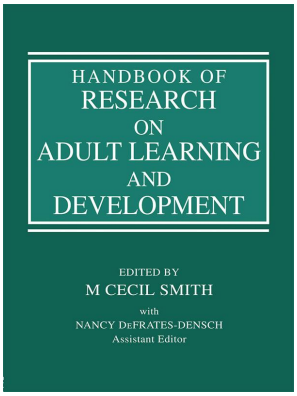

Handbook of Research on Adult Learning and Development

M Cecil Smith, Nancy DeFrates-Densch

Research Synthesis and Meta-Analysis

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9780203887882.ch6
Jeffrey C. Valentine, Harris Cooper
Published online on: 07 Nov 2008

HANDBOOK OF RESEARCH ON ADULT LEARNING AND DEVELOPMENT

EDITED BY

M CECIL SMITH

with

NANCY DEFRATES-DENSCH

Assistant Editor

First published 2009
by Routledge
270 Madison Ave, New York, NY 10016

Simultaneously published in the UK
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

This edition published in the Taylor & Francis e-Library, 2008.

Routledge is an imprint of the Taylor & Francis Group, an informa business

Library of Congress Cataloging in Publication Data
Handbook of research on adult learning and development / edited by M Cecil Smith with Nancy DeFrates-Densch.
p. cm.
Includes bibliographical references and index.
1. Adult learning—Research—Handbooks, manuals, etc. 2. Adult education—Research—Handbooks, manuals, etc. I. Smith, M Cecil. II, DeFrates-Densch, Nancy.

ISBN 0-203-88788-3 Master e-book ISBN

ISBN 10: 0-8058-5819-9 (hbk)
ISBN 10: 0-8058-5820-2 (pbk)
ISBN 10: 0-203-88788-3 (ebk)

ISBN 13: 978-0-8058-5819-8 (hbk)
ISBN 13: 978-0-8058-5820-4 (pbk)
ISBN 13: 978-0-203-88788-2 (ebk)

Chapter 6

Research Synthesis and Meta-Analysis

Jeffrey C. Valentine and Harris Cooper

Imagine that you are the head of a large non-profit foundation that has, as its mission, the goal of improving adult literacy. The foundation's board of directors is particularly interested in developing a new intervention specifically targeted at adults who are not functionally literate. What might such an intervention look like? To find this out, you ask your crack staff of researchers to search the literature for "best practices" in adult literacy programs. If your staff is thorough, it is likely that they will find multiple studies that claim to have implemented and/or tested best practices, and it is likely to be the case that the studies seem to reach different conclusions about which practices are most effective. How should you proceed?

This chapter addresses some of the possibilities for making sense out of a body of research evidence. Progress in any scientific field depends on the accumulation of knowledge, and the accumulation of knowledge depends greatly on the methods used to integrate individual studies into a coherent whole. You will see in this chapter that we are predisposed to a particular type of review, specifically one that is conducted with a similar degree of transparency and rigor as the best scientific studies. In the past, reviewing the literature on best practices in adult literacy would usually have involved a narrative review, in which a scholar would gather some studies that were relevant, read them, then pronounce on what those studies have to say. Typically, little attention has been paid to whether the studies could claim to be representative of the studies that had been conducted, and almost nothing was said about the standards of proof that were employed during the review. Further, results were often presented in impressionistic terms, with little insight provided about the magnitude of some relation (e.g., in the context of an adult literacy intervention, *how much* of an effect the intervention had on participants). Increasingly, however, scholars recognized that literature reviews did not meet the standards of rigor and transparency required in primary research.

At the same time, the amount of available research in the social and medical sciences increased dramatically. A savvy and conscientious scholar might be able to reasonably synthesize the results of a few studies. But what were Bob Rosenthal and Don Rubin (Rosenthal & Rubin, 1978), to do when, in the course of looking for studies that investigated expectancy effects, they found 345 studies? Or Jack Hunter and Frank Schmidt (Hunter & Schmidt, 1979, when they found over 900 estimates of the differential validity of employments tests by race? According to Glass (1976), when faced with this situation reviewers would likely "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. 6).

In contrast to the narrative review, a systematic research synthesis (or simply, a systematic review) employs a collection of methodological and statistical techniques designed to

improve upon the integration of empirical studies. Systematic research syntheses employ literature searching strategies that are meant to minimize differences between retrieved studies and studies that could not be located. Decisions about whether to include or exclude studies based on methodological quality or other criteria are made explicit prior to their application and are applied even-handedly. The extraction of information from research reports is carried out using coding rules similar to those developed for the scientific analysis of document content. Meta-analysis, or the statistical integration of the results of studies, is conducted using open standards of proof, and is approached with the same structure and rigor as is data analysis in primary studies.

From the outset, it is important to recognize that these procedures, including meta-analysis, are not a complete solution to the problems faced by research synthesists. Indeed, meta-analysis is unsuited to addressing certain kinds of questions. For example, theoretical analyses, taxonomic studies, and qualitative research cannot be analyzed using meta-analytic techniques. A scholar interested in tracing the historical development of the concept of literacy would not find meta-analysis a useful tool. In addition, a basic premise behind the use of statistics in research synthesis is that a research body of conceptually relevant studies exists. If this assumption is not met, a quantitative research synthesis cannot be done. And more generally, as Wachter and Straf (1990) point out, meta-analysis is no substitute for wisdom. A statistical method cannot generate theories or hypotheses that do not already exist, nor can it tell us what topics are important to study. Even less ambitiously, a statistical method cannot point out to its user what variables should be examined as moderators of relationships. These can only be accomplished through the thoughtful application of the human intellect.

In this chapter, we describe the ways in which systematic research synthesis can be an improvement over traditional narrative reviews. Also, we will describe briefly the processes by which a systematic review is carried out. We do not aim to cover all the details of how to conduct a systematic review. For such coverage, we can recommend texts by Cooper (1998), Cooper and Hedges (1994), and Lipsey and Wilson (2001). To help illustrate certain concepts and practices, we will refer to a systematic review by Torgerson, Porthouse, and Brooks (2005), as it is representative of the kind of reviews that can be done in the area of adult literacy.

A Brief History of Research Synthesis and Meta-Analysis

A century ago, Karl Pearson conducted what is believed to be one the first statistical syntheses of results of independent research (Pearson, 1904). Pearson gathered data from eleven studies of the effect of a vaccine against typhoid and for each study he calculated a new statistic called the correlation coefficient. He averaged the correlations and concluded that other vaccines were more effective. Three decades later, Ronald Fisher presented a technique for combining the probability values that came from statistically independent tests of the same hypothesis (Fisher, 1932). Early work on quantitative procedures for integrating results of independent studies was ignored for many years. However, the twin problems of (a) the lack of rigor in traditional literature reviews and (b) the dramatic increase in the numbers of studies available for review led to a revival of sorts.

Glass (1976) coined the term meta-analysis to stand for "the statistical analysis of a large collection of analysis results from individual studies for purposes of integrating the findings" (p. 3). Shortly thereafter, Cooper (1979) and Cooper and Rosenthal (1980) made the empirical case for meta-analysis by showing that narrative review procedures led to inaccurate or imprecise characterizations of the cumulative research

results. Glass, McGaw, and Smith (1981) then proposed that meta-analysis be viewed as a new application of analysis of variance and multiple regression procedures, with the outcomes of studies, in the form of effect sizes, treated as the criterion variable and the features of studies as the predictor variables. Hunter, Schmidt, and Jackson (1982) introduced an alternative meta-analytic model that (a) compared the observed variation in study outcomes to that expected by sampling error and (b) corrected the mean and variance of observed effect sizes for known sources of bias (e.g., measurement error, restrictions in sampled ranges). Rosenthal (1984) presented a compendium of meta-analytic methods including combining significance levels, effect size estimation, and the search for moderators of study results. Importantly, Rosenthal's procedures for testing moderators of effect sizes were not based on traditional inferential statistics, but on a new set of techniques involving assumptions tailored specifically for the analysis of study outcomes.

Simultaneous with the development of meta-analysis procedures, several attempts were undertaken to examine and place research synthesis in the context of a scientific process. In 1971, Feldman published an article entitled "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" (Feldman, 1971, p. 86). In the same year, Light and Smith (1971) presented a "cluster approach" to research synthesis that was meant to redress some of the deficiencies in the existing strategies. They 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 reviewing methods.

Two papers that appeared in the *Review of Educational Research* in the early 1980s brought the meta-analytic and review-as-research perspectives together. First, Jackson (1980) proposed six reviewing tasks "analogous to those performed during primary research" (p. 441). His paper employed a sample of 36 review articles from prestigious social science periodicals to examine the methods used in syntheses of empirical research. His conclusion was that "relatively little thought has been given to the methods for doing integrative reviews" (p. 459).

Cooper (1982) drew the analogy between research synthesis and primary research to its logical conclusion. He presented a five stage model of the review that viewed research synthesis as a data gathering exercise and, as such, applied to it criteria similar to those employed to judge primary research. Cooper argued that, similar to primary research, a research review involves problem formulation, data collection (the literature search), data evaluation, data analysis and interpretation (the meta-analysis), and public presentation. For each stage, Cooper codified the research question, its primary function in the review, and the procedural differences that might cause variation in the conclusions of different reviews. In addition, Cooper applied the notion of threats-to-inferential-validity—introduced by Campbell and Stanley (1966; also see Cook & Campbell, 1979) for evaluating the utility of primary research designs—to research synthesis. He identified numerous threats to validity associated with reviewing procedures that might undermine the trustworthiness of a research synthesis' findings.

Light and Pillemer (1984) offered a text that emphasized the use of research synthesis to inform social policy. Their approach highlighted the importance of meshing quantitative procedures and narrative descriptions in the interpretation and communication of synthesis results. Hedges and Olkin's (1985) text, titled *Statistical Procedures for Meta-*

Analysis, covered a wide array of meta-analytic procedures and established the procedures' legitimacy by presenting rigorous statistical proofs.

During and after the years that the works mentioned above were appearing, the use of meta-analysis spread from psychology and education through many disciplines, especially social policy analysis (Light, 1983) and the medical sciences (see *Statistics in Medicine*, 1987, Vol. 6, No. 3). These developments led Greenberg and Folger (1988) to state that "if the current interest in meta-analysis is any indication, then meta-analysis is here to stay" (p. 191).

Since the mid-1980s, several full-text treatments have appeared on meta-analysis. Some of these treat the topic generally (e.g., Cooper, 1998; Hunter & Schmidt, 1990; Lipsey & Wilson, 2001), some treat it from the perspective of particular research design conceptualizations (e.g., Eddy, Hasselblad, & Shachter, 1992), some are tied to particular software packages (e.g., Johnson, 1989), and some look at potential future developments in research synthesis (e.g., Wachter & Straf, 1990, Cook et al., 1992). In 1994, the first edition of *Handbook of Research Synthesis* was published (Cooper & Hedges, 1994). This book included 32 chapters contributed by specialists in information science, computer software, and statistics, as well as experts in the use of research synthesis for psychology, medicine, education, and public policy.

Components of Research Synthesis

As we discussed earlier, systematic research syntheses are conducted much like primary research (Cooper, 1998). A problem is formulated, data are collected, evaluated, analyzed, and interpreted, and results are presented to the public. In a sense, research synthesis may be thought of as a form of survey research, in which studies (as opposed to people) are the population of interest (Lipsey & Wilson, 2001). The "response" offered by the research report is called an effect size and the "demographics" related to the response are the characteristics of the study (its design, implementation characteristics, and sample). An effect size is a measure of the magnitude of the relationship in question. For example, a study investigating the relationship between literacy and socio-economic status (SES) might result in a correlation coefficient between these two variables; this correlation is the effect size. However, before the effect size is calculated, the researcher has made numerous decisions that unfold in systematic order. We will now discuss each stage of research synthesis.

Problem Formulation

In its most basic form, problem formulation involves identifying at least two relevant variables, specifying the anticipated relation between them, and providing a rationale for relating the variables to one another. So for example, my variables of interest might be adult literacy and SES. One might suspect that these are positively related (i.e., as wealth increases so does literacy), with two complementary reasons being (a) that having low literacy reduces one's value in the workforce and (b) being poor often involves a restriction in access to activities and materials that would enhance literacy.

While the problems addressed by primary researchers are limited only by their imaginations and funding levels, research synthesists have no choice but to study topics that have already appeared in the empirical literature. In fact, a topic may not be suitable for review unless a substantial number of studies have taken it as a problem. This does not

mean, however, that research synthesis is not a creative enterprise. Instead, the creativity enters when the synthesist must make sense of many related but not identical studies. More often than not, the cumulative results of studies are many times more complex than the result of any single study.

The most important issues that arise in defining the problem for research syntheses are (a) how broadly to define the conceptual variables of interest and (b) how to handle multiple operational definitions of the same conceptual variables. Although no two participants in any single study will be treated exactly alike, this variability within studies is small compared to the variability produced by differences in operationalizations across studies.

Torgerson et al. (2005) were interested in interventions designed to increase the literacy and/or numeracy of adults. This is a broad research question made necessary by the relative lack of research in this area. As such, their research problem formulates a very general question: For adults, is there a relationship between receiving an intervention designed to enhance literacy or numeracy and outcomes in those areas, relative to individuals who do not receive such an intervention?

Optimally, a research synthesist will specify in advance the conceptual definitions of a variable covered by the review and will then assess the fit between this definition and the operations employed in any given study. In practice, it is often necessary for the synthesist to retain some flexibility in judging the relevance of a given operation of the dependent variable. Variables can be operationalized in a surprising number of ways, and it is rare to have perfect knowledge of this diversity before undertaking a search of the literature. Thus, more so than is true in primary research, problem formulation in research synthesis is an iterative process. The synthesist begins with a defined concept and known operational realizations but these two ideated sets are refined as the subsequent stages of the review proceed. In the case of Torgerson et al. (2005), the research question is so broad that this activity was not necessary, but generally synthesists will have to wrestle with this problem.

The Literature Search

For the research synthesist, the data collection stage involves gathering studies and making judgments about their relevance to the question at hand. Too often, scholars conducting a narrative research review rely on a convenience sample of studies on which to base their conclusions. For example, the reviewer might only include studies published in a few selected journals plus those that have come to their attention through informal means. Thus, the studies reviewed are neither exhaustive nor representative of all studies that have been conducted on the topic of interest.

Systematic research synthesists are advised to operate with the goal of obtaining all relevant research (Cooper, 1998; Lipsey & Wilson, 2001), regardless of whether or where it was published. Of course, whether this goal is ever obtained is unknowable. Indeed, we can be reasonably certain that the goal is often unobtainable. Setting the goal, however, does lead to the use of comprehensive searching strategies that are meant to reduce biases in the results of obtained literature. Here, bias is defined as systematic differences in the outcomes of studies that do and do not come to the attention of the synthesist.

One important question faced by research synthesists involves whether to include unpublished research. Two reasons are frequently given for excluding unpublished research reports. The first is that the research base has the potential for becoming too large and unwieldy for a narrative review. Meta-analysts solve this problem by using computers to help store, sort, and analyze study results.

The second reason given is that unpublished research is often of lesser quality than published research. We would argue that this is too simple a conclusion. For example, researchers may not publish their results because publication is not their objective and if so this decision is independent of the quality of their work (see Cooper, DeNeve, & Charlton, 1997). Conversely, most researchers would agree that some low quality research does get published. Suffice it to say that the quality of research, published and not, is distributed along overlapping continua.

Moreover, research is often turned down for publication for reasons other than quality (Greenwald, 1975). In particular, research papers failing to achieve standard levels of statistical significance are frequently left in researcher's "file drawers," a problem known as publication bias. The concern here is that studies revealing smaller effect sizes will be systematically censured from the published literature and therefore estimates of effect in the published literature may make relationships appear stronger than if all estimates were available to the synthesist. For this reason, it is now accepted practice that rigorous research syntheses will include both published and unpublished research.

Torgerson et al. (2005) employed two main strategies to find relevant literature. They conducted keyword searches of several databases (including a specialty database that houses unpublished research), and also examined the reference sections of research reviews that were identified in their search for additional studies. Reviewers might also try contacting leading researchers, relevant professional organizations, and agencies known to fund work in the area of the research question.

Data Collection and Evaluation

In a typical research synthesis, trained personnel use standardized coding procedures to extract the desired information from research reports. The process is analogous to collecting data on a survey questionnaire or, more directly, to the content analysis of documents (Weber, 1990). The coding sheet contains information about the background characteristics of the reports (e.g., author, year of publication) as well as the specifics of the study, such as sample size and composition, research methodology, and study results. While some collected information will be the same regardless of the topic under consideration (for example, all reviewers will collect information on sample sizes), other coded study characteristics will be unique to the substantive topic of interest. Lipsey (1994) discusses how synthesists can go about identifying study features that might influence study outcomes.

Stock (1994) presented a detailed discussion of how to assemble a coding guide for a research synthesis. A good coding guide will facilitate complete, unambiguous, and reliable extraction of all relevant data in a research report. Among the issues Stock addressed, none is more important than the reliability of data extraction.

Reliability of Data Extraction. Traditional narrative reviewers give very little (if any) attention to the reliability of their descriptions of studies. On the other hand, systematic research reviewers pay special attention to errors that occur when extracting data from a research report. There are several sources of error in research reviews. One source of error occurs when primary researchers do not report or poorly report data that are of interest to the research synthesist. This occurs, for example, when researchers performing research on adult literacy neglect to describe how samples were obtained. To counter this problem, the synthesist may choose to contact the study authors directly for assistance. DuBois, Holloway, Valentine, and Cooper (2002) utilized this approach in a meta-

analysis of the effectiveness of youth mentoring programs and had moderate success. When a research report contained missing or incomplete data, these authors attempted to contact the primary author through electronic mail. Success in both contacting study authors and retrieving data were dependent on the amount of time that had passed since original publication of the research report; we had very little success with reports that were more than 10 years old, but were able to retrieve some information from newer reports. While labor intensive, this approach added data to the synthesis that would have been otherwise unavailable, increasing the representativeness of the data set. One can only hope that the current trend of decreasing costs for electronic storage will increase the probability that study authors will be willing and able to accommodate these requests in the future.

Research reviewers are not immune to making mistakes themselves. To assess the extent to which study information has been reliably extracted from research reports, most research synthesis texts suggest performing some sort of reliability assessment. This involves employing procedures akin to those used in assessing interjudge reliability in other research domains (e.g., Lipsey & Wilson, 2001; Orwin, 1994). At a minimum, a reliability assessment will involve at least two (and possibly more) coders working independently. If a substantial number of studies have met inclusion criteria, the reliability assessment could be conducted by having one of the researchers code a randomly sampled subset of studies. With a smaller number of studies, best practice is to have researchers independently code all studies. Disagreements then may be resolved in conference and/or by a third reader. This procedure raises the effective reliability of codes to very high levels, and is the approach adopted by Torgerson et al. (2005).

Empirical results suggest that different types of information about research are extracted from reports with different levels of reliability. For example, Stock, Okun, Haring, Miller, Kinney, and Ceurvorst (1982) found that, among 27 types of information extracted from research reports describing studies of the predictors of life satisfaction, the mean reliability coefficient (uncorrected for chance), was $r = .88$. Eighteen of the 27 characteristics had a reliability coefficient greater than or equal to .90, and an additional six characteristics had a reliability coefficient of $.80 \leq r \leq .90$.Two characteristics had reliabilities below .60, suggesting that they could not be reliably coded by the readers. One of these codes was "quality of study."

Judging Research Quality. It is a truism that a research synthesis is only as good as the studies that go into it. While multiple operationism can overcome some deficiencies in research design (see Cook & Campbell, 1979), if only low quality studies have been conducted on the topic of interest, one cannot expect a synthesis to yield interpretable findings. Of course, this holds true whether traditional narrative review procedures or meta-analysis are employed.

In some instances, research synthesists are interested in a variable that requires a high degree of inference on the part of the coders. Among these, the most frequent high inference codes occur because the synthesists wish to determine whether "study quality" moderates the estimated magnitude of the relationships in question. For example, do "high quality" studies reveal larger or smaller effects of an adult literacy intervention than "low quality" studies? Assessing whether a study is "good" or "bad" requires a complex judgment on the part of the coder, and the level of inference required may introduce error in the research review. As suggested by Stock et al. (1982), arriving at a dichotomous decision about study quality (good vs. bad) or even reducing study quality to a single continuous dimension can be difficult.

This is not to suggest that all high inference codes are unreliable. For example, Miller, Lee, and Carlson (1991) had judges read a description of an experimental manipulation. Judges made inferences about the affective, cognitive, and behavioral responses of study participants. Results suggested that judges effectively inferred participant affective responses, but did not reliably infer participant behavior or attitude change.

Research reviewers often employ quality scales designed to help them quantify the quality of a study's design and implementation. However, these scales unfortunately have little to recommend them. They generally lack sufficient operational specificity and they are based on criteria that lack empirical support. In a demonstration of these problems Jüni, Witschi, Bloch, and Egger (1999) applied 25 different quality scales to 17 studies reporting on trials comparing the effects of low-molecular weight heparin (LMWH) to standard heparin on the risk of developing blood clots after surgery. After applying the 25 quality scales to the 17 trials, the authors then performed 25 different meta-analyses examining, in each case, the relationship between study quality and the effect of LMHW (relative to standard heparin). Then, the authors examined the conclusions of the meta-analyses separately for "high" and "low" quality trials. For six of the quality scales, the "high quality" studies suggested no difference between LMWH and standard heparin, while the "low quality" studies suggested a significant positive effect for LMWH. For seven other quality scales, this pattern was reversed. That is, the "high quality" studies suggested a positive effect for LMWH, while the "low quality" studies suggested no difference between the two conditions. The remaining 12 quality scales resulted in conclusions that did not differ between "high" and "low" quality trials. Thus, Jüni et al. (1999) suggested that the clinical conclusion about the efficacy of the two types of heparin depended on the quality scale used.

The dependence between the outcomes of the synthesis and the quality scale used reveals a critical problem with the scales. A team of scholars could have chosen Quality Scale A, used it to exclude low quality studies, and then concluded that LMWH is effective. Starting with the same research question, another team of scholars could have chosen Quality Scale B, used it to exclude low quality studies, then concluded that LMWH is no more effective than standard heparin. As such, the scales appear to have been at best useless and at worst misleading (Berlin & Rennie, 1999).

Examination of the quality scales used makes it easy to see why this result occurred. The 25 quality scales often focused on different dimensions, and analyzed a differing number of dimensions (ranging from 3 to 34), including dimensions that are unrelated to validity in the traditional sense (e.g., whether or not an institutional review board approved the study). Even when the same dimension was analyzed, the weights assigned to the dimension varied. For example, one scale allocated 4% of total points on their scale to the presence of randomization, and 12% of the points to whether or not the outcome assessor was unaware of the condition the participants were in (called masking). Another scale very nearly reversed the relative importance of these two dimensions, allocating 14% of the total points to randomization and 5% to masking. Simply stated, there is no empirical justification for these weighting schemes, and the result in the absence of that justification is a hopelessly confused set of arbitrary weights.

Further, the scales reviewed by Jüni et al. (1999) share a reliance on single scores to represent a study's quality. Especially when scales focus on more than one aspect of validity, the single score approach results in a score that is summed from very different aspects of study design and implementation, many of which are not necessarily related to one another. For example, there is no necessary relation between the validity of outcome measures and the mechanism used to allocate participants to groups. When scales

combine disparate elements of study design into a single score, it is likely that important considerations of design are being obscured. For example, a study with strong internal validity but weak external validity can get a score identical to a study with weak internal validity and strong external validity. If the quality of these studies is expressed as a single number, how would one know the difference between these studies with such very different characteristics?

As such, we think the best strategy for addressing study quality in a research synthesis is for the reviewers to think deeply about the features of study design, implementation, and analysis that are likely to affect study results in their area and to carefully code studies on these dimensions. The reviewers can then explore the relation between these characteristics and study outcomes. Further, if the reviewers have a strong basis for believing that a characteristic biases a collection of studies, then studies with that characteristic could be excluded from the review (but note that, as much as possible, decisions like these should be made before data collection begins).

Torgerson et al. (2005) employed quality criteria in this manner, using items that were suggested by the CONSORT statement (Moher, Schulz, & Altman, 2001). The CONSORT statement addresses aspects of study design and implementation that ought to be reported in a study write up. Torgerson et al. then described where each of the included studies fell on each quality dimension.

The Statistical Analysis and Interpretation of Combined Study Results

In this section we discuss statistical methods that help reviewers summarize research results. Meta-analyses generally lead to more precise and reliable conclusions about a research base than does the narrative integration of research. An empirical example demonstrates this point. Cooper and Rosenthal (1980) randomly assigned 41 faculty and graduate students in a psychology department to read seven articles addressing the question of sex differences in task persistence. All participants read the same seven articles chosen to suggest that women have greater task persistence than men. Some participants were assigned to a meta-analysis condition, in which they were given detailed instructions about how to combine the study's significance levels and to obtain an overall estimate of the effect size for the seven studies. Participants assigned to the narrative review condition were simply told to use whatever procedures they typically would use to assess the literature.

After the participants completed their reviews, they were asked whether the articles they read supported the conclusion that women exhibit greater task persistence than men. Participants could respond "definitely yes," "probably yes," "can't tell," "probably no," or "definitely no." Among the narrative reviewers 73% found definitely or probably no support for the hypothesis compared with only 32% of participants using the meta-analytic technique. In fact, the combined probability that the null hypothesis was true was $p < .005$, indicating over twice as many inferential errors by the narrative than the quantitative reviewers. This study suggests that meta-analysis can be a superior data integration strategy.

As we mentioned in the introduction, traditional narrative reviewers are unable to generate effect sizes for the hypotheses they test, and do not employ strategies for weighting individual studies proportional to their size or quality. Rather, most narrative reviewers use, explicitly or not, a vote-count of studies. Often, this procedure leads to inferential errors.

The Vote-Count. Typically, when conducting a vote count, the reviewer assigns research reports to one of three categories: either the relation between the two variables is significantly positive, significantly negative, or the null hypothesis cannot be rejected (Bushman, 1994). The number of reports in each category is counted, and a decision rule is applied to the count to determine what the research base suggests about the relation. The category with the largest number of votes wins.

The vote count is intuitively appealing. However, it has few defenders as a final data analysis strategy. Hedges and Olkin (1985) have demonstrated that in many instances in the social sciences, vote counts have power characteristics that are inversely related to the number of studies contained in the review. That is, counterintuitively, when a real but moderate effect exists in a population of studies that are carried out with less-than-ideal statistical power characteristics, the *more* studies that a review covers, the *less* likely it is that a vote count will reject the null hypothesis. In addition, the vote counting strategy does not differentially weight studies based on sample size. This is a problem because a study with 100 participants and a study with 1,000 participants are given equal weight, even though the larger study provides the more precise and reliable estimate of the effect. Further, the effect size of the studies reviewed is not considered. A study showing small negative effects is given the same weight as a study showing large positive ones. For these reasons, vote counting is not considered a credible analytic strategy for drawing inferences (Cooper & Dorr, 1995).

Effect Size Metrics. As an alternative to vote counting procedures, meta-analysts will (a) calculate an effect size for the outcomes of hypothesis tests in every study, (b) average these effect sizes across hypothesis tests to estimate general magnitudes of effect, and (c) compare effect sizes to discover if variations in outcomes exist and, if so, what features of comparisons might account for them.

Cohen (1988) defined an effect size as "the degree to which the phenomenon is present in the population, or the degree to which the null hypothesis is false" (pp. 9–10). For the meta-analyst, estimates of effect size are the most crucial output of studies. Because the reporting of effect sizes in primary research is not yet universal, a meta-analyst often must estimate or calculate the effect size from other statistics present in a research report (see Rosenthal, 1984, for many of these approaches). Often, they also must adjust an effect size estimate to remove certain sampling biases.

Two effect size metrics are most applicable to research on adult literacy. The first, known as the standardized mean difference (or *d*-index, Cohen's *d*, or simply as *d*) by is a scale-free measure of the separation between two group means (Cohen, 1988). Calculating the basic standardized mean difference for any comparison involves dividing the difference between the two group means by their average (or pooled) standard deviation. This calculation results in a measure of the difference between the two group means expressed in terms of their common standard deviation. For example, a standardized mean difference of .25 indicates that one-quarter standard deviation separates the two means.

Because the interpretation of the standardized mean difference effect size is not intuitively transparent in many contexts, Cohen (1988) presented a measure associated with it called U_3. U_3 describes the percentage of the sample with the lower mean that was exceeded by the average (or 50th percentile) score in the higher-meaned group. When $d = 0$, $U_3 = 50\%$, suggesting that half of the scores in the lower-meaned group were exceeded by the mean of the higher-scoring group. Of course, this suggests no difference

between the groups. When $d = .50$, or when the higher-scoring group mean is one-half standard deviation above the lower-scoring group mean, $U_3 = 69\%$. This suggests that 69% of the scores in the higher-meaned group exceed the average score in the lower-meaned group.

A related way to think about the interpretation of the standardized mean difference effect size is that it represents the probability that a randomly selected member of the treatment group will outscore a randomly selected member of the comparison group. For example, assume that a meta-analysis of adult literacy programs found that the average effect size was $d = +.50$, with the positive sign in front of the d-index, indicating that the participants in the intervention outscored the participants in the comparison condition. The $d = +.50$ can be interpreted as a 69% chance that a randomly-chosen participant receiving the literacy intervention would outscore a randomly-chosen comparison group member. If the studies in this review were all randomized experiments, we would have a strong basis for believing that, if the intervention were not effective at all, a randomly-drawn intervention group member would have a 50% chance of outperforming a randomly-drawn comparison group member.

The second effect size metric well-suited to research in adult literacy is the r-index, or correlation coefficient. The correlation coefficient is familiar to most researchers and students. However, interpretations of the correlation coefficient that rely on the proportion of variance explained (i.e., r^2), are frequently misinterpreted even by experienced researchers (cf, Abelson, 1985; Rosenthal & Rubin, 1982). The physicians' aspirin study (Steering Committee of the Physicians' Health Study Research Group, 1988) provides a particularly compelling example. Over a 5-year period, approximately 22,000 physicians with a history of heart attack took either 325 mg. of aspirin every other day or they took a placebo. The data showed that 1.7% of the placebo group and .09% of the aspirin group had a second heart attack during the course of the study. The correlation associated with these data was $r = .03$, and $r^2 = .0009$, a seemingly trivial relationship. However, this effect means that taking an aspirin every other day was associated with a significant reduction in the risk of having a second heart attack. In fact, risk of a second heart attack was reduced by more than one-third (Rosenthal & Rosnow, 1991). Olkin (1992), citing Chalmers (n.d.) suggests that had research synthesis techniques been applied to the efficacy of aspirin in reducing the risk of a second heart attack, sufficient evidence demonstrating the clinical benefits of aspirin existed as early as 1976, with an estimated savings of 10,000 to 20,000 lives over the next decade.

Typically, the correlation coefficient is used to express the relationship between two continuous variables, such as continuous measures of literacy and earnings per year. On the other hand, the d-index is used to relate one dichotomous variable to a continuous variable, such as comparing groups of qualitatively different participants (e.g., males and females, Whites and African Americans), on a continuous measure, such as literacy. The choice of metric is determined by which fits best with the characteristics of the variables under consideration.

Identifying Independent Samples. A statistical problem arises when a single study contains multiple effect size estimates taken on the same sample of participants. There are several approaches meta-analysts use to handle such dependent effect sizes. Some treat each effect size as independent, regardless of the number that comes from the same sample of people. They assume that the effect of violating the independence assumption is not substantial. Other meta-analysts use the study as the unit of analysis. They calculate the mean effect size or take the median result and use this value to represent the study.

Sophisticated statistical models also have been suggested as a solution to the problem of dependent effect size estimates (Gleser & Olkin, 1994; Raudenbush, Becker, & Kalaian, 1988), but due to their complexity they are yet rarely found in practice.

Examining Effect Size Distributions. Inspecting the effect size distribution is an important part of a complete meta-analysis. This information often includes stem-and-leaf displays and/or funnel plots (Greenhouse & Iyengar, 1994) that help convey information about the magnitude and variation in effect sizes, as well as help assess whether publication bias might exist in the sample of comparison outcomes (see below for a discussion of publication bias).

The distribution of effects will also be examined to determine whether it contains statistical outliers (Barnett & Lewis, 1978). As with data based on primary research, if outliers are present meta-analysts then must decide how to treat them. Some synthesists remove them from the data set entirely while others will modify the outlying values so as to make them conform more closely to the general distribution of results. For either option, the meta-analysts wish both to make the distribution of effects more normal and to mitigate the effect of a few extreme values on measures of central tendency and dispersion, and on subsequent moderator analyses. Whichever option is chosen, it is important to inform the readers so they may make their own judgments about the adequacy of the strategy.

Averaging Effect Sizes and Measuring Their Dispersion. The most pivotal outcomes of a meta-analysis are the average effect sizes and measures of dispersion that accompany them. Weighted procedures are typically used to calculate average effect sizes across comparisons. In this procedure, each independent effect size is first multiplied by the inverse of its variance and the sum of these products is then divided by the sum of the inverses.

The weighting procedure is generally preferred to not weighting because it gives greater weight to effect sizes based on larger samples. Also, as noted above, confidence intervals are calculated for weighted average effects. Many texts, including Hedges and Olkin (1985), Shadish and Haddock (1994), and Lipsey and Wilson (2001) provide procedures for calculating the appropriate weights and confidence intervals.

In addition to the confidence interval as a measure of dispersion, research synthesists usually carry out a highly informative procedure called a "homogeneity analysis." A homogeneity analysis compares the amount of variance in an observed set of effect sizes with the amount of variance that would be expected by sampling error alone. If there is greater variation in effects than would be expected by chance, then the meta-analyst begins the process of examining moderators of hypothesis test outcomes. (Note also that the meta-analyst may search for moderators in the absence of a statistically significant homogeneity analysis if there are good theoretical reasons for doing so.)

Artifactual Influences of the Size of Observed Effects. There are numerous factors that influence the magnitude of an observed relation between two variables other than the "true" relationship. In particular, meta-analysts must pay attention to factors that might attenuate the magnitude of the effect size. For example, both (a) unreliability in the scores used to measure the dependent variable and (b) artificially dichotomizing continuous scores on the dependent variable into groups (that is, turning a continuous measure of SES into middle and low wealth groups) will lead to smaller observed effect sizes than the "actual" effect in the population. The synthesist must decide how this attenuation

should be treated. One approach is to estimate the degree to which the effect size has been attenuated due to these and other related issues. Procedures for carrying this out are described in Hunter and Schmidt (1990).

A final influence on effect sizes is the number of factors employed in a research design. If the research design includes more factors than the reviewer is interested in, the reviewer is faced with the choice of either using a standard deviation reduced by the inclusion of additional factors (inflating the estimate of the effect size), or attempting to retrieve the standard deviation that would have been obtained if all extraneous factors had been ignored. Whenever possible, this latter strategy should be employed. In practice, however, it is often difficult to retrieve this estimate. In such cases, we recommend examining whether or not the number of factors involved in the studies is correlated with effect size.

Fixed and Random Effects Models of Error. Another aspect of conducting a meta-analysis that recently has received considerable attention involves the decision about whether a fixed-effects or random-effects model of error influences the generation of study outcomes. In a fixed-effect analysis, each effect size's variance is assumed to reflect only sampling error (i.e., error solely due to participant differences) and thus can be taken into account through the procedures described previously for weighting effect sizes by sample size. When a random-effect analysis is carried out, a study-level variance component is assumed to be present as an additional source of random influence. Hedges and Vevea (1998) state that fixed-effect models of error 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). A further statistical consideration is that in the search for moderators, fixed-effect models may seriously underestimate, and random-effects models seriously overestimate, error variance when their assumptions are violated (Overton, 1998). In view of these competing sets of concerns, one approach is to consider applying both models in all primary study analyses (e.g., Cooper et al., 2000). Specifically, all analyses could be conducted twice, once employing fixed-effect assumptions and once using random-effect assumptions. Differences in results based on which set of assumptions is used can be incorporated into the interpretation and discussion of findings.

Moderator Analysis. The search for why the outcomes of hypothesis tests differ is the most interesting and informative part of conducting a meta-analysis. Because effect sizes are sample statistics, they will vary somewhat even if they all estimate the same underlying population value. Homogeneity analysis allows the meta-analyst to test whether sampling error alone accounted for this variation or whether features of studies, samples, treatment designs, or outcome measures also play a role. The synthesist calculates average effect sizes for subsets of studies, comparing the average effect sizes for different methods, types of programs, outcome measures, and participants and compares these to determine if they provide insight into what influences the strength and/or direction of the relationship. In fact, the synthesist can ask questions about variables that moderate outcomes even if no individual study has included the moderator variable. For example, a meta-analyst can ask whether the effects of a literacy intervention differ for low vs. middle SES individuals, even if no single study ever included both of these groups. The results of such a comparison of average effect sizes can suggest whether this student characteristic would be important to look at in future research and/or as a guide to policy.

After calculating the average effect sizes for different subgroups of comparisons, the meta-analyst statistically tests whether the group factor is reliably associated with different magnitudes of effect. Three statistical procedures for examining variation in effect sizes have appeared in the literature. The first approach applies statistical procedures typically used on primary research data, like ANOVA or multiple regression. The effect sizes serve as the dependent variable and comparison features serve as independent or predictor variables. This approach has been criticized based on the questionable tenability of the underlying assumptions (see Hedges & Olkin, 1985). Most notably, traditional inferential statistics assume that the error in measurement is relatively homogeneous across data points (the assumption of homoscedacity). This assumption is often violated in meta-analytic data sets.

The second approach compares the variation in obtained effect sizes with the variation expected due to sampling error (Hunter & Schmidt, 1990). This approach involves calculating not only the observed variance in effects but also the expected variance, given that all observed effects are estimating the same population value. A formal statistical test of the difference between these two values is typically not carried out. Rather, the meta-analyst adopts a critical value for the ratio of observed-to-expected variance to use as a means for rejecting the null hypothesis. The meta-analyst might also adjust effect sizes to account for methodological artifacts such as sampling error, range restrictions, or unreliability of measurements.

The third approach involves the homogeneity statistic described above. Analogous to ANOVA, comparisons are grouped by features and the average effect sizes for these groups are tested to determine if the averages are drawn from the same population. If this hypothesis is rejected, the grouping variable remains a plausible potential moderator of effect.

Sensitivity Analysis. An additional step in meta-analysis which is gaining popularity is the performance of sensitivity analyses. A sensitivity analysis is used to determine if and how the conclusions of an analysis might differ if it was conducted using different statistical procedures or assumptions. There are numerous points at which a meta-analyst might decide a sensitivity analysis is appropriate. For example, there might be a set of comparisons that fall at the edge of the conceptual definition of what constitutes an acceptably valid measure of SES. The effects of SES might be tested with and without the inclusion of these comparisons. Or, some evaluations of the relation between SES and literacy have missing data. These comparisons might be omitted from one analysis and included in another analysis that makes conservative assumptions about what those values might be.

In sum then, meta-analysts have a wide array of techniques at their disposal. Some of these techniques have been developed specifically for analysis of meta-analytic data sets and others have been adopted from other research methodologies. The specific techniques used in any meta-analysis will differ somewhat depending on the characteristics of the data set and the questions asked by the research synthesist.

Assessment of the Potential for Publication Bias. As defined earlier, publication bias refers to the well-documented tendency for studies lacking statistically significant effects to go unpublished. All else being equal, studies with "less" statistical significance have smaller intervention effects. Given that these studies are less likely to be published, the risk is that a review might present an overly optimistic (i.e., biased) picture of the evidence. As such, it is increasingly common for reviewers to assess the likelihood that publication bias has affected their results. Rothstein, Sutton, and Borenstein (2005) present a

comprehensive overview of the problem of publication bias as well as its assessment. One relatively common technique involves a trim and fill analysis (Duval & Tweedie, 2000). This analysis is based on the assumption that the effects in the observed studies (i.e., the studies included in the review) are drawn from a normally distributed population of effects. With this assumption in mind, the trim and fill analysis essentially examines the distribution of observed effects and "fills" in any effects that appear to be missing. The analysis will generate a new distribution of effects that conforms to the assumption of a normal distribution, and also provides a new estimated effect size that can be compared to the effect size estimated from the observed studies. If the conclusion about the effects of the intervention is similar in both analyses, it lends greater weight to the confidence that can be placed in the conclusion. If the trim and fill analysis suggests that publication bias might be playing a role, caution is warranted.

Unfortunately, the trim and fill analysis is not a perfect solution to the problem of publication bias. The analysis exploits a frequent negative correlation between effect size and study size (ideally, these would be independent). However, there may be good reasons for, say, larger studies to yield smaller effect sizes. For example, interventions may be more difficult to implement as the number of participants increases. If implementation quality is positively related to study outcomes (i.e., better implementation = better outcomes), then this might show up as a negative correlation between study size and study outcomes (i.e., larger studies have smaller effects because they are more difficult to implement with good fidelity). As such, in this case a trim and fill analysis might mistake a problem with implementation fidelity as a problem with publication bias. Despite this limitation, techniques to help assess the plausibility of publication bias should be routinely undertaken in research syntheses. Torgerson et al. (2005) did not carry out a trim and fill analysis, but did produce graphics called funnel plots. These are plots of effects sizes (on the x-axis) and their standard errors (on the y-axis). Just like with the trim and fill analysis, the distribution of effect sizes is analyzed for gaps (the difference is that the analysis is done visually instead of statistically). A good source for issues related to publication bias (including a more complete treatment of funnel plots and the trim and fill procedure) is Rothstein, Sutton, & Borenstein (2005).

When Should a Meta-Analysis Not Be Done? Recall that Torgerson et al. (2005) developed a very broad question for their research synthesis. They did this because they anticipated that they would find relatively few studies; their intuitions were correct. In all, they identified only 18 studies of an experimental or quasi-experimental nature, and only 9 of these presented data. Given that these 9 studies included interventions designed to affect literacy, numeracy, or both, they reasonably felt that combining these studied was not appropriate. Instead, these reviewers presented tables that detailed for readers important characteristics of studies, such as design, sample size and characteristics, intervention characteristics, and outcome measures. They also provide paragraph descriptions of each study that met inclusion criteria. They conclude by stating:

> It is of concern that the few trials that have been undertaken tend to be of low methodological quality … substantial heterogeneity between the included studies … makes it difficult to draw either quantitative or qualitative conclusions about which particular forms of intervention are effective. (p. 99)

We agree with this general approach. Ideally, the conditions under which synthesists

will meta-analytically combine data will be operationalized before the data have been collected.

Public Presentation

Research is not complete until results are shared with the scientific community (American Psychological Association, 2001). We have already touched on some of the issues related to public presentation in previous sections. For example, the research synthesist must take care to interpret and report effect sizes in such a manner as to be understandable to the intended audience. Two other potential sources of invalidity are relevant to the public presentation stage. First, leaving out evidence regarding possible moderating influences on main effects can jeopardize the trustworthiness of the conclusions presented by the research synthesist. Second, the research synthesist should ensure that the synthesis techniques employed are transparent to the reader. That is, the synthesist should provide enough information so that readers can critically assess the methods, strengths, and weaknesses of the research synthesis. Halvorsen (1994) and Light, Singer, and Willett (1994) present numerous suggestions regarding effective public presentation techniques.

Other Threats to Validity in Conducting Research Synthesis

As we suggest above, using rigorous and systematic rules for synthesizing a literature does not ensure that the resulting inferences will be infallible. Cooper (1982) pointed to several threats to the validity of research synthesis conclusions. For example, during problem formulation threats to the validity of a synthesis could occur if the synthesist did not pay proper attention to conceptual distinctions in definitions and hypotheses that were viewed as important by others in the field. The validity of a literature search could be compromised by the use of a few selective sources of research reports or by publication bias. The validity of data evaluation can be threatened if information from research reports is missing, or if the individuals extracting information from documents are poorly trained.

Many more threats to the validity of a research synthesis can be identified (Matt & Cook, 1994). We will elaborate on two here. One obvious threat to the validity of meta-analytic conclusions involves the rules of inference employed by a synthesist. The possibility always exists that the meta-analyst has used an invalid rule for inferring a characteristic of the target population. This occurs because the target population does not conform to the assumptions underlying the analysis techniques. Of course, this is not a shortcoming unique to the use of quantitative integration techniques. In non-quantitative syntheses, rules of inference also must be used but it is difficult to gauge their appropriateness because they are not very often made explicit. For meta-analyses, the suppositions of statistical tests are generally known and some statistical biases in reviews can be removed.

Another threat to validity is that the meta-analyst might capitalize on or suffer because of the probabilitistic nature of statistical findings. First, as in primary research, the meta-analyst might conduct many statistical tests without adjusting for "synthesis-wise" error rates. Second, because of gaps in the literature, a meta-analyst might discover so few tests of a particular hypothesis that the statistical power of the meta-analysis is low. Unlike a primary researcher, the meta-analyst cannot sample more participants (or in this case,

generate more studies) so as to increase the sensitivity of tests. It is possible, however, to expand the search for relevant research.

The Contribution of Systematic Research Reviews

Systematic research reviewing is not a perfect solution to the problems faced by social and behavioral scientists. But still, it can make important—indeed essential—contributions to the scientific enterprise. The use of proper procedures for the synthesis of multiple studies does more then simply ameliorates the problems associated with the traditional narrative review. Systematic research synthesis transforms the difficulties into strengths. Variation in the context, design, and sampling characteristics of individual studies are sources of consternation when studies are examined individually, serially, and narratively. When multiple studies, each limited in their representation of context, design, and sample, are treated as data points in a second round of scientific investigation they contribute jointly to confident, general, and properly contextualized conclusions about relationships and hypotheses.

Because of the potential value of systematic research reviews in the policy domain, both the producers and consumers of reviews now agree they must think about what distinguishes good from bad reviews. Further, they agree that without high-quality reviews, both theoreticians and practitioners will question the value of research for assisting the development of effective explanations for behavior and behavioral interventions.

References

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin, 97*, 129–133.

American Psychological Association (2001). *Publication manual of the American Psychological Association* (5th ed.). New York: Author.

Barnett, V., & Lewis, T. (1978). *Outliers in statistical data*. Chicester, England: Wiley.

Berlin, J. A., & Rennie, D. (1999). Measuring the quality of trials. *Journal of the American Medical Association, 282,* 1083–1085.

Bushman, B. J. (1994). Vote-counting procedures in meta-analysis. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 193–213). New York: Russell Sage.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Cohen, J. (1988). *Statistical power analysis in the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field setting*s. Chicago: Rand McNally.

Cook, T. D., Cooper, H. M., Cordray, D. S., Hartmann, H., Hedges, L. V., Light, R. J., Louis, T., & Mosteller, F. (1992). *Meta-analysis for explanation: A casebook*. New York: Russell Sage.

Cooper, H. M. (1979). Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *Journal of Personality and Social Psychology, 37*, 131–146.

Cooper, H. M. (1982). Scientific guidelines for conducting integrative research reviews. *Review of Educational Research, 52*, 291–302.

Cooper, H. M. (1998). *Synthesizing research: A guide for literature reviews* (3rd ed.). Thousand Oaks, CA: Sage.

Cooper, H., Charlton, K., Valentine, J. V., & Muhlenbruck, L. (2000). Making the most of summer school: A meta-analytic and narrative review. *Monographs on Child Development, 65*(1). Malden, MA: Blackwell Press.

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., & Dorr, N. (1995). Race comparisons on need for achievement: A meta-analytic alternative to Graham's narrative review. *Review of Educational Research, 65*, 483.

Cooper, H., & Hedges, L. V. (1994). *Handbook of research synthesis.* New York: Russell Sage.

Cooper, H. M., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin, 87*, 442–449.

DuBois, D. L., Holloway, B. E., Valentine, J. C., & Cooper, H. (2002). Effectiveness of meta-analytic programs for youth: A meta-analytic review. *American Journal of Community Psychology, 30,* 157–197.

Duval S. J., & Tweedie R. L. (2000). A non-parametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*, 89–98.

Eddy, D. M., Hasselblad, V., & Schachter, R. (1992). *Meta-analysis by the confidence profile method.* Boston, MA: Academic Press.

Feldman, K. A. (1971). Using the work of others: Some observations on reviewing and integrating. *Sociology of Education, 4*, 86–102.

Fisher, R. A. (1932). *Statistical methods for research workers.* London: Oliver & Boyd.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3–8.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Glass, G. V., & Smith, M. L. (1979). Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis, 1*, 2–16.

Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. V. Hedges (Eds.). *Handbook of research synthesis.* New York: Russell Sage.

Greenberg, J., & Folger, R. (1988). *Controversial issues in social research methods.* New York: Springer-Verlag.

Greenhouse, J. B. & Iyengar, S. (1994). Sensitivity analysis and diagnostics. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 383–398). New York: Russell Sage.

Greenwald, A. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1–20.

Halvorsen, K. T. (1994). The reporting format. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 425–438). New York: Russell Sage.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed and random effects models in meta-analysis. *Psychological Methods, 3*, 486–504.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Beverly Hills, CA: 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. (1989). *DSTAT: Software for the meta-analytic review of research literatures.* Hillsdale, NJ. Erlbaum.

Jüni, P., Witschi, 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.

Light, R. J. (Ed.) (1983). *Evaluation studies review annual* (Vol. 8). Beverly Hills, CA: Russell Sage.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of research reviewing.* Cambridge, MA: Harvard University Press.

Light, R. J., Singer, J. D., & Willett, J. B. (1994). The visual presentation and interpretation of meta-analyses. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 439–453). New York: Russell Sage.

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. (1994). Identifying potentially interesting variables and analysis opportunities. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 111–123). New York: Russell Sage.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Matt, G. E. & Cook, T. D. (1994). Threats to the validity of research syntheses. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 503–520). New York: Russell Sage.

Miller, N., Lee, J., & 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., Schulz, K. F., & Altman, D. G. for the CONSORT Group (2001). The CONSORT Statement: Revised recommendations for improving the records of parallel-group randomised trials. *The Lancet, 357*, 1191–1194.

Olkin, I. (1992, July/August). Reconcilable differences: Gleaning insight from conflicting scientific studies. *The Sciences*, 30–36.

Orwin, R. G. (1994). Evaluating coding decisions. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 139–162). New York: 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

Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *British Medical Journal, 3*, 1243–1246.

Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin, 103*, 111–120.

Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Beverly Hills, CA: Sage.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw Hill.

Rosenthal, R., & Rubin, D. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences, 3*, 377–415.

Rosenthal, R., & Rubin, D. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin, 92*, 500–504.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.) (2005). *Publication bias in meta-analysis: Prevention, assessment, and adjustments*. Chichester, England: Wiley.

SAS Institute. (1992). *SAS user's guide: Statistics* (Version 6). Cary, NC: Author.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.). *Handbook of research synthesis* (pp. 261–282). New York: Russell Sage.

Steering Committee of the Physicians' Health Study Research Group (1988). Preliminary report: Findings from the aspirin component of the ongoing physicians' health study. *New England Journal of Medicine, 318,* 262–264.

Stock, W. A. (1994). Systematic coding for research synthesis. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 125–138). New York: Russell Sage.

Stock, W. A., Okun, M. A., Haring, M. J., Miller, W., Kinney, C., & Ceurvorst, R. W. (1982). Rigor in data synthesis: A case study of reliability in meta-analysis. *Educational Researcher, 11*(6), 10–14, 20.

Torgerson, C., Porthouse, J., & Brooks, G. (2005). A systematic review of controlled trials evaluating interventions in adult literacy and numeracy. *Journal of Research in Reading, 28*, 87–107.

Wachter, K. W., & Straf, M. L. (Eds.). (1990). *The future of meta-analysis*. New York: Russell Sage.

Weber, R. P. (1990). *Basic content analysis* (2nd ed.). Thousand Oaks, CA: Sage.