studies needed to be conducted, if applicable
	• Other efforts to retrieve all available studies, e.g.,
	   ○ Listservs queried
	   ○ Contacts made with authors (and how authors were chosen)
	   ○ Reference lists of reports examined
	• Method of addressing reports in languages other than English
	• Process for determining study eligibility
	• Aspects of reports examined (i.e., title, abstract, and/or full text)
	   ○ Number and qualifications of relevance judges
	   ○ Indication of agreement
	   ○ How disagreements were resolved
	• Treatment of unpublished studies

Coding Procedures	• Number and qualifications of coders (e.g., level of expertise in the area, training)
	• Intercoder reliability or agreement
	• Whether each report was coded by more than one coder and, if so, how disagreements were resolved
	• Assessment of study quality
	   ○ If a quality scale was employed, a description of criteria and the procedures for application
	   ○ If study design features were coded, what these were
	• How missing data were handled

Paper Section and Topic	Description
Statistical Methods	- Effect size metric(s) - Effect sizes calculating formulas (e.g., means and SDs, use of univariate *F-to-r* transform, etc.) - Corrections made to effect sizes (e.g., small sample bias, correction for unequal sample sizes, etc.) - Effect size averaging and/or weighting method(s) - How effect size confidence intervals (or standard errors) were calculated - How effect size credibility intervals were calculated, if used - How studies with more than one effect size were handled - Whether fixed- and/or random-effects models were used and the model choice justification - How heterogeneity in effect sizes was assessed or estimated - Means and SDs for measurement artifacts, if construct-level relationships were the focus - Tests and any adjustments for data censoring (e.g., publication bias, selective reporting) - Tests for statistical outliers - Statistical power of the meta-analysis - Statistical programs or software packages used to conduct statistical analyses
Results	- Number of citations examined for relevance - List of citations included in the synthesis - Number of citations relevant on many but not all inclusion criteria excluded from the meta-analysis - Number of exclusions for each exclusion criteria (e.g., effect size could not be calculated), with examples - Table giving descriptive information for each included study, including effect size and sample size - Assessment of study quality, if any

Table 8.1 (Continued)

Paper Section and Topic	Description

Results	• Tables and/or graphic summaries
	  ○ Overall characteristics of the database (e.g., number of studies with different research designs)
	  ○ Overall effect size estimates, including measures of uncertainty (e.g., confidence and/or credibility intervals)
	• Results of moderator and mediator analyses (analyses of subsets of studies)
	  ○ Number of studies and total sample sizes for each moderator analysis
	  ○ Assessment of interrelations among variables used for moderator and mediator analyses
	• Assessment of bias including possible data censoring

Discussion	• Statement of major findings
	• Consideration of alternative explanations for observed results
	  ○ Impact of data censoring
	• Generalizability of conclusions, e.g.,
	  ○ Relevant populations
	  ○ Treatment variations
	  ○ Dependent (outcome) variables
	  ○ Research designs, etc.
	• General limitations (including assessment of the quality of studies included)
	• Implications and interpretation for theory, policy, or practice
	• Guidelines for future research

SOURCE: APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008).

SOURCE: APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008).

Abstract

The abstract for a research synthesis follows the same rules as abstracts for primary research. Because an abstract is short, you can spend only a sentence or two on stating the problem, the kinds of studies that were included in the meta-analysis, your method and results, and major conclusions. As with the title, it is important to think about people doing literature searches when writing your abstract. Remember to include the terms you think searchers who are interested in your topic are likely to pick when they construct their computer searches. Also, remember that many people will read only your abstract, so you must tell them the most important things about your meta-analysis.

The Introduction Section

The introduction to a research synthesis sets the stage for the empirical results that follow. It should contain a conceptual presentation of the research problem and a statement of the problem's significance. Introductions are typically short in primary research reports. In research syntheses, introductions should be considerably more detailed. You should attempt to present a complete overview of the research question, including its theoretical, practical, and methodological history. Where do the concepts involved in the research come from? Are they grounded in theory (as is, for example, the notion of intrinsic motivation) or in practical circumstances (as is the notion of homework)? Are there theoretical debates surrounding the meaning or utility of the concepts? How do theories predict that the concepts will be related to one another? Are there conflicting predictions associated with different theories? What variables do different theories, scholars, or practitioners suggest might influence the strength of the relation?

The introduction to a research synthesis must contextualize the problem under consideration. Especially when the synthesist intends to report a meta-analysis, it is crucial that ample attention be paid to the

qualitative and historical debates surrounding the research question. Otherwise, you will be open to the criticism that numbers have been crunched together without ample appreciation for the conceptual and contextual underpinnings that give empirical data their meaning.

Once the context of the problem has been laid out, the introduction then should describe how the important issues you have identified have guided your decisions about how the meta-analysis was conducted. How did you translate the theoretical, practical, and historical issues and debates into your choices about what moderator variables to explore? Were there issues of concern regarding how studies were designed and implemented, and were these represented in your meta-analysis?

The introduction to a research synthesis is also where you should discuss previous efforts to integrate the research on the topic. This description of past syntheses should highlight what has been learned from these efforts as well as point out their inconsistencies and methodological strengths and weaknesses. The contribution of your new effort should be emphasized by a clear statement of the unresolved empirical questions and controversies addressed by your new work.

In sum, the introduction to a research synthesis should present a complete overview of the theoretical, conceptual, and/or practical issues surrounding the research problem. It should present the controversies in the topic area still to be resolved, and indicate which of these were the focus of the new synthesis effort. It should present a description of prior syntheses, what their contribution and shortcomings were, and why your synthesis is innovative and important.

The Method Section

The purpose of a method section is to describe operationally how the research was conducted. The method section of a research synthesis will be considerably different from that of a primary research report. The MARS suggests that a meta-analysis method section will need to address five separate sets of questions—(a) inclusion and exclusion criteria, (b) moderator and mediator analyses, (c) search strategies, (d) coding procedures, and (e) statistical methods. The order in which they are presented can vary, but you should consider using these topics as subheadings in the report.

Inclusion and exclusion criteria.

The method section should address the criteria for relevance that were applied to the studies uncovered by the literature search. What characteristics of studies were used to determine whether a particular effort was relevant to the topic of interest? For example, in the synthesis of research on the effects of choice on intrinsic motivation, three criteria had to be met by every study included in the synthesis: (a) the study had to include an experimental manipulation of choice (not a naturalistic measure of choice); (b) the study had to use a measure of intrinsic motivation or a related outcome, such as effort, task performance, subsequent learning, or perceived competence; and (c) the study had to present enough information to allow us to compute an effect size.

Next, you need to describe what characteristics of studies would have led to their exclusion from the synthesis, even if they otherwise met the inclusion criteria. You should also state how many studies were excluded for any given reason. For example, the meta-analysis of the effect of choice on intrinsic motivation excluded studies that met the three inclusion criteria but were conducted on populations with a special characteristic or in a country other than the United States or Canada. This led to the exclusion of two studies conducted on children with learning disabilities or behavior disorders and eight studies conducted outside North America.

When readers examine the relevance criteria employed in a synthesis, they will be critically evaluating your notions about how concepts and operations fit together. Considerable debate about the outcomes of a particular synthesis may focus on these decisions. Some readers may find that your relevance criteria were too broad—operational definitions of concepts were included that they believe were irrelevant. Of course, you can anticipate these concerns and rather than exclude studies based on them use the debatable criteria as distinctions between studies and then analyze them as potential moderators of study results. Other readers may find that your operational definitions were too narrow. For example, some readers might think that we should have included samples from countries outside North America in our synthesis on choice and intrinsic motivation. However, we justified our decision by pointing out that very few studies were found that used non–North American samples and only a few countries were represented among these few studies. Therefore, we believed that including these studies still would not have warranted generalizing our conclusions to people other than those living in North America. Moderator analyses could have been used to determine whether the effect of choice varied depending on the country sampled, but we believed there were too few studies to reliably conduct such an analysis. Still, these exclusion criteria might lead readers to examine excluded studies to determine if including their findings would affect the synthesis outcome.

In addition to this general description of the included and excluded evidence, this subsection is a good place to describe the typical methodologies found in primary research. The presentation of prototype

studies is a good way to present methods that are used in many studies. You can choose several studies that exemplify the methods used in many other studies and present the specific details of these investigations. In instances where only a few studies are found to be relevant, this exercise may not be necessary—the description of the methods used in each study can be combined with the description of the study's results. In our meta-analysis of homework, we took this approach to describing the methods and results of the few studies that used experimental manipulations of homework.

Moderator and mediator analyses.

Similar to the inclusion and exclusion criteria, the descriptions you give of the variables you tested as moderators or mediators of study results let readers know how you defined these variables and, especially, how you chose to distinguish among studies based on the studies' different status on these variables. So, for example, the meta-analysis on choice and intrinsic motivation identified "the number of options per choice" as a potential variable that might mediate the effect of choice. Our method section defined this variable and told readers that we grouped the studies into those that provided (a) two options per choice, (b) three to five options, or (c) more than five options.

Searching strategies.

Information on the procedures, sources, keywords, and years covered by the literature search allows the reader to assess the thoroughness of your search and therefore how much credibility to place in the conclusions of the synthesis. In terms of attempted replication, it is the description of the literature search that first would be examined when other scholars attempt to understand why different syntheses on the same topic have come to similar or conflicting conclusions. It is also good to include a rationale for the choice of sources, especially with regard to how different sources were used to complement one another in order to reduce bias in the sample of studies. The MARS lists 16 different aspects of a literature search for you to address in the methods section. Atkinson, Koenka, Sanchez, Moshontz, and Cooper (2015) expanded this list to cover more specific details about who did the coding, results of the application of inclusion and exclusion criteria, and aspects of the initial screening for relevance. Their summary is presented in Table 8.2.

The results of a search often can be neatly summarized in a table. For example, Brunton and Thomas (2012) used a diagram suggested by PRISMA (2015, which was developed for syntheses in health but is more generally relevant) to present the results of a search looking for studies on the effectiveness of personal development planning (reflecting, recording, planning and actions) to improve learning. A copy of their table is presented in Figure 8.1.

Brunton and Thomas note that many times more documents will be examined than will be included in the synthesis. This is typical. Also, the boxes in the table are not standard, though most are nearly universally used. You can change these to help your reader understand what you did in your particular circumstance.

Coding procedures.

A third subsection of methods should describe the characteristics of the people who retrieved information from the studies, the procedures used to train them, and how the reliability of the retrieved information was assessed, as well as what this assessment revealed. Often, these will be the same people who searched the literature and made relevance decisions. If so, this should be mentioned.

It is also important to discuss in the coding procedures section how missing data were handled. For example, in our meta-analysis of choice and intrinsic motivation, we used our ability to calculate an effect size as an inclusion criterion. So studies were examined to determine whether an effect size could be calculated from them before the process of coding other information began. If no effect size was retrievable, no further coding occurred. In other meta-analyses, estimation procedures might be used to fill in these blanks. The same is true for other missing study characteristics. For example, if a study lacks information on whether random assignment was used, the study might be represented as giving no such information. Other times, you might develop a convention that says if the use of random assignment was not mentioned, it was assumed the study did not use random assignment. These kinds of rules should be described in this section of your report.

The coding section also can be where you described how you made judgments about study quality. The decision about where to put information on study quality really can fit in several sections, so it is best placed where it provides for the clearest exposition. If studies were excluded based on features of their design or implementation, this would be reported with other inclusion and exclusion criteria.

Table 8.2 Reporting Standards for Literature Searches in Research Syntheses

Searcher Characteristics	• Number of searchers • Level of education and/or training • Past experience conducting searches • Were initial screeners different from second screeners? ○ If yes, how?
Initial Screening for Relevance	• Elements of reports used in initial screening decision ○ e.g., title, abstract, full report • Criteria for passing from first to second screen

Final Inclusion Criteria	• Definitions
	  ○ Researched variables
	  ○ Participants
	  ○ Researched settings
	  ○ Dates
	    ▫ For conduct of research
	    ▫ For report appearance
	• Types of reports, e.g.,
	    ▫ Publication status
	    ▫ Peer review status
	• Treatment of studies reported in foreign languages
	• Adequate reporting
	  ○ Information needed to include the report

Excluded Articles	• Listing of studies that met most inclusion criteria but were ultimately excluded
	  ○ At least one criteria that each study failed to meet
	    ▫ For example, data presented in unusable way

General search reporting guidelines	• Total number of documents in the search results • Number of reports retained after initial screening • Number of reports meeting relevance criteria ○ Number of reports excluded due to insufficient information

Reference Database & Registry Searches	• Search software version
	• Full names of databases searched
	   o Justification of database choice
	• Search terms used
	   o Boolean connectors
	• Whether automatic explosion terms occurred
	• Parts of text searched
	• Date database search(es) were conducted

Journal-Bibliography and Registry Search	• Journals or bibliographies scanned for relevant reports
	   o names
	   o yes/volumes scanned
	   o document elements used for decision
	   o e.g., tables of contents, titles, abstracts

Backward (Reference List) Search	• Whether reference lists were examined • Criteria for reference lists chosen • Element used for relevance decision (e.g., references, mention in Introduction, abstract, full report)
Forward (Citation) Search	• Reports for which citation searches were conducted o justification for why these reports were chosen

Direct Contact Searches	• Mass communications to formal or informal distribution lists
	   o group names or defining characteristics
	   o dates of contact
	• Communications to individual researchers
	   o criteria for decisions to contact
	   o number of researchers contacted
	   o response rate
	• Communications with colleagues (leading to reports found no other way)

Other Search Strategies	• Describe search strategies in addition to those above and the results of these searches

Figure 8.1 A PRISMA Flowchart Describing the Outcomes of a Literature Search

SOURCE: Brunton, J. & Thomas, J. (2012). Information management in reviews. In D. Gough, S. Oliver & J. Thomas (Eds.). An introduction to systematic reviews. Thousand Oaks, CA: Sage.

Statistical methods.

The final topics described in the method section of a research synthesis are the procedures and conventions used to carry out any quantitative analysis of results. Why was a particular effect size metric chosen and how was it calculated? What analyses techniques were used to combine results of separate tests of a hypothesis and to examine the variability in findings across tests? This section should contain a rationale for each procedural choice and convention you use and should describe what the expected impact of each choice might be on the outcomes of the research synthesis.

Another important topic to cover in this subsection concerns how you identified independent findings (see Chapter 4). You should carefully spell out the criteria used to determine how to treat multiple hypothesis tests from the same laboratory, report, or study.

The Results Section

The results section should present a summary description of the literature and the findings of the meta-analysis. It should also present any results of the synthesis used to test the implications of different assumptions about the data, such as different models of error and different patterns of missing data. While the results sections of syntheses will vary considerably depending on the nature of the research topic and evidence, the MARS provides a good general strategy for presenting results. Next, I suggest some possible subsections for organizing the presentation of results, along with some suggestions regarding how to visually display your findings in tables and figures. Additional suggestions regarding the presentation of data in meta-analysis can be found in Borman and Grigg (2009).

Results of the literature search.

Often, synthesists will present a table that lists all the studies included in the meta-analysis. This table will also describe a few critical characteristics of each study. For example, Table 8.3 reproduces the table we used in the synthesis of homework research to describe the six studies that tested the effects of homework using an experimental manipulation. We decided that the most important information to include in the table, along with the name of the first author and year of report appearance, was the

research design, the number of classes and students included in the study, the students' grade level, the subject matter of the homework assignment, the achievement outcome measure, and the effect size. Nearly all tables of this sort include information on author and year, sample size, and effect size, so this is a pretty simple example. Sometimes the information you want to present in this table will be extensive. If so, you may want to use abbreviations. Table 8.4 presents such a table reproduced from our meta-analysis on choice and intrinsic motivation. In this case, we resorted to an extensive footnote to describe the abbreviations.

Table 8.3   Studies That Manipulated Homework Versus No-Homework Conditions

Author (Year)	Research Design	Classes Students ESS[a]	Grade Level	Subject Matter	Type of Achievement Measure	Effect Size
Finstad (1987)	Nonequivalent control with no pretest differences	2 39 5.2	2	Places to 100	Unit test developed by Harcourt Brace Jovanovich	+.97
Foyle (1984)	Randomized by class Analyzed by student	6 131 15.8	9-12	American history	Unit test developed by the teacher	+.46
Foyle (1990)	Randomized by class Analyzed by student	4 64 10.2	5	Social studies	Unit test developed by the teacher	+.90
McGrath (1993)	Randomized within class Analyzed by student	3 94 8.0	12	Shakespeare	Unit test developed by Harcourt Brace Jovanovich	+.39
Author (Year)	**Research Design**	**Classes Students ESS[a]**	**Grade Level**	**Subject Matter**	**Type of Achievement Measure**	**Effect Size**
Meloy (1987)	Unlucky random assignment followed by nonequivalent control with pretest	5 70 12.6 3 36 7.4	3 4	English skills	Unit test developed by McDougal Littell	+
					Researcher-shortened version of the Iowa Test of Basic Skills Language subtest	–
					Unit Test developed by McDugal, Little	+
					Researcher-shortened version of the Iowa Test of Basic Skills Language subtest	–
Townsend (1995)	Nonequivalent control without equating	2 40 5.2	3	Vocabulary	Unit test developed by the teacher	+.71

a. ESS stands for effective sample size based on an assumed intraclass correlation of .35.

SOURCE: Cooper, H., Robinson, J. C., & Patall, E. A. (2006). Does homework improve academic achievement? A synthesis of research, 1987–2003. Review of Educational Research, 76, 1–62. Copyright 2006 by the American Educational Research Association. Reprinted with permission.

ESS stands for effective sample size based on an assumed intraclass correlation of .35.

SOURCE: Cooper, H., Robinson, J. C., & Patall, E. A. (2006). Does homework improve academic achievement? A synthesis of research, 1987–2003. Review of Educational Research, 76, 1–62. Copyright 2006 by the American Educational Research Association. Reprinted with permission.

Table 8.4 Characteristics of Selected Experimental Studies Examining the Effects of Choice on Intrinsic Motivation

Author (Year)	Type of Document	Sample	Number of Choices	Options	Choice Type	Control Group Type	Knowledge of Alternatives	Design	Setting	Reward Condition	Outcome	Measure Type	Effect Size
Abrahams 1 (1988)	D	48a A	4 SC	IND	IR	RAC	UAW	Y	TUL	NRW	FCTS	B	+.90
											I/E/L	S	+.84
											TP	B	+.18
Abrahams 2 (1988)	D	42b A	4 SC	IND	IR	RAC	UAW	Y	TUL	NRW	FCTS	B	+.51
											I/E	S	+.12
											WTE	S	+.41
Amabile & Gitomer (1984)	J	28 C	5 MC	10	IR	NSOC	AW	Y	LNS	NRW	FCTS	B	+.79
											CR	S	+1.06
Bartleme (1983)	D	104 A	8 MC	IND	IR	RAC	UAW	Y	TUL	CLPSD	E/L	S	+.07
											E/L	S	−.11
											E/L	S	+.08
											WTE	S	−.16
											TP	B	−.05
											SL	B	−.22
		34 A								NRW	E/L	S	+.46
											E/L	S	−.53
											E/L	S	+.15
											WTE	S	+.10
											TP	B	+.17
											SL	B	−.22

Author (Year)	Type of Document	Sample	Number of Choices	Options	Choice Type	Control Group Type	Knowledge of Alternatives	Design	Setting	Reward Condition	Outcome	Measure Type	Effect Size
		70 A								RWd	E/L	S	−.17
											E/L	S	−.05
											E/L	S	+.10
											WTE	S	−.56
											TP	B	−.17
											SL	B	−.17
Becker (1997)	J	41 A	1	2	IR	NSOC	UAW	M	NS	NRW	GIM	S	+.58
											TP	B	+1.25

As Atkinson et al. and (2014) suggest, you may also want to provide a table that describes the studies that were potentially relevant but were excluded. The MARS suggests these studies include those that were relevant on many but not all criteria used to define a study as relevant. This table might look like Table

8.3 or 8.4; it is usually not as extensive and contains columns that identify the relevance criteria or at least a column that explains the criteria that led to the study's exclusion.

Table 8.4 contains only a small portion of the studies that appeared in the actual table. Because tables that describe the studies that went into a meta-analysis can be quite long, journals are now providing auxiliary websites on which this and other material can be placed, rather than including it in the printed version of the manuscript. In electronic versions of articles, the tables may reside on separate web pages but be linked to the article at the point in the report that they would otherwise appear. When you submit your report for publication, you should be sure to include these tables (in the report or in a separate document); when your paper is accepted, you and the editor will decide what the best strategy is to present your results.

Assessment of study quality.

If you conducted an assessment of the quality of each study, this can be included in the described tables. Or, if the judgments were complex, you might consider presenting them in a table of their own. For example, the information in Table 5.3 could be presented in a table in which the quality dimensions are presented in columns and quality ratings (the "yes" and "no" in Table 5.3) are given in separate rows devoted to each study.

Aggregate description of the literature.

Certain aggregate descriptive statistics about the literature should be reported as well. Table 8.5 presents the section of our homework meta-analysis that presented the aggregate results for studies that correlated a measure of the amount of homework a student did and the student's achievement. This subsection includes the following elements:

- The number of studies, effect sizes, and samples that went into the meta-analysis
- A description of studies that caused any differences in these numbers—that is, studies with more than one sample and/or outcome measure
- The range of years in which reports appeared[2]
- The total number of participants across all studies and the range, median, mean, and variance of sample sizes within studies
- A test for statistical outliers among the sample sizes
- The variables that could not be tested as moderators because either (a) too many studies were missing this information or (b) there was insufficient variation across studies
- The number of positive and negative effect sizes
- The range of and median effect size
- The unweighted and weighted mean effect size and the confidence interval for the weighted mean
- A test for statistical outliers among the effect sizes
- The results of a test for missing data and how adjusting for missing data affected the cumulative results

You may also consider putting some of this information in a table, if you believe some of the nuances in the data and the rationales need no additional explanation that might be lost in a tabular presentation, for example they are included in the methods text. Table 8.6 presents the results of a meta-analysis that asked the question "What is the correlation between college student self-grades and instructor grades when they mark the same test?"

Graphic presentation of results.

A good way to present the results of your meta-analysis is to use what is called a Forrest Plot. Figure 8.2 presents a Forrest Plot of the results of the hypothetical meta-analysis I used in Chapter 6 to illustrate the mechanics of the calculations (Table 6.4). This figure was generated by the Comprehensive Meta-Analysis software package (2015; Borenstein, Hedges, Higgins, & Rothstein, 2005). The first three columns of the figure present the study number, whether it was a member of Moderator Group A or B, and its total sample size.[3] The next three columns give each study's correlation and the lower and upper limits of its 95% confidence interval. The Comprehensive Meta-Analysis program would let me report other statistics here as well. The Forrest Plot part of the figure is on the right. This graph presents each correlation in what is called a *box-and-whiskers* display. The box is centered on the value of the study's correlation. The size of the box is proportional to the study's sample size relative to the other studies in the meta-analysis. The length of the whiskers depicts the correlation's confidence interval. Note as well that the figure includes the weighted average correlations and confidence intervals for the Group A and B studies and for the overall set of studies (using a fixed-effect model; a random-effect model could also have been requested). These averages are depicted on the Forrest Plot as diamonds rather than as boxes and whiskers.[4] This type of figure is growing in popularity for the presentation of meta-analytic results.

Table 8.5 An Example of a Text Summary of Aggregate
Meta-Analysis Results

The literature search uncovered 32 studies that described the correlations between the time that a student spent on homework, as reported by either the student or a parent, and a measure of academic achievement. These studies are listed in Table 8. The 32 studies reported 69 separate correlations based on 35 separate samples of students. Cooper et al. (1998) reported 8 correlations, separating out effects for elementary and secondary students (two independent samples) on both class grades and standardized tests with time-on-homework reported by either students or parents. Drazen (1992) reported 12 correlations: for reading, for math, and for multiple subjects for three national surveys (three independent samples). Bents-Hill (1988) reported eight correlations: for language arts, for math, for reading, for multiple subjects for both class grades and for a standardized test of achievement. Epstein (1988), Olson (1988), and Walker (2002) each reported two effect sizes, for math and for reading. Fehrmann et al. (1992), Wynn (1996), and Keith and Benson (1992) each reported two correlations: for class grades and for achievement test results. Hendrix et al. (1990) reported three correlations: for multiple subjects, for verbal ability, and for nonverbal ability. Mau & Lynn (2000) reported three correlations: for math, for reading, and for science. Singh et al. (2002) reported two correlations: one for math and one for science.

The 32 studies appeared between the years 1987 and 2004. The sample sizes ranged from 55 to approximately 58,000, with a median size of 1,584. The mean sample size was 8,598 with a standard deviation of 12,856, suggesting a non-normal distribution. The Grubbs test revealed a significant outlier, $p < .05$. This sample was the largest in the dataset, reported by Drazen (1992) for six correlations obtained from the 1980 High School and Beyond longitudinal study. As a result, we replaced these six sample sizes with the next largest sample size in the dataset: 28,051. The mean sample size for the adjusted dataset was 7,742 with a standard deviation of 10,192.

Only three studies specifically mentioned that students were drawn from regular education classrooms; one of these studies included learning disabled students as well (Deslandes, 1999). The remaining studies did not report information on the students' achievement or ability level.

Seventeen studies did not report information on the socioeconomic status (SES) of students, eleven reported that the sample's SES was "mixed," 3 described the sample as middle SES, and one as lower SES. Seventeen studies did not report the sex make-up of the sample while 14 reports said the sample was comprised of both sexes. Only one study reported correlations separately for males and females. Because of a lack of reporting or variation across categories, no analyses were conducted on these variables.[1]

Of the 69 correlations, 50 were in a positive direction and 19 in a negative direction. The mean unweighted correlation across the 35 samples (averaging multiple correlations within each sample) was $r = .14$, the median was $r = .17$, and the correlations ranged from $- .25$ to .65.

The weighted average correlation was $r = .24$ using a fixed-error model with a 95% confidence interval (95% CI) from .24 to .25. The weighted average correlation was $r = .16$ using a random-error model with a 95% confidence interval from .13 to .19. Clearly, then, the hypothesis that the relationship between homework and achievement is $r = 0$ can be rejected under either error model. There were no significant outliers among the correlations so all were retained for further analysis.

The trim-and-fill analyses were conducted in several different ways. We performed the analyses looking for asymmetry using both fixed and random error models to impute the mean correlation and creating graphs using both fixed and random models (see Borenstein et al., 2005) while searching for possible missing correlations on the left side of the distribution (those that would reduce the size of the positive correlation). None of the analyses produced results different from those described above. Using a random error model, there was evidence that three effect sizes might have been missing; imputing them would lower the mean correlation to $r = .23$ (95% CI = .22/.23). The random error results of this analysis were $r = .14$ (95% CI = .11/.17).

Table 8.6 Correlations Between Student and Instructor Grades

# of studies contributing correlations: 28
# of correlations (effect sizes): 62
# of independent samples: 37
Range of sample sizes: 16–3588
    Outliers: 3588, 490
    Moved to next nearest neighbor: Both changed to 230
Range of correlations using outcome values as the unit of analysis: –0.03 to 0.98
    # Positives: 61
    # Negatives: 1
    Outliers: None
Range of correlations using the independent sample as unit of analysis: .10–0.98
    # Positives: 37
    # Negatives: 0
    Outliers: None
Weighted average r-index using the independent sample as unit: 0.71
    CI 95% (random effects model)
    High: 0.79
    Low: 0.62
    Tau: 0.52
    I-squared: 97.04

SOURCE: Atkinson, Sanchez, Koenka, Moshontz, and Cooper (2015). Reproduced with permission.

SOURCE: Atkinson, Sanchez, Koenka, Moshontz, and Cooper (2015). Reproduced with permission.

Another good way to graphically present the effect sizes that contribute to a meta-analytic database is in the form of a *stem-and-leaf display.* In a simple stem-and-leaf display the first decimal place of each effect size acts as the stem, which is placed on the left side of a vertical line. The second decimal place acts as the leaf, placed on the right side of the vertical line. Leaves of effect sizes sharing the same stems are placed on the same line.

Figure 8.2 Forrest Plot of Hypothetical Meta-Analysis Conducted in Chapter 6

Model	Study name	Group by Moderator	Total	Correlation	Lower limit	Upper limit
	1.000	A	3505	0.06	0.03	0.09
	2.000	A	3606	0.12	0.09	0.15
	3.000	A	4157	0.22	0.19	0.25
Fixed		A		0.14	0.12	0.16
	4.000	B	1021	0.08	0.02	0.14
	5.000	B	1955	0.27	0.23	0.31
	6.000	B	12146	0.26	0.25	0.28
Fixed		B		0.25	0.24	0.27
Fixed		Overall		0.20	0.19	0.22

Correlation and 95% CI: scale from -0.35, -0.18, 0.00, 0.18, 0.35. Favors A — Favors B
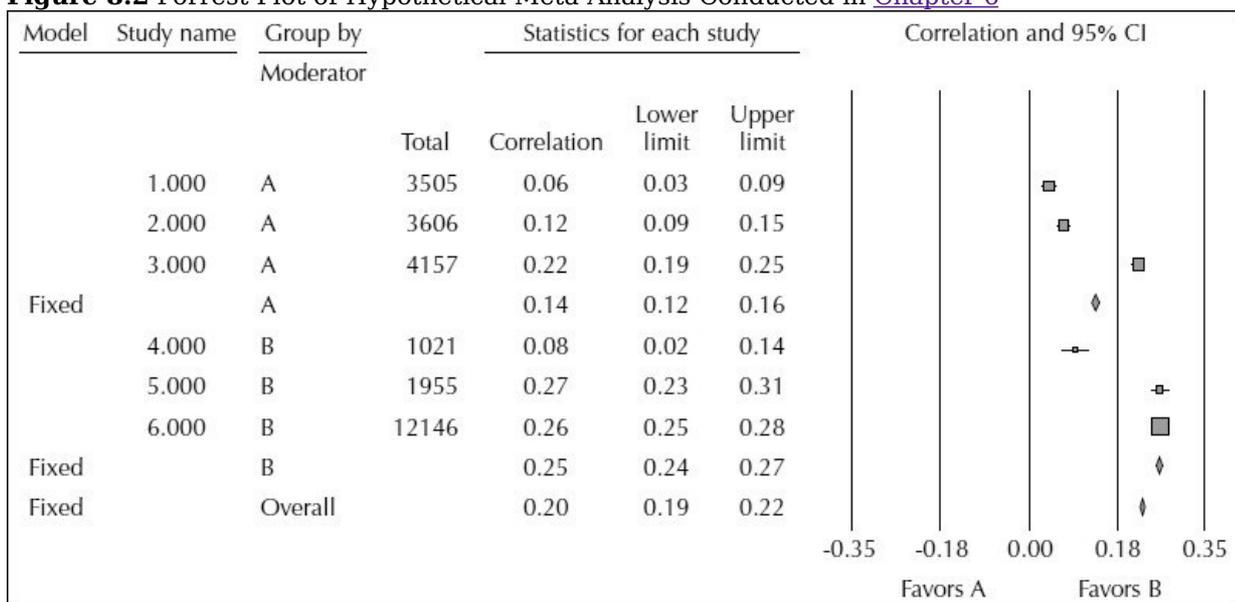
NOTE: This figure was generated using Comprehensive Meta-Analysis, Version 2.1 (Borenstein et al., 2005).

Figure 8.3 Distribution of Correlations Between Time on Homework and Achievement as a Function of Grade Level

Lower Grades	Stem	Upper Grades
5	+.6	
	+.3	00
6	+.2	998665
1	+.2	32200000
5	+.1	877
1	+.1	
	+.0	4
689	−.0	38
1	−.1	
5	−.2	3

SOURCE: Cooper, Robinson, and Patall (2006, p. 43). Copyright 2006 by the American Educational Research Association. Reprinted with permission.

NOTE: Lower grades represent grades 1 through 6. Upper grades represent grades 7 through 12 or samples that were described as middle or high school.

Our example meta-analyses on homework used a stem-and-leaf display, so I have reproduced it here in Figure 8.3. This is a somewhat more complex stem-and-leaf display. Here, we used this graphic to present the results of 33 studies that correlated the amount of homework students reported doing each night with a measure of their achievement. The stems are the first digit of the correlations and are presented in the middle column of the figure. The leaves are the second digit of each correlation. In the left side of the center column we have represented each of the 10 correlations we found that were calculated based on responses from children in elementary school, grades 1 through 6. On the right side of the center column, we represented the 23 correlations based on secondary school samples. So, with no loss in the precision of the information presented, this figure allows the reader to see the shape and dispersion of the 33 correlations and to note that the correlations are most often positive. But they can also visually detect a relationship between the magnitude of correlations and the grade level of students.

In general, then, the subsection that describes the aggregate results of the meta-analysis should give the reader a broad quantitative overview of the literature. This should complement the qualitative overviews contained in the introduction and method sections. It should provide the reader with a sense of the kinds of people, procedures, and circumstances contained in the studies. This subsection of results gives readers an opportunity to assess for themselves the representativeness of the sampled people and circumstances relative to the target populations. Also, it provides the broad overview of the findings regarding the main hypothesis under investigation.

Analyses of moderators of study results.

Another subsection should describe the results of analyses meant to uncover study characteristics that might have influenced their outcomes. For each moderator tested, the report should present results on whether the study characteristic was statistically significantly associated with variance in effect sizes. If the moderator proved significant, the report should present an average effect size and confidence interval for each grouping of studies. For example, we used a table to report the results from our search for moderators of the effects of choice on intrinsic motivation. This table is partially reproduced here as Table 8.7. Note that the number of findings differed slightly for each moderator variable we tested due to our use of a shifting unit of analysis.

Finally, the section describing moderator and mediator analyses should give readers some idea of the interrelationships among the different predictors of effect sizes. So, for example, in the report of our meta-analysis on the effects of choice on intrinsic motivation, we included a table that presented a matrix of the relationships between each pair of moderator variables. These interrelationships were used in the discussion of results to caution readers about possible confounds among our results.

In sum, the results section should contain your overall quantitative description of the covered literature, a

description of the overall findings regarding the hypotheses or relationships of primary interest, and the outcomes of the search for moderators and mediators of relationships. This lays the groundwork for the substantive discussion that follows.

Table 8.7 Table of Results of Moderator Analyses Examining the Effect of Choice on Intrinsic Motivation

Moderators	k	d	95% Confidence Interval		
			Low Estimate	High Estimate	
Publication type					14.98**(4.04)*
Published	28	.41**(.46)**	.33(.31)	.48(.60)	
Unpublished	18	.20**(.26)**	.13(.14)	.28(.38)	
Choice type					21.61**(5.63)
Choice of activities	11	.16**(.20)**	.06(.04)	.26(.35)	
Choice of versions	8	.27**(.26)**	.15(.06)	.38(.46)	
Instructionally irrelevant	8	.59**(.61)**	.43(.29)	.74(.94)	
Instructionally relevant	9	.24**(.33)**	.14(.14)	.34(.51)	
Choice of reward	3	.35**(.34)	.09(-.03)	.60(.71)	
Number of options per choice					5.62+(3.29)
Two	10	.20**(.19)**	.10(.05)	.29(.33)	
Three to five	13	.38**(.43)**	.26(.16)	.50(.69)	
More than five	18	.26**(.34)**	.18(.19)	.34(.49)	

Moderators	k	d	95% Confidence Interval		
			Low Estimate	High Estimate	
Number of choices (Analysis 1)					32.01**(11.15)**
One choice	21	.21**(.23)**	.14(.12)	.28(.33)	
Multiple choices	5	.18**(.25)	.04(-.02)	.31(.53)	
Successive choices	18	.54**(.58)**	.44(.40)	.64(.77)	
Number of choices (Analysis 2)					27.66**(10.28)**
One choice	21	.21**(.23)**	.14(.12)	.28(.33)	
Two to four choices	12	.61**(.63)**	.48(.38)	.75(.88)	
More than five choices	12	.32**(.45)**	.22(.23)	.43(.66)	
Reward					24.41**(12.16)**
No reward	40	.35**(.40)**	.29(.27)	.41(.52)	
Reward internal to choice manipulation	5	.35**(.36)**	.16(.08)	.54(.64)	
Reward external to choice manipulation	5	-.01(-.02)	-.15(-.22)	.12(.18)	

SOURCE: Adapted from "The effects of choice on intrinsic motivation and related outcomes: A meta-analysis of research findings," by E. A. Patall, H. Cooper, and J. C. Robinson, 2008, *Psychological Bulletin, 134,* 289. Copyright 2008 by the American Psychological Association.
NOTE: Random effects Q values and point estimates are presented in parentheses. +p <.10, *p <.05, **p <.01.

SOURCE: Adapted from "The effects of choice on intrinsic motivation and related outcomes: A meta-analysis of research findings," by E. A. Patall, H. Cooper, and J. C. Robinson, 2008, *Psychological Bulletin, 134,* 289. Copyright 2008 by the American Psychological Association.
NOTE: Random effects Q values and point estimates are presented in parentheses. +p <.10, *p <.05, **p <.01.

The Discussion Section

The discussion section of a research synthesis serves the same functions served by discussions in primary research. Discussions typically contain at least five components.

First, your discussion should present a summary of the major findings of the synthesis. This should not be too long and should focus primarily on the results you will spend time interpreting.

Second, you should interpret the major findings. The interpretation should describe the magnitude of the important effect sizes and their substantive meaning. This will involve examining the results in relation to the predictions you made in the introduction. Also, you need to examine the results for what they tell us about the theories and theoretical debates presented in the introduction. How to do this was the principal purpose of Chapter 7.

Third, your discussion should consider alternative explanations for your data. Typically, these will include, at a minimum, consideration of the possible impact of (a) missing data, (b) correlations among moderator variables, and (c) issues arising from methodological artifacts shared by the studies going into the meta-analysis.

Fourth, you will need to examine the generalizability of your findings. This will require you to consider (a) whether participants from all the relevant subpopulations have been included in the studies that make up your meta-analytic database, (b) whether important variations in independent or predictor variables and dependent or outcome variables are represented (or not) in the studies, and (c) the match between the research designs used in the individual studies and the inferences you wish to draw.

Finally, you should include a discussion of topics that need to be examined in future research. These should include new questions raised by the outcomes of the synthesis, and old questions left unresolved because of ambiguous synthesis results or a lack of prior primary research.

In general, then, the discussion section of a research synthesis report is used to make substantive interpretation of findings, to assess the generalizability of findings, to appraise whether past controversies have been resolved, and to suggest fruitful directions for future research.

I have rarely, if ever, seen a report of a research synthesis that included everything I have mentioned and everything listed in the MARS. Sometimes this is understandable; the relevance of the information is minimal given the nature of the literature being described. Other times, the omission is more concerning. It leaves the reader wondering how to interpret the results and, ultimately, whether the results are to be trusted. So, it is important to ask, when you consider the report of the results of a research synthesis,

Were the procedures and results of the research synthesis documented clearly and completely?

Exercises

Find a report of a meta-analysis that interests you. As you read it, check off the items in the MARS that are included in the report. What items are missing? Are they important to how you might interpret the findings? If so, how might their omission change your confidence in the conclusions of the synthesis?

Notes

1. In the interest of full disclosure, I should mention that I served as chair of this committee.

2. In the meta-analysis of individual differences in rape attitudes, a chart was presented that illustrated the number of studies of rape attitudes reported in each of consecutive years because we wanted to show how interest in the topic was growing.

3. Had this been an actual meta-analysis, I would have substituted the first author's last name and the year of the report for the study number.

4. The calculations in Figure 8.1 are based on r-indexes transformed to z-scores then transformed back to rs. So, the results differ very slightly from those in Table 6.4.

9 **Conclusion** Threats to the Validity of Research Synthesis Conclusions

> **This chapter describes**
>
> For each stage of a research synthesis
>
> - The general validity issues associated with the methodological choices made during that stage
> - The specific threats to validity
> - What synthesists can do to lessen the chances that the threats will be plausible alternative explanations to their conclusions
> - The cost and feasibility of conducting research syntheses
> - The value of disconfirmation in science
> - Creativity in the research synthesis process

In order to help you keep in mind the implications of decisions you make as you carry out your research synthesis, in this chapter I present some of the major threats to validity you will encounter at each stage of your project. I also summarize some of the practices you can implement to lessen the plausibility of these threats. Also, there are several issues related to research synthesis that involve more-general and more-philosophical considerations in applying the guidelines set forth in the previous chapters. I end the chapter, and the text, by briefly addressing these issues.

Validity Issues

Recall that Campbell and Stanley's (1963) list of validity threats to primary research was expanded and rearranged by Bracht and Glass (1968), Campbell (1969), Cook and Campbell (1979), and, most recently, by Shadish et al. (2002). This same expansion and rearranging of threats to validity has also occurred for research synthesis. In 1984 the first edition of this book (Cooper, 1984) suggested 11 threats to validity. Matt and Cook (1994) expanded this list to 21 threats; Shadish et al. expanded the list to 29 threats; and then Matt and Cook (2009) pared the list back to 28 threats. Not only does the list of threats expand and contract, but list providers also differ somewhat in their construal of the general class of validity (construct, internal, external, or statistical) that each specific threat might be related to. This is not a bad thing, but a good one. It serves to emphasize that what we are working with here is a dynamic theory of evidence. As such, it is okay for theorists to disagree. It is a sign of vitality and suggests the future will hold even more progress and refinement in thinking.

In Tables 9.1 through 9.7 I provide a summary of the validity issues associated with each step in research synthesis. At the top of each table are general statements about the threats to validity associated with that step. Next, I provide a list of more specific threats to validity taken from Matt and Cook (2009; also found in Shadish et al., 2002) that I have tried to align with the seven steps. Just as these previous list makers disagreed somewhat about the placement of these threats into different broader classes, I am certain that others will disagree with my classification (I moved several into different steps myself before settling on final resting places). Also, I have listed only 24 of the threats offered by previous list makers. I found that some of the threats appeared to be at least partially redundant in the concerns they covered.

Several of the threats to validity cataloged by Matt and Cook (1994, 2009) and by Shadish et al. (2002) that arise in the course of research synthesis are simply holdovers that represent pervasive problems in primary research. For instance, two threats to the validity of a synthesis' conclusions when data are being collected are that (a) the data from studies might not support conclusions about causal relations and (b) the people sampled in the covered studies might not be representative of the target population. This suggests that any threat associated with a particular primary research design is applicable to a synthesis' conclusions if the design characteristic appears in a substantial portion of the covered research. So, research designs should be examined carefully as potential moderators of study results. The creation of these *nomological nets* (Cronbach & Meehl, 1955) can be one of your synthesis' most valuable contributions. However, if an assortment of research designs (and participants, settings, and outcomes) is not contained in a synthesis, then threats associated with weaknesses in the dominant design features also threaten the synthesis' conclusions.

The last entries in Tables 9.1 through 9.7 summarize many of the good practices I mentioned in the previous chapters. Here, they are phrased in a way that shows how the practice will help protect your synthesis from the threats listed above them. You can use these tables along with Table 1.3, which lists the questions to ask about how a research synthesis was conducted, as summary guides to help you as you plan and execute your project.

Table 9.1 Questions Concerning the Validity of Research Synthesis Conclusions: Formulating the Problem

---

General validity issues:

1. Poorly defined (a) constructs (both abstractly and operationally) and (b) relationships between constructs can lead to ambiguity in and/or misapplication of results to circumstances to which they are not relevant.

Specific validity threats:

Unreliability (of treatments and/or measures) of primary studies

Explanation: If individual studies are poorly implemented, it will be difficult to accurately define their treatments and outcomes.

Underrepresentation of prototypical attributes

Explanation: Synthesists may define their concepts in ways that suggest greater generality than the operations actually used in studies would warrant.

Protecting validity:

1. Undertake your literature searches with the broadest possible conceptual definitions in mind. Begin 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, err toward making overly inclusive decisions, at least in the early stages of your project.

2. As the literature is being searched, reevaluate the fit between your conceptual definitions and the operations you are finding. Adjust the conceptual definitions accordingly, so they accurately reflect what operations have been used in studies.

3. To complement conceptual broadness, be thorough in your attention to distinctions in study characteristics. Any suggestion that a difference in study results is associated with a distinction in study characteristics should receive some testing, if only in a preliminary analysis.

SOURCE: Validity threats in italics are taken from Shadish et al. (2002), and Matt and Cook (2009).

SOURCE: Validity threats in italics are taken from Shadish et al. (2002), and Matt and Cook (2009).

Table 9.2 Questions Concerning the Validity of Research Synthesis Conclusions: Searching the Literature

---

General validity issues:

1. Studies found by the literature search might be different in methods and results from the entire population of studies and therefore might lead to inaccurate portrayals of the cumulative evidence.

Specific validity threats:

Publication bias

Explanation: If only published studies are used in the meta-analysis, the findings may overestimate the strength of the relationship.

Table 9.2 (Continued)

Protecting validity:

1. Conduct a broad and exhaustive search of the literature. A complete literature search has to include at least a search of reference databases, a perusal of relevant journals, an examination of references in related past primary research and research syntheses, a forward (citation) search, and contact with active and interested researchers. The more exhaustive a search, the more confident you can be that another synthesist using similar, but perhaps not identical, sources of information will reach the same conclusions.

2. Present indices of potential retrieval bias, if they are available. For instance, many research syntheses examine the potential effects of missing evidence and whether any difference exists in the results of studies that are published versus those that are unpublished.

SOURCE: Validity threat in italics is taken from Shadish et al. (2002), and Matt and Cook (2009).

Table 9.3 Questions Concerning the Validity of Research Synthesis Conclusions: Gathering Information From Studies

---

General validity issues:

1. Coders might incorrectly retrieve information from study reports that then misrepresent the studies in the cumulative analysis.

Specific validity threats:

Unreliability of coding in meta-analysis
Explanation: Unreliable coding can attenuate meta-analytic effect size estimates.

Rater drift
Explanation: Coders change their criteria for codes from one study to another (because of practice effects, fatigue, etc.).

Biased effect size sampling
Explanation: Only some of the plausibly relevant effect sizes are coded, and these favor one direction of findings.

Protecting validity:

1. Use coder training and coder evaluation procedures to minimize unreliable retrieval of information from studies.

2. Quantify intercoder agreement and continue training until an acceptable level of agreement is reached.

3. Have codes that lead to disagreement or low confidence discussed by multiple parties.

4. When possible, have more than one coder examine each study.

SOURCE: Validity threats in italics are taken from Shadish et al. (2002), and Matt and Cook (2009).

Table 9.4 Questions Concerning the Validity of Research Synthesis Conclusions: Evaluating the Quality of Studies

---

General validity issues:

1. Causal relationships are inferred when they are not supported by the evidence.

2. The use of nonquality factors to evaluate studies might result in the biased exclusion of studies or the improper weighting of studies in cumulative results.

Specific validity threats:

Absence of studies with successful random assignment
Primary study attrition

Explanation: These two threats to validity occur within each study, so there is little the research synthesists can do to address them. If they are present in all or most studies, they create the threat that the synthesist may infer causal relationships that are not supported by the evidence.

Reactivity effects

Explanation: Coders are influenced by the expectations of the principal investigators.

Table 9.4 (Continued)

Protecting validity:

1. Make every effort to ensure that only *a priori* conceptual and methodological judgments influence the decision to include or exclude studies from your synthesis, and not the results of the study.

2. If studies are to be weighted differently, your weighting scheme should be explicit and justifiable.

3. The approach used to categorize study methods should exhaust as many design characteristics as possible. Detail each design distinction that was related to study results and describe the outcome of the analysis.

Table 9.5 Questions Concerning the Validity of Research Synthesis Conclusions: Analyzing and Integrating the Outcomes of Studies

---

General validity issues:

1. Rules for summarizing and integrating data from the individual studies might be inappropriate and lead to incorrect cumulative results.

Specific validity threats:

Capitalization on chance

Explanation: Meta-analysts can test many relationships. If they do not adjust significance levels accordingly, they can inflate the likelihood that chance findings will appear statistically significant.

Lack of statistical independence among effect sizes

Explanation: If meta-analysts treat nonindependent effect sizes as if they were independent, they will overestimate the precision and power of their analyses.

Failure to weight study-level effect sizes proportional to their precision

Explanation: Meta-analysts who do not weight effect sizes (by the inverse of their sampling error) introduce imprecision into their estimates of average effect sizes.

Inaccurate (low power) homogeneity tests

Explanation: Meta-analysts may conduct homogeneity tests that suffer from low statistical power.

Underjustified use of fixed effect models

Explanation: Meta-analysts may use a fixed effect model when the heterogeneity in effect sizes suggests a random-effects model is more appropriate.

Protecting validity:

Recommendations concerning what assumptions are appropriate for synthesists to make about their data will depend on the data in a particular problem area and the purposes of a synthesis. Therefore,

1. Be as explicit as possible about the assumptions that guided your analyses when you convey your conclusions and inferences to readers.

2. Your decision about the proper unit of analysis for your synthesis should be based both on statistical considerations and the nature of the particular problem under study. The approach you choose should be carefully described and justified.

3. If there is any evidence bearing on the validity of your interpretation rules, it should be presented.

SOURCE: Validity threats in italics are taken from Shadish et al. (2002), and Matt and Cook (2009).

SOURCE: Validity threats in italics are taken from Shadish et al. (2002), and Matt and Cook (2009).

Table 9.6 Questions Concerning the Validity of Research Synthesis
Conclusions: Interpreting the Evidence

General validity issues:

1. Cumulative results might be different if different statistical assumptions are used.
2. Missing data might cause results to be different than if all data were available.
3. Synthesis-generated evidence might be used to infer that moderator variables have a causal effect when a causal inference is not warranted.
4. The generality, magnitude, and/or importance of cumulative findings might be misrepresented.

Table 9.6 (Continued)

Specific validity threats:

Missing effect sizes in primary studies

Bias in computing effect sizes (that must be estimated from other statistics)

Explanation: When data are missing, meta-analysts must omit or approximate effect sizes, and these approximation procedures can vary in accuracy.

Restricted range in primary studies

Explanation: If there is a restricted range of outcomes in the primary studies, these will attenuate the effect size estimates when they are combined in meta-analysis.

Moderator variable confounding

Explanation: Meta-analysts make claims about the causal relationship between a moderator and effect sizes when the moderator is correlated with other moderators.

Sampling biases associated with the persons, settings, treatments, outcomes, and times entering a meta-analysis

Restricted heterogeneity of substantively irrelevant third variables

Explanation: The above two threats occur because meta-analysts may overgeneralize to persons, settings, treatments, outcomes, and times not contained in the study samples.

Failure to test for heterogeneity in effect sizes

Explanation: Meta-analysts may not test moderator variables that account for systematic variation in effect sizes.

Lack of statistical power for homogeneity tests

Lack of statistical power for studying disaggregated groups

Explanation: When meta-analysts test for differences in effect sizes (a) averaged across subgroups of studies or (b) within subgroups of studies, they may not have sufficient statistical power to uncover important findings.

Restricted heterogeneity of irrelevancies

Explanation: A threat to the generality of the findings of a meta-analysis occurs when there is not enough variation in study attributes that are irrelevant to the relationship of interest, meaning the meta-analysts cannot test whether the effect holds across many situational variations.

Protecting validity:

1. State explicitly what conventions you used when incomplete or erroneous research reports were encountered.

2. Whenever possible, analyze your data using multiple procedures that require different assumptions. (Much greater confidence can be placed in results that do not vary across different analyses based on different assumptions.)

3. Summarize the sample characteristics of individuals used in the separate studies. Note important missing samples that may restrict generality.

SOURCE: Validity threats in italics are taken from Shadish et al. (2002), and Matt and Cook (2009).

Table 9.7 Questions Concerning the Validity of Research Synthesis Conclusions: Presenting the Results

General validity issues:

1. Omission of synthesis procedures might make conclusions hard to evaluate with regard to plausible threats to validity, and difficult to reproduce.

Specific validity threats:

None

Protecting validity:

1. When writing your report, employ the Meta-Analysis Reporting Standards (MARS) presented in Table 8.1 and the search strategies presented in Table 8.2.

SOURCE: Validity threats in italics are taken from Shadish et al. (2002), and Matt and Cook (2009).

Criticism of Research Synthesis and Meta-Analysis

Another way to summarize the issues I have covered in this text is to look at the criticisms that have been leveled against research synthesis and meta-analysis and see how we might respond to them. Four recent texts have compiled lists of such criticisms (Borenstein et al., 2009; Card, 2012; Littell, Corcoran, & Pillai, 2008; Petticrew & Roberts, 2006) and answered them. In Table 9.8, I have integrated these lists, categorized them by the stage of the synthesis process they pertain to, and provided my own responses (similar to those you will find in the above references).

There are two general and important questions to ask when a criticism is leveled against research synthesis and meta-analysis. First, "Could this criticism be leveled against research synthesis regardless of whether the new standards of evidence were being employed or a more traditional form of research review was conducted?" You will find that in many instances the criticism really pertains to research reviews in general, warranted or not. Perhaps the major difference is that when scientific standards are applied to the conduct of a research synthesis the weaknesses in evidence become more transparent; simply because weaknesses are less evident in traditional reviews does not mean they are not there.

The second question to ask is, "Does this criticism relate to the methods themselves or to the way a particular research synthesis was conducted?" You will find that in many instances critics point to weaknesses in syntheses they have seen published but were flawed in their execution. This is bound to happen, and is almost inevitable. It does not mean the methods are flawed; the methods provide the yardstick for assessing the trustworthiness of the synthesis. It is the execution that leaves something to be desired. A similar assessment could be made of every primary study that has ever been conducted. There is always room for improvement.

Feasibility and Cost

It is considerably more expensive, in terms of both time and money, for synthesists to undertake a project using the guidelines set forth in this book than to conduct syntheses in a less rigorous manner. More people are involved who need to be compensated for their time. More time and resources are needed to search the literature, develop coding frames, run analyses, and prepare reports.

Given these costs, should a potential synthesist with limited resources be discouraged from undertaking such a project? Certainly not; just as the perfect, irrefutable primary study has never been conducted, so, too, the perfect synthesis remains an ideal. My guidelines represent more a yardstick for evaluating syntheses than a set of absolute requirements. In fact, you should be aware of several instances in which the syntheses I used as examples fell short of complete adherence to the guidelines. You should not hold the guidelines as absolute criteria that must be met but rather as targets that help you refine procedures until you strike a good balance between rigor and feasibility. It is critical, though, that you acknowledge in your report where the weaknesses in your synthesis exist.

Table 9.8 Common Criticisms of Research Synthesis and Meta-Analysis and Responses to Them

Formulating the Problem

- *Meta-analysts mix together studies that are not related; this is the apples-and-oranges problem.*

  It is true that meta-analytic methods allow the study of broad concepts, such as fruit, that might not be amenable to synthesis without the aid of statistical combining procedures. However, meta-analysis can be applied to narrow areas of research (e.g., randomized experiments on aerobic exercise effects on neurocognitive functioning) as well as broad areas (e.g., any research designs looking at any of the effects of homework). Whether a meta-analysis might be too broad or too narrow is a function of the questions a field needs answered and the capability of the existing research to answer them, and not the method used for research synthesis.

Searching the Literature

- *Research syntheses cannot be done without the assistance of experienced information scientists.*

  Not true. A good search can be conducted without an advanced library degree. Still, it is always a good idea to involve people with searching experience in the area under study to ensure all the potential sources of studies are included and there is no bias in the selection of sources, whether or not meta-analysis is used. This is an issue of execution, not inherent to the methods themselves.

- *Studies with nonsignificant results are hard to find; this is the file drawer problem.*

  True, but a comprehensive search can use methods (e.g., prospective registers, direct contact with researchers) to mitigate the problem. Also, the impact of publication bias can be assessed and estimated (with reasonable assumptions) using the literature that is located.

- *Study eligibility decisions in research syntheses can be biased.*

  As long as the eligibility criteria are transparent and fairly applied, this should not be a problem. Disagreements can always occur about what should and should not be included in a synthesis. Sensitivity analyses can be used to assess whether inclusion criteria matter.

Table 9.8 (Continued)

- *Studies conducted in English are easier to locate; this is the dissemination bias problem.*

  Synthesists can search non-English sources for studies and have them translated. This is an issue of execution, and is not inherent to the methods themselves.

Gathering Information From Studies

- *Coding of information from studies needs to be done reliably but study reports are often ambiguous and are missing needed information.*

  True, reports of primary studies are often ambiguous and do not include the information a research synthesist wants. This is an issue no matter what synthesis procedures are used. The reliability of codes can be improved by using established techniques for sound content analysis. Synthesists can use low-inference codes to improve reliability. Estimation of missing data can be done using the same techniques used in primary research.

Evaluating the Quality of Studies

- *Meta-analyses often include inadequate assessment of study quality; this is the "garbage in, garbage out" problem.*

  This is an issue of execution, and is not inherent to the methods themselves. Syntheses can be restricted to include only high-quality studies or can use variations in the methods of studies to examine their effects on outcomes.

- *Primary research reports often lack transparency about their methods.*

  True (see above). But this is a problem in how primary studies are reported, and not in the methods of research synthesis. Numerous efforts are attempting to improve the reporting of primary studies.

Analyzing and Integrating the Outcomes of Studies

- *The amount of expertise needed to understand a meta-analysis makes it hard for many synthesists to do.*

  Training in conducting meta-analysis has improved greatly. It is no longer a mysterious process known to only a few.

- *Meta-analyses are often performed poorly.*

  This is an issue of execution, not inherent to the methods themselves. This criticism can be leveled against primary studies as well as research syntheses.

Interpreting the Evidence

- *One number, an effect size, cannot summarize a research field.*

  True, but it is very rare for a meta-analysis to result in a single number. Effect sizes are reported using different assumptions about the underlying data. Reports include multiple effect sizes for different groupings of studies, based on both methodological and conceptual variations in studies.

- *Meta-analyses may overlook the nuances in a literature captured by qualitative review.*

  Again, this is an issue of execution, and is not inherent to the methods themselves. Doing a meta-analysis does not preclude the examination of individual studies to highlight their uniqueness. Important studies can, and should, be examined separately, as well as included in the meta-analysis. These studies can be highlighted for their unique contributions. Qualitative studies and studies with unique methodological or conceptual variations can be used to flesh out the quantitative results.

- *Meta-analysis results can disagree with randomized trials; this is the gold standard of evidence.*

  What does disagreement mean? Often these arguments point to differences in statistical significance that might be explained by differences in statistical power (e.g., differences in sample size or design). This is a wrong definition of the term *inconsistency* because the underlying effects can be very similar. Randomized studies can be included in meta-analyses (see next argument) and the effects of different methods can be tested for their effect on outcomes.

- *Research syntheses include only randomized controlled trials.*

  Not true. This criticism is the opposite of the criticism above. It is the nature of the question and the type of studies that have been conducted that dictate what evidence should be included in a research synthesis.

- *Research syntheses are no substitute for doing high-quality individual studies.*

  True. Research syntheses are a complement to high-quality primary studies. Often, it is the research synthesis of past studies that makes clear what the most valuable future primary studies should look like.

- *Meta-analyses rely too heavily on effect sizes, a crude measure.*

  Effect sizes are a better measure of a study's outcomes than statistical significance. There is no more precise measure.

Table 9.8 (Continued)

General

- *Research syntheses are the same as ordinary reviews, only bigger.*

  True, if research syntheses are prepared in accord with scientific guidelines and "ordinary" reviews are not. This is a good character of rigorous research syntheses. If more exhaustive (and representative) databases are used in research syntheses, more primary studies should be found. This should reduce bias, result in more precise estimates of effect, and test the effects of moderators more formally.

- *Meta-analysis requires the adoption of a biomedical model of health.*

  This impression is caused by the widespread use of meta-analyses in biomedical fields, not by anything inherent in the methods. Meta-analyses first appeared in the social sciences and are used frequently therein. There is no preferred theoretical model underlying the use of meta-analytic techniques.

- *Research syntheses are of no relevance to the real world; they have poor external validity across populations, treatment and outcome variation, and time.*

  No more so than any primary study. In fact, research syntheses can have greater external validity than any single primary study. If a study is used in a research synthesis along with other studies, how can a synthesis have less external validity than its constituent studies? It can only have more. That said, all research, whether primary or research synthesis, needs to be carefully evaluated for all threats to validity.

- *Newer meta-analytic techniques have not yet been validated.*

  True, but this shows that the development of research synthesis methods is dynamic and improving. Similar to all methodological procedures in the social and behavioral sciences, the fundamental methods are well established and newer approaches must remain under scrutiny until they are understood clearly.

The Scientific Method and Disconfirmation

While the practical aspects of conducting research syntheses may mean the investigator must settle for a less-than-perfect product, the ideals of science still must be strictly applied to the research synthesis process. The most crucial scientific element missing from haphazard synthesis procedures is the potential for the disconfirmation of the synthesist's prior beliefs. In most instances, primary researchers undertake their work with some recognition that the results of their study may alter their beliefs. By extending the scientific method to research syntheses, we also expand the potential for disconfirmation. Ross and Lepper (1980) have stated this position nicely:

> We know all too well that the scientific method is not immune to the diseases of biased assimilation, causal explanation, and a host of other nagging afflictions; scientists can be blind, sometimes deliberately so, to unanticipated or uncongenial interpretations of their data and recalcitrant in their theoretical allegiances. . . . Nevertheless, it is the scientific method . . . that has often been responsible for increasing human understanding of the natural and social world. Despite its flaws, it remains the best means of delivering us from the errors of intuitive beliefs and intuitive methods for testing those beliefs. (p. 33)

Creativity in Research Synthesis

Early in this text I mentioned that one objection to the use of scientific guidelines for research synthesis is that this system stifles creativity. Critics who raise this issue think the rules for conducting and reporting primary research are a straitjacket on innovative thinking. I cannot disagree more. Rigorous criteria will not produce syntheses that are mechanical and uncreative. Your expertise and intuition will be challenged to capitalize on or create opportunities to obtain, evaluate, and analyze information unique to your problem area. I hope the syntheses examples have demonstrated the diversity and complexity of issues that confront those who adopt the scientific method. These challenges are created, not solved, by

the rules of science.

Conclusion

I began this book with the supposition that research synthesis was a data-gathering exercise that needed to be evaluated against scientific criteria. Because of the growth in empirical research and the increased access to information, the conclusions of research syntheses will become less trustworthy unless we systematize the process and make it more rigorous and unbiased. I hope that the concepts and techniques presented here have convinced you that it is feasible and desirable for social scientists to require rigorous syntheses. Such rules bring with them greater potential for creating consensus among scholars and for focusing discussion on specific and testable areas of disagreement when conflict does exist. Because of the increasing role that research syntheses play in our definition of knowledge, these adjustments in procedures are inevitable if social scientists hope to retain their claim to objectivity as well as their credibility with those who turn to scientists to help solve social problems and increase our understanding of the world.

References

American Psychological Association (APA). (2010). Publication manual of the American Psychological Association (6th ed.). Washington, DC: Author.

American Psychological Association (APA). (2015). PsycINFO. Retrieved from www.apa.org/pubs/databases/psychinfo/index.aspx

American Psychological Association's Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. American Psychologist, 61, 271–283.

Anderson, K. B., Cooper, H., & Okamura, L. (1997). Individual differences and attitudes toward rape: A meta-analytic review. Personality and Social Psychology Bulletin, 23, 295–315.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? American Psychologist, 63, 839–851.

Arthur, W., Jr., Bennett, W., Jr., & Huffcutt, A. I. (2001). Conducting meta-analysis using SAS. Mahwah, NJ: Lawrence Erlbaum.

Atkinson, D. R., Furlong, M. J., & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship? Journal of Counseling Psychology, 29, 189–194.

Atkinson, K. M., Koenka, A. C., Sanchez, C. E., Moshontz, H., & Cooper, H. (2015). Reporting standards for literature searches and report inclusion criteria: Making research syntheses more transparent and easy to replicate. Research Synthesis Methodology, 6, 87–95.

Atkinson, K. M., Sanchez, C. E., Koenka, A. C., Moshontz, H., & Cooper, H. (2014). Who makes the grades?: A synthesis of research comparing student, peer and instructor grades in college classrooms. Manuscript under review.

Barber, T. X. (1978). Expecting expectancy effects: Biased data analyses and failure to exclude alternative interpretations in experimenter expectancy research. The Behavioral and Brain Sciences, 3, 388–390.

Barnett, V., & Lewis, T. (1984). Outliers in statistical data (2nd ed.). New York: John Wiley & Sons.

Becker, B. J. (2005, November). Synthesizing slopes in meta-analysis. Paper presented at the meeting on Research Synthesis and Meta-Analysis: State of the Art and Future Directions, Durham, NC.

Becker, B. J. (2009). Model-based meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), The handbook of research synthesis and meta-analysis (2nd ed., pp. 377–395). New York: Russell Sage.

Becker, B. J., & Wu, M. (2007). The synthesis of regression slopes in meta-analysis. Statistical Science, 22, 414–429.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. Biometrics, 50, 1088–1101.

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. Psychological Review, 74, 183–200.

Berlin, J. A., & Ghersi, D. (2005). Preventing publication bias: Registers and prospective meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), Publication bias in meta-analysis: Prevention, assessment and adjustments (pp. 35–48). West Sussex, UK: John Wiley & Sons.

Bohning, D., Kuhnert, R., & Rattanasiri, S. (2008). Meta-analysis of binary data using profile likelihood. Boca Raton, FL: Taylor & Francis.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2005). Comprehensive meta-analysis (Ver. 2.1) [Computer software]. Englewood, NJ: Biostat.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Introduction to meta-analysis. West Sussex, UK: John Wiley & Sons.

Borman, G. D., & Grigg, J. A. (2009). The visual and narrative interpretation of research synthesis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), The handbook of research synthesis and meta-analysis (2nd ed., pp. 497–519). New York: Russell Sage.

Bourque, L. B., & Clark, V. A. (1992). Processing data. Newbury Park, CA: Sage.

Bracht, G. H., & Glass, G. V. (1968). The external validity of experiments. American Educational Research Journal, 5, 437–474.

Brunton, J., & Thomas, J. (2012). Information management in reviews. In D. Gough, S. Oliver, & J. Thomas (Eds.), An introduction to systematic reviews (pp. 83–106). Thousand Oaks, CA: Sage.

Bushman, B. J., & Wang, M. C. (2009). Vote counting procedures in meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), The handbook of research synthesis and meta-analysis (2nd ed., pp. 207–220). New York: Russell Sage.

Campbell Collaboration. (2015). What helps? What harms? Based on what evidence? Retrieved from www.campbellcollaboration.org/

Campbell, D. T. (1969). Reforms as experiments. American Psychologist, 24, 409–429.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. Chicago: Rand McNally.

Card, N. A. (2012). Applied meta-analysis for social science research. New York: Guilford Press.

Carlson, M., & Miller, N. (1987). Explanation of the relation between negative mood and helping. Psychological Bulletin, 102, 91–108.

Chalmers, I., Hedges, L. V., & Cooper, H. (2002). A brief history of research synthesis. Evaluation & the Health Professions, 25, 12–37.

Chen, D.-G., & Peace, K. E. (2013). Applied meta-analysis with R. Boca Raton, FL: Taylor & Francis.

Cheung, M. W.-L. (2015). Meta-analysis: A structural equation modelling approach. Chichester, UK: John Wiley & Sons.

Christensen, L. (2012). Types of designs using random assignment. In H. Cooper (Ed.), APA handbook of research methods in psychology (pp. 469–488). Washington, DC: American Psychological Association.

Coalition for Evidence-Based Policy. (2015). Coalition for evidence-based policy. Retrieved from http://coalition4evidence.org/

Cochrane Collaboration. (2015). Reliable source of evidence in health care. Retrieved from http://www.cochrane.org/

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). New York: Academic Press.

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

Comprehensive Meta-Analysis. (2015). Comprehensive meta-analysis. Retrieved from http://www.meta-analysis.com/index.php

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation. Chicago, IL: Rand McNally.

Cook, T. D., Cooper, H., Cordray, D. S., Hartmann, H., Hedges, L. V., & Light, R. J. (1992). Meta-analysis for explanation: A casebook. New York, NY: Russell Sage.

Cooper, H. (1982). Scientific guidelines for conducting integrative research reviews. Review of Educational Research, 52, 291–302.

Cooper, H. (1984). The integrative research review: A systematic approach. Beverly Hills, CA: Sage.

Cooper, H. (1986). On the social psychology of using research reviews: The case of desegregation and black achievement. In R. Feldman (Ed.), The social psychology of education (pp. 341–364). Cambridge, UK: Cambridge University Press.

Cooper, H. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews. Knowledge in Society, 1, 104–126

Cooper, H. (1989). Homework. New York, NY: Longman.

Cooper, H. (2006). Research questions and research designs. In P. A. Alexander, P. H. Winne, & G. Phye (Eds.), Handbook of research in educational psychology (2nd ed., pp. 849–877). Mahwah, NJ: Lawrence Erlbaum.

Cooper, H. (2007). Evaluating and interpreting research syntheses in adult learning and literacy. Boston, MA: National College Transition Network, New England Literacy Resource Center/World Education.

Cooper, H. (2009). The search for meaningful ways to express the effects of interventions. Child Development Perspectives, 2, 181–186.

Cooper, H., Charlton, K., Valentine, J. C., & Muhlenbruck, L. (2000). Making the most of summer school.

Malden, MA: Blackwell Publishing.

Cooper, H., DeNeve, K., & Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. Psychological Methods, 2, 447–452.

Cooper, H., & Hedges, L.V. (Eds.). (1994). The handbook of research synthesis. New York, NY: Russell Sage.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). The handbook of research synthesis and meta-analysis (2nd ed.). New York, NY: Russell Sage.

Cooper, H., Jackson, K., Nye, B., & Lindsay, J. J. (2001). A model of homework's influence on the performance evaluations of elementary school students. Journal of Experimental Education, 69, 181–202.

Cooper, H., & Koenka, A. C. (2012). The overview of reviews: Unique challenges and opportunities when research syntheses are the principal elements of new integrative scholarship. American Psychologist, 67, 446–462.

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis using individual participant data and aggregate data. Psychological Method, 14, 165–176.

Cooper, H. M., Patall, E. A., & Lindsay, J. J. (2009). Research synthesis and meta-analysis. In L. Bickman & D. Rog (Eds.), Applied social research methods handbook (2nd ed., pp. 344–370). Thousand Oaks, CA: Sage.

Cooper, H., & Ribble, R. G. (1989). Influences on the outcome of literature searches for integrative research reviews. Knowledge: Creation, Diffusion, Utilization, 10, 179–201.

Cooper, H., Robinson, J. C., & Patall, E. A. (2006). Does homework improve academic achievement? A synthesis of research, 1987–2003. Review of Educational Research, 76, 1–62.

Cooper, H., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. Psychological Bulletin, 87, 442–449.

Coursol, A., & Wagner, E. E. (1985). Effects of positive findings on submission and acceptance rates: A note on meta-analysis bias. Professional Psychology, 17, 136–137.

Crane, D. (1969). Social structure in a group of scientists: A test of the "invisible college" hypothesis. American Sociological Review, 34, 335–352.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281–302.

Cuadra, C. A., & Katter, R. V. (1967). Opening the black box of relevance. Journal of Documentation, 23, 291–303.

Cumming, G. (2012). Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis. New York, NY: Routledge.

Davidson, D. (1977). The effects of individual differences of cognitive style on judgments of document

relevance. Journal of the American Society for Information Science, 8, 273–284.

Deci, E. L., & Ryan, R. M. (2013). The handbook of self-determination research. Rochester, NY: University of Rochester Press.

Deeks, J. J., Higgins, J. P. T., & Altman, D. G. (2008). Analysing data and undertaking meta-analyses. In J. T. P. Higgins & S. Green (Eds.), Cochrane handbook for systematic reviews of intervention (pp. 243–296). Chichester, UK: John Wiley & Sons.

Dickerson, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), Publication bias in meta-analysis: Prevention, assessment and adjustments (pp. 11–33). West Sussex, UK: John Wiley & Sons.

Duval, S. (2005). The trim-and-fill method. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), Publication bias in meta-analysis: Prevention, assessment and adjustments (pp. 127–144). Chichester, UK: John Wiley & Sons.

Duval, S., & Tweedie, R. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. Journal of the American Statistical Association, 95, 89–98.

Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics, 56, 276–284.

Eddy, D. M., Hasselblad, V., & Schachter, R. (1992). Meta-analysis by the confidence profile approach. Boston, MA: Academic Press.

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. British Medical Journal, 315, 629–634.

Eid, M., & Diener, E. s. (Ed). (2006). Handbook of multimethod measurement in psychology. Washington, DC: American Psychological Association.

Ellis, P. (2009). Effect size calculator. Retrieved from http://www.polyu.edu.hk/mm/effectsizefaqs/calculator/calculator.html

Equator Network. (2015). Enhancing the quality and transparency of health research. Retrieved from http://www.equator-network.org/

Eysenck, H. J. (1978). An exercise in mega-silliness. American Psychologist, 33, 517.

Feldman, K. A. (1971). Using the work of others: Some observations on reviewing and integrating. Sociology of Education, 4, 86–102.

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. Journal of Abnormal and Social Psychology, 58, 203–210.

Fisher, R. A. (1932). Statistical methods for research workers. London: Oliver & Boyd.

Fiske, D. W., & Fogg, L. (1990). But the reviewers are making different criticisms of my paper! American Psychologist, 45, 591–598.

Fleiss, J. L., & Berlin, J. A. (2009). Measures of effect size for categorical data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), The handbook of research synthesis and meta-analysis (2nd ed., pp. 237–253). New York: Russell Sage.

Fowler, F. J. (2014). Survey research methods (5th ed.). Thousand Oaks, CA: Sage.

Gale Directory Library. (n.d.). Retrieved from http://www.gale.cengage.com/DirectoryLibrary/

Garvey, W. D., & Griffith, B. C. (1971). Scientific communication: Its role in the conduct of research and creation of knowledge. American Psychologist, 26, 349–361.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5, 3–8.

Glass, G. V. (1977). Integrating findings: The meta-analysis of research. In L. S. Shulmar (Ed.), Review of Research in Education (Vol. 5, pp. 35–79). Itasca, IL: Peacock.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage.

Glass, G. V., & Smith, M. L. (1978). Reply to Eysenck. American Psychologist, 33, 517–518.

Glass, G. V., & Smith, M. L. (1979). Meta-analysis of research on the relationship of class size and achievement. Educational Evaluation and Policy Analysis, 1, 2–16.

Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), The handbook of research synthesis and meta-analysis (2nd ed., pp. 357–376). New York: Russell Sage.

Gottfredson, S. D. (1978). Evaluating psychological research reports. American Psychologist, 33, 920–934.

Greenberg, J., & Folger, R. (1988). Controversial issues in social research methods. New York: Springer-Verlag.

Greenwald, A. G. (1975). Consequences of prejudices against the null hypothesis. Psychological Bulletin, 82, 1–20.

Greenwald, R., Hedges, L. V., & Laine, R. (1996). The effects of school resources on student achievement. Review of Educational Research, 66, 411–416.

Grubbs, F. E. (1950). Sample criteria for testing outlying observations. Journal of the American Statistical Association, 21, 27–58.

Harris, M. J., & Rosenthal, R. (1985). Mediation of interpersonal expectancy effects: 31 meta-analyses. Psychological Bulletin, 97, 363–386.

Hartung, J., Knapp, G., & Sinha, B. K. (2008). Statistical meta-analysis with applications. Hoboken, NJ: John Wiley & Sons.

Hattie, J. (2008). Visible learning, A synthesis of over 800 meta-analyses relating to achievement. New York, NY: Routledge.

Hedges, L. V. (1980). Unbiased estimation of effect size. Evaluation in Education: An International Review Series, 4, 25–27.

Hedges, L. V. (1982). Fitting categorical models to effect sizes from a series of experiments. Journal of Educational Statistics, 7, 119–137.

Hedges, L. V. (1994). Fixed effects models. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 286–299). New York: Russell Sage.

Hedges, L. V., & Olkin, I. (1980). Vote-counting methods in research synthesis. Psychological Bulletin, 88, 359–369.

Hedges, L., & Olkin, I. (1985). Statistical methods for meta-analysis. Orlando, FL: Academic Press.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. Research Synthesis Methods, 1, 39–65.

Hedges, L. V., & Vevea, J. L. (1998). Fixed and random effects models in meta-analysis. Psychological Methods, 3, 486–504.

Hunt, M. (1997). How science takes stock: The story of meta-analysis. New York, NY: Russell Sage.

Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. Psychological Bulletin, 86, 721–735.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). Meta-analysis: Cumulating research findings across studies. Beverly Hills, CA: Sage.

Jackson, G. B. (1980). Methods for integrative reviews. Review of Educational Research, 50, 438–460.

Johnson, B. T., & Eagly, A. H. (2000). Quantitative synthesis of social psychological research. In H. T. Reis & C. M. Judd (Eds.), Handbook of research methods in social and personality psychology (pp. 496–528). New York: Cambridge University Press.

Jonas, D. E., Wilkins, T. M., Bangdiwala, S., Bann, C. M., Morgan, L. C., Thaler, K. J., . . . , & Gartlehner, G. (2013). Findings of Bayesian mixed treatment comparison meta-analyses: Comparison and exploration using real-world trial data and simulation. (Prepared by RTI-UNC Evidence-based Practice Center under Contract No. 290-2007-10056-I). AHRQ Publication No. 13-EHC039-EF. Rockville, MD: Agency for Healthcare Research and Quality.

Jüni, P., Witshci, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. Journal of the American Medical Association, 282, 1054–1060.

Justice, A. C., Berlin, J. A., Fletcher, S. W., & Fletcher, R. A. (1994). Do readers and peer reviewers agree on manuscript quality? Journal of the American Medical Association, 272, 117–119.

Kane, T. J. (2004). The impact of after-school programs: Interpreting the results of four recent evaluations. New York, NY: William T. Grant Foundation.

Kazdin, A., Durac, J., & Agteros, T. (1979). Meta–meta analysis: A new method for evaluating therapy outcome. Behavioral Research and Therapy, 17, 397–399.

Kline, R. B. (2011). Principles and practices of structural equation modelling (3rd ed.). New York, NY: Guilford Press.

Kromrey, J. D., & Rendina-Gobioff, G. (2006). On knowing what we do not know: An empirical comparison of methods to detect publication bias in meta-analysis. Educational and Psychological Measurement, 66, 357–373.

Leong, F. T. L., & Austin, J. T. (Eds.). (2006). The psychology research handbook: A guide for graduate students and research assistants. Thousand Oaks, CA: Sage.

Levin, H. M. (2002). Cost-effectiveness and educational policy. Larchmont, NY: Eye on Education.

Light, R. J., & Pillemer, D. B. (1984). Summing up: The science of reviewing research. Cambridge, MA: Harvard University Press.

Light, R. J., & Smith, P. V. (1971). Accumulating evidence: Procedures for resolving contradictions among research studies. Harvard Educational Review, 41, 429–471.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. American Psychologist, 48, 1181–1209.

Lipsey, M. W., & Wilson, D. B. (2001). Practical meta-analysis. Thousand Oaks, CA: Sage.

Littell, J. H., Corcoran, J., & Pillai, V. (2008). Systematic reviews and meta-analysis. Oxford, UK: Oxford University Press.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. Journal of Personality and Social Psychology, 37, 2098–2109.

Macaskill, P., Walter, S., & Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. Statistics in Medicine, 20, 641–654.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. Cognitive Therapy and Research, 1, 161–175.

Mansfield, R. S., & Bussey, T. V. (1977). Meta-analysis of research: A rejoinder to Glass. Educational Researcher, 6, 3.

Marsh, H. W., & Ball, S. (1989). The peer review process used to evaluate manuscripts submitted to academic journals: Interjudgmental reliability. Journal of Experimental Education, 57, 151–170.

Matt, G. E., & Cook, T. D. (1994). Threats to the validity of research syntheses. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 503–520). New York, NY: Russell Sage.

Matt, G. E., & Cook, T. D. (2009). Threats to the validity of generalized inferences from research syntheses. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.). The handbook of research synthesis and meta-analysis (2nd ed., pp. 537–560). New York, NY: Russell Sage.

May, H. (2012). Nonequivalent comparison group designs. In H. Cooper (Ed.), APA handbook of research methods in psychology (pp. 489–510). Washington, DC: American Psychological Association.

McGrath, J. B. (1993). Student and parental homework practices and the effects of English homework on student test scores (Doctoral dissertation, United States International University). Dissertation Abstracts International, 53, 3490.

McPadden, K., & Rothstein, H. R. (2006, August). Finding the missing papers: The fate of best paper proceedings. Paper presented at AOM Conferences, Academy of Management Annual Meeting, Atlanta, GA.

Merriam-Webster. (2015a). Significant. Retrieved from http://www.merriamwebster.com/dictionary/significant

Merriam-Webster. (2015b). Promising. Retrieved from http://www.merriamwebster.com/dictionary/promising

Miller, N., Lee, J. Y., & Carlson, M. (1991). The validity of inferential judgments when used in theory-testing meta-analysis. Personality and Social Psychology Bulletin, 17, 335–343.

Moher, D., Tetzlaff, J., Liberati, A., Altman, D. G., & the PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analysis: The PRISMA statement. Annals of Internal Medicine, 151, 264–269.

Neyeloff, J. L., Fuchs, S. C., & Moreira, L. B. (2012). Meta-analyses and Forest plots using a Microsoft Excel spreadsheet: Step-by-step guide focusing on descriptive data analysis. BMC Research Notes, 5, 1–6.

Nickerson, R. S. (1998). Confirmatory bias: A ubiquitous phenomenon in many guises. Review of General Psychology, 2, 175–220.

Olkin, I. (1990). History and goals. In K. Wachter & M. Straf (Eds.), The future of meta-analysis (pp. 3–10). New York, NY: Russell Sage.

Orwin, R. G., & Vevea, J. L. (2009). Evaluating coding decisions. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), The handbook of research synthesis and meta-analysis (2nd ed., pp. 177–203). New York, NY: Russell Sage.

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. Psychological Methods, 3, 354–379.

Patall, E. A., Cooper, H., & Robinson, J. C. (2008). The effects of choice on intrinsic motivation and related outcomes: A meta-analysis of research findings. Psychological Bulletin, 134, 270–300.

Pearson, K. (1904). Report on certain enteric fever inoculation statistics. British Medical Journal, 3, 1243–1246.

Peek, P., & Pomerantz, J. (1998). Electronic scholarly journal publishing. In M. E. Williams (Ed.), Annual review of information science and technology (pp. 321–356). Medford, NJ: Information Today.

Peters, D. P., & Ceci, S. J. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. Behavioral and Brain Sciences, 5, 187–255.

Petticrew, M., & Roberts, H. (2006). Systematic reviews in the social sciences: A practical guide. Malden, MA: Blackwell Publishing.

Pigott, T. D. (2009). Methods for handling missing data in research synthesis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), The handbook of research synthesis and meta-analysis (2nd ed., pp. 399–416). New York: Russell Sage.

Pigott, T. D. (2012). Advances in meta-analysis. New York: Springer.

Pope, C., Mays, N., & Popay, J. (2007). Synthesizing qualitative and quantitative health evidence: A guide to methods. Berkshire, UK: Open University Press.

Popper, K. (2002). The logic of scientific discovery. London, UK: Routledge.

Price, D. (1965). Networks of scientific papers. Science, 149, 510–515.

PRISMA. (2015). Transparent reporting of systematic reviews and meta-analysis. Retrieved from http://www.prisma-statement.org/statement.htm

Promising Practices Network (PPN). (2014). How programs are considered. Retrieved from http://www.promisingpractices.net/criteria.asp

Randolph, J. J., & Shawn, R. (2005). Using the binomial effect size display (BESD) to present the magnitude of effect sizes to the evaluation audiences. Practical Assessment, Research & Evaluation, 10(14), 1–7.

Raudenbush, S. W. (2009). Random effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), The handbook of research synthesis and meta-analysis (2nd ed., pp. 295–315). New York, NY: Russell Sage.

Raudenbush, S. W., & Bryk, A. S. (2001). Hierarchical linear models: Applications and data analysis methods. Thousand Oaks, CA: Sage.

Reed, J. G., & Baxter, P. M. (2009). Using reference databases. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), The handbook of research synthesis and meta-analysis (2nd ed., pp. 73–101). New York, ny: ssell Sage.

Rosenthal, R. (1978). How often are our numbers wrong? American Psychologist, 33, 1005–1008.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. Psychological Bulletin, 86, 638–641.

Rosenthal, R. (1984). Meta-analytic procedures for social research. Beverly Hills, CA: Sage.

Rosenthal, R. (1990). How are we doing in soft psychology? American Psychologist, 45, 775–777.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. Behavioral and Brain Sciences, 3, 377–386.

Rosenthal, R., & Rubin, D. (1982). A simple, general purpose display of magnitude of experimental effect. Journal of Educational Psychology, 74, 166–169.

Ross, L., & Lepper, M. R. (1980). The perseverance of beliefs: Empirical and normative considerations.

New Directions for Methodology of Social and Behavioral Science, 4, 17–36.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analysis: Prevention, assessment and adjustment. West Sussex, UK: John Wiley & Sons.

Sandelowski, M., & Barroso, J. (2007). Handbook for synthesizing qualitative research. New York: Springer.

Scarr, S., & Weber, B. L. R. (1978). The reliability of reviews for the American Psychologist. American Psychologist, 33, 935.

Schmidt, F. L., & Hunter, J. E. (2015). Methods of meta-analysis: Correcting error and bias in research findings (3rd ed.). Thousand Oaks, CA: Sage.

Schram, C. M. (1989). An examination of differential-photocopying. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Scott, J. C., Matt, G. E., Wrocklage, K. M., Crnich, C., Jordan, J., Southwick, S. M., . . . & Schweinsberg, B. C. (2015). A quantitative meta-analysis of neurocognitive functioning in posttraumatic stress disorder. Psychological Bulletin, 141, 105–140.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin.

Shadish, W. R., & Haddock, K. (2009). Combining estimates of effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), The handbook of research synthesis and meta-analysis (2nd ed., pp. 257–277). New York, NY: Russell Sage.

Shadish, W. R., & Rindskopf, D. M. (2007). Methods for evidence-based practice: Quantitative synthesis of single-subject designs. New Directions for Evaluation, 113, 95–109.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. American Psychologist, 32, 752–760.

Smith, P. J., Blumenthal, J. A., Hoffman, B. M., Cooper, H., Strauman, T. J., Welsh-Bohmer, K., . . . Sherwood, A. (2010). Aerobic exercise and neurocognitive performance: A meta-analytic review of randomized clinical trials. Psychosomatic Medicine, 72, 239–252.

Stock, W. A., Okun, M. A., Haring, M. J., Miller, W., & Kinney, C. (1982). Rigor and data synthesis: A case study of reliability in meta-analysis. Educational Researcher, 11(6), 10–14.

Suhls, J., & Martin, R. (2009). The air we breathe: A critical look at practices and alternatives in the peer-review process. Perspectives on Psychological Science, 4, 40–50.

Sutton, A. J., & Abrams, K. R. (2013). Bayesian methods in meta-analysis and evidence synthesis. Statistical Methods in Medical Research, 10, 277–303.

Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., & Song, F. (2000). Methods for meta-analysis in medical research. Chichester, UK: John Wiley & Sons.

Tamim, R. M., Bernard, R. M., Borokhovski, E., Abrami, P. C., & Schmid, R. F. (2011). What forty years of

research says about the impact of technology on learning: A second-order meta-analysis and validation study. Review of Educational Research, 81, 4–28.

Taveggia, T. C. (1974). Resolving research controversy through empirical cumulation. Sociological Methods and Research, 2, 395–407.

United States Department of Health and Human Services Agency for Healthcare Research and Quality. (2013). Findings of Bayesian mixed treatment comparison meta-analyses: Comparison and exploration using real-world trial data and simulation. Washington, DC: Author.

Valentine, J. C., & Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The study design and implementation assessment device (study DIAD). Psychological Methods, 13, 130–149.

Webb, E. J., Campbell, D. T., Schwartz, R. D., Sechrest, L., & Grove, J. B. (2000). Nonreactive measures in the social sciences (REv. ed.). Thousand Oaks, CA: Sage.

What Works Clearinghouse. (2014). Review process: Standards. Retrieved from http://e/wwc/DocumentSum.aspx?sid=19ies.ed.gov/ncee/wwc/

Wikipedia. (2015). Standard normal table. Retrieved July 28, 2015, from http://en.wikipedia.org/wiki/Standard_normal_table

Wilson, D. B. (2009). Systematic coding for research synthesis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), The handbook of research synthesis and meta-analysis (2nd ed., pp. 159–176). New York, NY: Russell Sage.

Wilson, D. B. (2015). Practical meta-analysis effect size calculator. Retrieved from http://www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-Home.php

Xhignesse, L. V., & Osgood, C. (1967). Bibliographical citation characteristics of the psychological journal network in 1950 and 1960. American Psychologist, 22, 779–791.

Xu, J., & Corno, L. (1998). Case studies of families doing third-grade homework. Teachers College Record, 100, 402–436.

Younger, M., & Warrington, M. (1996). Differential achievement of girls and boys at GCSE: Some observations from the perspective of one school. British Journal of Sociology of Education, 17(3), 299–313.

Author Index

Subject Index