SAGE researchmethods

Research Synthesis and Meta-Analysis

In: The SAGE Handbook of Applied Social Research Methods

By: Harris M. Cooper, Erika A. Patall & James J. Lindsay
Edited by: Leonard Bickman & Debra J. Rog
Pub. Date: 2013
Access Date: June 20, 2017
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks
Print ISBN: 9781412950312
Online ISBN: 9781483348858
DOI: http://dx.doi.org/10.4135/9781483348858
Print pages: 344-370
©2009 SAGE Publications, Inc.. All Rights Reserved.
This PDF has been generated from SAGE Research Methods. Please note that the pagination of the online version will vary from the pagination of the print book.

©2009 SAGE Publications, Ltd.. All Rights Reserved.

Research Synthesis and Meta-Analysis

As the volume of primary research across all fields of social science continues to grow at rapid rates, research synthesis has become more important today than at any other time in history. With the development of metaanalysis, a set of procedures for summarizing the quantitative results from multiple studies, the rigor, systematicity, and transparency of research syntheses was greatly improved. However, a number of developments, including the creation of the Cochrane Collaboration and Campbell Collaboration, have heightened the profile of meta-analysis in recent years. Furthermore, recent advancements in analytic strategies, including the use of a random effects model of error, the development of meta-regression, and improved methods for dealing with missing data and data censoring, have enhanced the popularity, efficiency, and trustworthiness of meta-analyses.

We begin this chapter with a brief history of meta-analysis and research synthesis. We then describe the different stages of a rigorous research synthesis. Next, we outline a set of generally useful meta-analytic techniques and follow this with a discussion of some of the difficult decisions that research synthesists face in carrying out a meta-analysis. We conclude by addressing some broader issues concerning criteria for evaluating the quality of knowledge syntheses in general and meta-analyses in particular.

A general theme of the chapter is that social scientists who are conducting research syntheses need to think about what distinguishes a good synthesis from a bad synthesis. This kind of effort is crucial for assessing the value of existing research syntheses and for promoting high-quality research synthesis in the future.

A Brief History of Research Synthesis and Meta-Analysis

In 1904, Karl Pearson published what is believed to be the first meta-analysis. Having been asked to synthesize the evidence on a vaccine against typhoid, Pearson gathered data from 11 relevant studies, and for each study, he calculated a recently developed statistic called the correlation coefficient. He averaged these measures of the treatment's effect across two groups of studies distinguished by the nature of their outcome variable. Based on the average correlations, Pearson concluded that other vaccines were more effective.

In 1932, Ronald Fisher, in his classic text *Statistical Methods for Research Workers*, noted, “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 then presented a technique for combining the *p* values that came from statistically independent tests of the same hypothesis. His work would

©2009 SAGE Publications, Ltd.. All Rights Reserved.

be followed by more than a dozen papers published prior to 1960 on the same topic (see Olkin, 1990).

This early development of procedures for statistically combining results of independent studies went largely unused. However, beginning in the 1960s, with the tremendous growth in social scientific research and increasing interest in its social policy implications, these methods began to gain widespread use (Chalmers, Hedges, & Cooper, 2002). By the mid-1970s, when Robert Rosenthal and Donald Rubin undertook a synthesis of research studying the effects of interpersonal expectations on behavior, they found 345 studies that pertained to their hypothesis (Rosenthal & Rubin, 1978). Almost simultaneously, Gene Glass and Mary Lee Smith were conducting a synthesis of the relation between class size and academic achievement (Glass & Smith, 1979). They found 725 estimates of the relation, based on data from nearly 900,000 students. Smith and Glass (1977) also gathered assessments of the effectiveness of psychotherapy; this literature revealed 833 tests of the treatment. Likewise, John Hunter and Frank Schmidt uncovered 866 comparisons of the differential validity of employment tests for black and white workers (Hunter, Schmidt, & Hunter, 1979).

Each of these research teams realized that for some topic areas, prodigious amounts of empirical evidence had been amassed on why people act and feel the way they do and on the effectiveness of psychological, social, educational, and medical interventions. These researchers concluded that the traditional research synthesis simply would not suffice. Largely independently, the three research teams rediscovered and reinvented Pearson's and Fisher's solutions to their problem.

In discussing his solution, 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, other proponents of meta-analysis demonstrated that traditional synthesis procedures led to inaccurate or imprecise characterizations of the literature, even when the size of the literature was relatively small (Cooper, 1979; Cooper & Rosenthal, 1980).

The first half of the 1980s witnessed the appearance of five books devoted primarily to meta-analytic methods. The first, by 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. In 1982, Hunter, Schmidt, and Jackson 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).

©2009 SAGE Publications, Ltd.. All Rights Reserved.

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.

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

Finally, in 1985, with the publication of *Statistical Methods for Meta-Analysis*, Hedges and Olkin helped to elevate the quantitative synthesis of research to an independent specialty within the statistical sciences. This book, summarizing and expanding nearly a decade of programmatic developments by the authors, not only covered the widest array of meta-analytic procedures but also established their legitimacy by presenting rigorous statistical proofs.

Meta-analysis did not go uncriticized. Some critics opposed quantitative synthesis, using arguments similar to those used to oppose primary data analysis (Barber, 1978; Mansfield & Bussey, 1977). Others linked meta-analysis with more general synthesis procedures that are inappropriate, but not necessarily related to the use of statistics in synthesis. We address several of these issues later in this chapter.

Since the mid-1980s, several other books have appeared on meta-analysis. Some of these treat the topic generally (e.g., Cooper, 1998; Hunter & Schmidt, 2004; Lipsey & Wilson, 2001), some treat it from the perspective of particular research design conceptualizations (e.g., Eddy, Hassleblad, & Schachter, 1992; Mullen, 1989), some are tied to particular software packages (e.g., Johnson, 1993; Wang & Bushman, 1999), and some look to the future of research synthesis as a scientific endeavor (e.g., Cook et al., 1992; Wachter & Straf, 1990).

During and after the years that the works mentioned above were appearing, literally thousands of meta-analyses were published. In 1994, the first edition of *Handbook of Research Synthesis* was published (Cooper & Hedges, 1994). Through the 1990s, the use of meta-analysis spread from psychology and education (see Hunt, 1997, for a history of these efforts) through many disciplines, especially social policy analysis and the medical sciences (see Chalmers, Hedges, & Cooper, 2002, for a history of meta-analysis in medicine). One of the most notable events in medicine was the establishment of the U.K. Cochrane Center in 1992. The Center was meant to

©2009 SAGE Publications, Ltd.. All Rights Reserved.

facilitate the creation of an international network to prepare and maintain systematic synthesis of the effects of interventions across the spectrum of health care practices. At the end of 1993, an international network of individuals, called the Cochrane Collaboration (http://www.cochrane.org/index.htm), emerged from this initiative (Bero & Rennie, 1995; Chalmers, 1993). By 2006, the Cochrane Collaboration was an internationally renowned initiative with 11,000 people contributing to its work, in more than 90 countries. The Cochrane Collaboration is now the leading producer of research syntheses in health care and is considered by many to be the gold standard for determining the effectiveness of different health care interventions. Its library of systematic synthesis numbers in the thousands. In 2000, an initiative called the Campbell Collaboration (http://www.campbellcollaboration.org) was begun with similar objectives for the domain of social policy analysis, focusing initially on policies concerning education, social welfare, and crime and justice.

Research Synthesis as a Scientific Process

Several early attempts that framed the integrative research synthesis in the terms of a scientific process occurred independent of the meta-analysis movement. 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 maybe 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) 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 synthesis methods.

Three years later, Taveggia (1974) struck a complementary theme:

> A methodological principle overlooked by [synthesists] … is that research results are probabilistic … 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)

Taveggia described six common problems in literature syntheses; selecting research; retrieving, indexing, and coding studies; analyzing the comparability of findings; accumulating

©2009 SAGE Publications, Ltd.. All Rights Reserved.

comparable findings; analyzing the resulting distributions; and reporting the results.

Two articles that appeared in the *Synthesis 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). Jackson portrayed the limitations of meta-analysis as well as its strengths. His article employed a sample of 36 synthesis 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) took the analogy between research synthesis and primary research to its logical conclusion. He presented a five-stage model of the integrative synthesis that viewed research synthesis as a data-gathering exercise and, as such, applied to it criteria similar to those employed to judge primary research. Similar to primary research, a research synthesis 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 asked, its primary function in the synthesis, and the procedural differences that might cause variation in synthesis' conclusions. In addition, Cooper applied the notion of threats to inferential validity—introduced by Campbell and Stanley (1966; expanded by Cook and Campbell, 1979, and further refined in Shadish, Cook, & Campbell, 2002) for evaluating the utility of primary research designs—to research synthesis. He identified numerous threats to validity associated with synthesis procedures that might undermine the trustworthiness of a research synthesis' findings. He also suggested that other threats might exist and that any particular synthesis' validity could be threatened by consistent deficiencies in the set of studies that formed its database.

Table 11.1 presents Cooper's (1982) conceptualization of the research synthesis process. In the next section, we describe briefly the critical decisions that characterize each stage.

The Stages of Research Synthesis

The Problem Formulation Stage

During the problem formulation stage, research synthesists must (a) define the variables of interest both conceptually and operationally and (b) clearly state the relationship of interest.

Conceptual definitions describe qualities of the variables that are independent of time and space but can be used to distinguish relevant from irrelevant events (Shoemaker, Tankard, & Lasorsa, 2004). The first source of variation in synthesis conclusions enters during this concept

©2009 SAGE Publications, Ltd.. All Rights Reserved.

identification. Two synthesists using an identical label for an abstract concept can employ different definitions or levels of

Table 11.1 Research Synthesis Conceptualized as a Research Process

	Stage of Research				
Stage Characteristics	*Problem Formulation*	*Data Collection*	*Data Evaluation*	*Analysis and Interpretation*	*Public Presentation*
Research question asked	What evidence should be included in the review?	What procedures should be used to find relevant evidence?	What retrieved evidence should be included in the review?	What procedures should be used to make inferences about the literature as a whole?	What information should be included in the review report?
Primary function in review	Constructing definitions that distinguish relevant from irrelevant studies	Determining which sources of potentially relevant studies to examine	Applying criteria to separate "valid" from "invalid" studies	Synthesizing valid retrieved studies	Applying editorial criteria to separate important from unimportant information
Procedural differences that create variation in review conclusions	1. Differences in included operational definitions 2. Differences in operational detail	Differences in the research contained in sources of information	1. Differences in quality criteria 2. Differences in the influence of nonquality criteria	Differences in rules of inference	Differences in guidelines for editorial judgment
Sources of potential invalidity in review conclusions	1. Narrow concepts might make review conclusions less definitive and robust. 2. Superficial operational detail might obscure interacting variables.	1. Accessed studies might be qualitatively different from the target population of studies. 2. People sampled in accessible studies might be different from target population.	1. Nonquality factors might cause improper weighting of study. 2. Omissions in study reports might make conclusions unreliable.	1. Rules for distinguishing patterns from noise might be inappropriate. 2. Review-based evidence might be used to infer causality.	1. Omission of review procedures might make conclusions irreproducible. 2. Omissions of review findings and study procedures might make conclusions obsolete.

SOURCE: From *Synthesizing Research: A Guide for Literature Synthesis,* 3rd ed., by H. M. Cooper, 1998. Reprinted with pemission of SAGE.

abstraction. That is, conceptual definitions can differ in breadth, or in the number of events to which they refer. Let's take as an example the concept of homework. One synthesist may consider as homework only assignments meant to have students practice what they have learned in class, whereas another may include assignments to visit museums or to watch certain television programs. In such a case, the second synthesist employs a broader conception of homework, and this synthesis will likely contain more research than will the first.

As in primary research, in order to relate concepts to concrete events, the variables of interest in a research synthesis also must be operationally defined. An operational definition provides a description of the characteristics of observable events that are used to determine whether the event represents an occurrence of the conceptual variable. Synthesists can also vary in the way operations are treated *after* the relevant research has been retrieved. Thus, synthesists who employ identical conceptual definitions of homework and who include the same set of studies can still reach decidedly different conclusions if one synthesist retrieved more information about the features of studies and recognized a relation between a study feature and outcome that the

©2009 SAGE Publications, Ltd.. All Rights Reserved.

other synthesist did not test. One synthesist might discover that the outcomes of homework studies depended on whether textbook or teacher-developed tests were used to assess impact, whereas another synthesist never even coded studies based on this feature of the outcome measure.

Each difference in how a problem is formulated introduces a potential threat to the trustworthiness of a synthesis' conclusions. First, synthesists who focus on very narrow conceptualizations provide little information about how many different contexts a finding applies to. Therefore, synthesists who employ broad conceptual definitions can *potentially* produce more valid conclusions than ones using narrow definitions. However, broad definitions can lead to the erroneous conclusion that research results are insensitive to variations in a study's context. We can assume, therefore, that synthesists who examine more operational details within their broader constructs will produce more trustworthy conclusions. These synthesists present more information about contextual variations that do and do not influence the synthesis outcome.

The Literature Search Stage

The decisions a synthesist makes during the literature search determine the nature of studies that will ultimately form the basis for conclusions. Identifying populations for research syntheses is complicated by the fact that syntheses involve two targets. First, a synthesist wants the findings to reflect the results of *all previous research* on the problem. The synthesist can exert some control over whether this goal is achieved through their choice of information sources. Second, the synthesist hopes that the included studies will allow generalizations to the *individuals or other units that interest researchers in the topic area.* The synthesist's influence is constrained at this point by the types of individuals or units who were sampled by the primary researchers. Thus, a synthesis of the homework research first should include as many of the previous studies as the synthesist can find, and it is hoped that these studies will include all the types of students for whom homework is a relevant issue.

Some discrepancies in synthesis conclusions are created by differences in the sources synthesists use to retrieve studies, such as journal networks, reference databases, listservs, and personal communications. The studies available through different sources are often different from one another. The first concern with the literature search is that the synthesis may not include, and probably will not include, all studies pertinent to the topic of interest. Synthesists who have used the broadest sources of information are most likely to retrieve a set of results that resembles the entire population of previous research. However, methodologists do differ in their opinions about how exhaustive a literature search needs to be, especially as it

©2009 SAGE Publications, Ltd.. All Rights Reserved.

pertains to the inclusion of unpublished research. We take up this debate in the following sections.

The second concern that arises during the literature search is that the participants or other units in the retrieved studies may not represent all units in the target population. For instance, it may be that little or no research has been conducted that examines the effects of homework on first- or second-grade students. The synthesist cannot be faulted for the existence of this gap *if* the retrieval procedures used were exhaustive. However, synthesists who qualify conclusions with information about the kinds of units missing or overrepresented in studies probably run less risk of making overly broad generalizations.

The Data Evaluation Stage

After the literature is collected, the synthesist makes critical judgments about the quality of individual studies. Each study is examined to determine whether it is contaminated by factors irrelevant to the problem under consideration. Then, trained personnel use standardized coding procedures to extract the desired information from research reports.

Differences in syntheses are created by differences in synthesists' criteria for evaluating the quality of research. Just how this evaluation ought to proceed is another source of disagreement among researchers that we will address more fully below. Relatedly, variation in conclusions is created when factors other than research quality affect synthesists' decisions, for example, the reputation or institution of the primary researchers, or the research findings. The use of any criteria other than methodological quality ought to be considered a threat to the validity of a research synthesis (e.g., Mahoney, 1977).

A second threat to trustworthiness during research evaluation is completely beyond the control of the synthesist. This threat involves incomplete reporting by primary researchers. If the synthesist must estimate or omit what happened in these studies, wider confidence intervals must be placed around synthesis conclusions. We will examine some solutions to the problem of missing data below.

A third threat to the validity of conclusions drawn from a synthesis can result because synthesists are not immune to making mistakes themselves in coding information from reports. To address this issue, it is recommended that two or more synthesists code either all or a subset of studies in the synthesis. The extent to which study information has been reliably extracted from research reports can then be assessed by 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). A second

©2009 SAGE Publications, Ltd.. All Rights Reserved.

strategy involves having two synthesists independently code all the studies in the synthesis. Disagreements then may be resolved in conference or by a third reader. This procedure raises the effective reliability of codes to very high levels.

The Analysis and Interpretation Stage

During analysis and interpretation, the separate research reports collected by the synthesist are integrated into a unified statement about the research problem. It is at this stage that the synthesist must decide whether or not to use meta-analysis. Synthesis conclusions can differ because synthesists employ different analytic interpretation techniques. A systematic relation that cannot be distinguished from noise under one set of rules may be discernible under another set.

One source of concern during the analysis and interpretation of studies involves the rules of inference employed by the synthesist. In nonquantitative syntheses, it is difficult to gauge the appropriateness of inference rules because they are not very often made explicit. For meta-analyses, the suppositions of statistical tests are generally known, and some statistical biases can be removed. Regardless of the strategy used for analysis and interpretation, the possibility always exists that the synthesist has used an invalid rule for inferring a characteristic of the target population. For this reason, the number of primary studies available, the degree of statistical detail presented in research reports, and the frequency of methodological replications need to be assessed before determining whether to perform a meta-analysis.

Meta-analysis should be the default option when the goal of a synthesis is to summarize a research literature for purposes of making a general statement about the support for, or size of, a relationship between variables. However, there are some instances in which the use of meta-analysis might be less appropriate, or perhaps completely unnecessary. First, meta-analysis is improper if the goal of the synthesis is to critically appraise a research literature study-by-study or to identify particular studies central to a field. Second, meta-analysis maybe inappropriate in cases where conceptual and methodological approaches to research on a topic have changed over time. Third, under certain conditions meta-analysis might not lead to the kinds of generalizations the synthesist wishes to make. Under these circumstances, the synthesist might convincingly establish the generalization of a finding using conceptual and theoretical bridges rather than statistical ones. Finally, even if the synthesist wishes to summate statistical results across studies on the same topic, the studies might have been conducted using decidedly different methodologies, participants, and outcome measures. In such cases, statistical combinations might mask important differences in research findings. In these instances, it may make the most sense not to use meta-analysis, or to conduct several discrete

©2009 SAGE Publications, Ltd.. All Rights Reserved.

meta-analyses within the same synthesis. Regardless of the technique used to analyze and integrate the results of individual studies, all research synthesists should provide justification for their methods and ensure that the synthesis techniques employed are transparent to the reader.

A second concern involves the misinterpretation of synthesis-based evidence as supporting statements about causality. For example, it might be that a study finding a larger-than-normal effect of homework on achievement was conducted at an upper-income school. However, it might also be the case, known or unknown to the synthesist, that this study used unusually long homework assignments. The synthesist cannot discern, therefore, which characteristic of the study, if either, produced the larger effect. Thus, when different study characteristics are found associated with the effects of a treatment, the synthesist should recommend that future researchers examine these factors within a single experiment.

The Public Presentation Stage

Finally, the production of a document describing the synthesis is a task with important implications for the accumulation of knowledge. Two threats to validity accompany report writing. First, the omission of details about how the synthesis was conducted reduces the possibility that others can replicate the conclusions. The second threat involves the omission of evidence that others find important. A synthesis will quickly become obsolete if it does not address the variables and relations that are (or will be) important to an area.

The Elements of Meta-Analysis

Suppose a research synthesist is interested in whether fear-arousing advertisements can be used to persuade adolescents that smoking is bad. Suppose further that the (hypothetical) synthesist is able to locate eight studies, each of which examined the question of interest. Of these, six studies reported nonsignificant differences between attitudes of adolescents exposed and not exposed to fear-arousing ads and two reported significant differences indicating less favorable attitudes held by adolescent viewers. One was significant at $p < .05$ and one at $p < .02$ (both two-tailed). Can the synthesist reject the null hypothesis that the ads had no effect?

There are multiple methods a research synthesist could employ to answer this question. First, the synthesist could cull through the eight reports, isolate those studies that present results counter to their own position, discard these disconfirming studies due to methodological limitations, and present the remaining supportive studies as presenting the truth of the matter. Such a research synthesis would be viewed with extreme skepticism. It would contribute little to answering the question.

©2009 SAGE Publications, Ltd.. All Rights Reserved.

The Vote Count

As an alternative procedure, the synthesist could take each report and place it into one of the three piles: statistically significant findings that indicate that ads were effective, statistically significant findings that indicate that the ads created more positive attitudes toward smoking (in this case, the pile would have no studies), and nonsignificant findings that do not permit rejection of the hypothesis that the fear-arousing ads had no effect. The synthesist then would declare the largest pile the winner. In our example, the null hypothesis wins.

This vote count of significant findings has much intuitive appeal and has been used quite often. However, the strategy is unacceptably conservative. The problem is that chance alone should produce only about 5% of all reports falsely indicating that viewing the ads created more negative attitudes toward smoking. Therefore, depending on the number of studies, 10% or less of positive and statistically significant findings might indicate a real difference due to the ads. However, the vote-counting strategy requires that a minimum 34% of findings be positive and statistically significant before the hypothesis is ruled a winner. Thus, the vote counting of significant findings could, and often does, lead to the suggested abandonment of hypotheses (and effective treatment programs) when, in fact, no such conclusion is warranted.

Hedges and Olkin (1980) describe a different way to perform vote counts in research synthesis. This procedure involves (a) counting the number of positive and negative results, regardless of significance, and (b) applying the sign test to determine if one direction appears in the literature more often than would be expected by chance. This vote-count method has the advantage of using all studies but suffers because it does not weight a study's contribution by its sample size. Thus, a study with 100 participants is given weight equal to a study with 1,000 participants. This is a potential problem because large samples are likely to provide more precise answers to questions. Therefore, results from larger samples should be given more weight. Furthermore, the revealed magnitude of the hypothesized relation (or impact of the treatment under evaluation) in each study is not considered—a study showing a small positive attitude change is given equal weight to a study showing a large negative attitude change. 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 effect size estimate (see Bushman, 1994; Hedges & Olkin, 1985).

Estimating Effect Sizes

While vote-counting addresses the question of whether or not an effect exists; it gives no information about whether that effect is large or small, important or trivial. Therefore, the question of greatest importance is often not “Do fear-arousing ads create more negative

©2009 SAGE Publications, Ltd.. All Rights Reserved.

attitudes toward smoking in adolescents, yes or no?" Instead, the question should be "How much of an effect do fear-arousing ads have?" The answer might be zero or it might be either a positive or a negative value. Furthermore, the synthesist is likely interested in what factors influence the effect of fear-arousing ads. Given these new questions, the synthesist would turn to the calculation of average effect sizes.

Cohen (1988) has 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). In meta-analysis, effect sizes are (a) calculated for the outcomes of studies (or sometimes comparisons within studies), (b) averaged across studies to estimate general magnitudes of effect, and (c) compared between studies to discover if variations in study outcomes exist and, if so, what features of studies might account for them.

Although numerous estimates of effect size are available, three dominate the literature. The first, called the *d*-index by Cohen (1988; also see Hedges & Olkin, 1985; Rosenthal, 1994), is a scale-free measure of the separation between two group means. Calculating the *d*-index for any study involves dividing the difference between the two group means by either their average standard deviation or the standard deviation of the control group. For example, Cooper, Robinson, and Patall (2006) examined the difference in academic achievement of students who did and did not do homework. Across five studies that manipulated the presence of homework, the average *d*-index was 0.60 favoring the homework doers. Thus, the average academic achievement of students who did homework was 0.60 standard deviations above the average score of students who did not.

Figure 11.1 presents the *d*-indices associated with three hypothetical studies. In Figure 11.1a, the fear-arousing ad has no effect on adolescents' reported attitudes toward smoking, thus $d = 0$. In Figure 11.1b, the average adolescent viewing the ad has an attitude score that is four tenths of a standard deviation more negative than the average adolescent viewing control ads. Here, $d = 0.40$. In Figure 11.1c, d = 0.85, indicating an even greater separation between the two group means.

In many instances, synthesists will find that primary researchers do not report the means and standard deviations of the separate groups. For such cases, meta-analysts can use one of a number of computational formulas that do not require means and standard deviations. The interested reader may refer to Rosenthal (1994) or Lipsey and Wilson's (2001) for listings of algebraically equivalent formulas that can be used to compute an effect size from various statistical information.

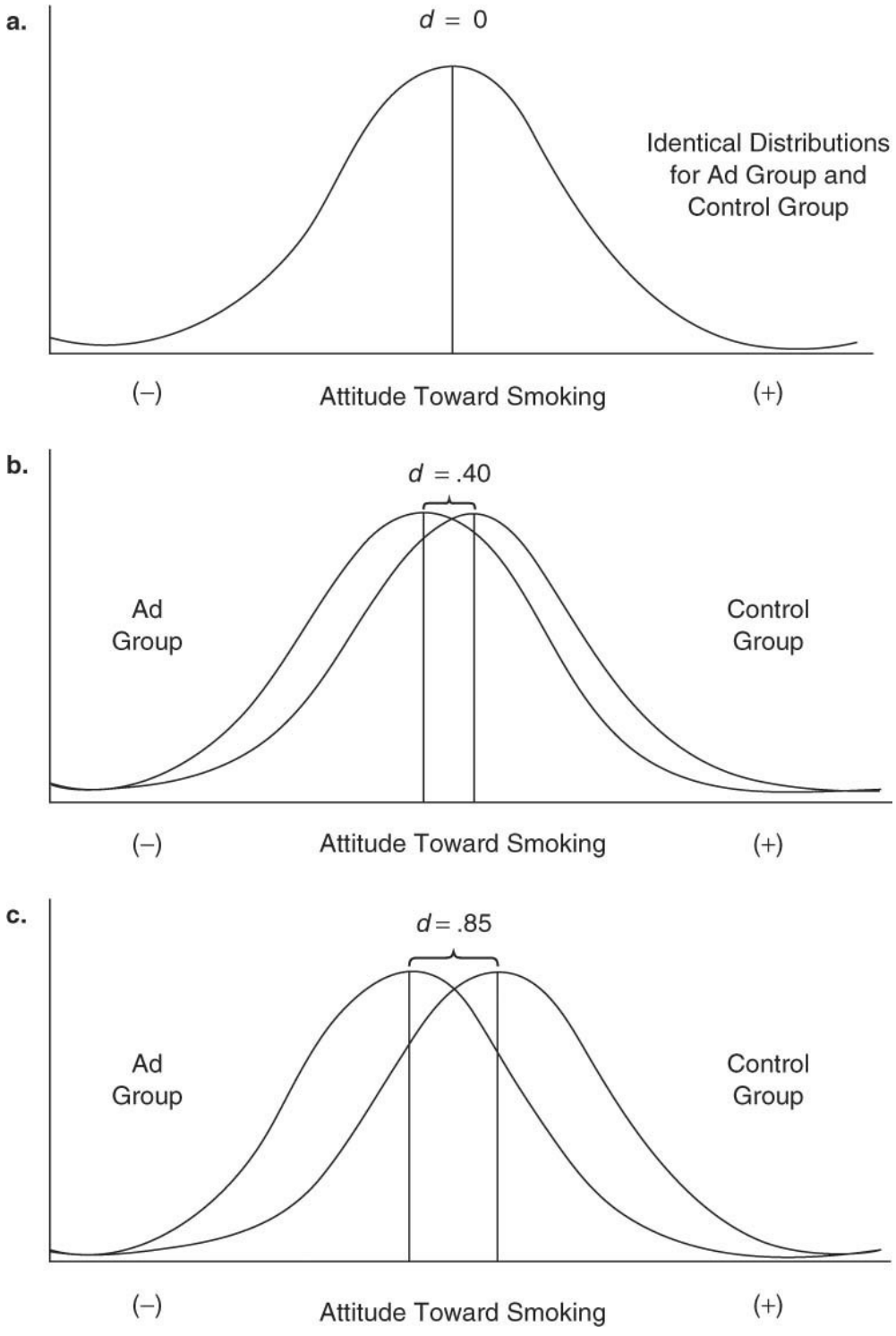
©2009 SAGE Publications, Ltd.. All Rights Reserved.

Another effect size metric is the *r*-index, or the Pearson product-moment correlation coefficient. Typically, it is used to measure the degree of linear relation between two variables. The correlation coefficient is familiar to most researchers and is most appropriate when describing the relationship between two continuous variables. For example, Cooper and colleagues (2006) found 32 studies that described the correlations between the time a student spent on homework and a measure of academic achievement. The average correlation for the 32 studies was $r = 0.24$, suggesting that more time spent on homework is related to greater academic achievement.

The third effect size metric is the odds ratio. The odds ratio is applicable when both variables are dichotomous and findings are presented as frequencies or proportions. This measure of effect is used most in medical sciences, in which the researcher is often interested in the effect of a treatment on mortality or the appearance or disappearance of disease. It also appears frequently in studies of educational interventions when the outcome of interest is drop-out or retention rates or criminal justice studies where the outcome is recidivism. For example, if the synthesist was interested in whether exposure to fear-arousing ads led adolescents to continue or quit smoking, then an odds ratio would be an appropriate effect size metric. First, the odds of smoking must be determined for each condition, when participants are exposed to fear-arousing advertisements versus control advertisements. Then, the ratio of the odds for being exposed to fear-arousing advertisements over control advertisements is then calculated as the ratio of the odds.

©2009 SAGE Publications, Ltd.. All Rights Reserved.

Figure 11.1 Three Relations Between Fear-Arousing Ads and Attitudes Toward Smoking Expressed by the *d*-Index



Averaging Effect Sizes and Measuring Dispersion

The most pivotal outcomes of meta-analyses are the average effect sizes and measures of dispersion that accompany them. State-of-the-art meta-analytic procedures call for the

©2009 SAGE Publications, Ltd.. All Rights Reserved.

weighting of effect sizes when they are averaged across studies. In the weighted 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 because it gives greater weight to effect sizes based on larger samples, and larger samples give more precise population estimates. Confidence intervals are then calculated to test the null hypothesis that the difference between two means, or the size of a correlation or odds ratio, is zero (Hedges, Cooper, & Bushman, 1992). Going back to the meta-analysis conducted by Cooper and colleagues (2006) looking at the effect of homework on academic achievement, the average *d*-index was 0.60 favoring homework doers, with a 95% confidence interval of 0.38 to 0.82. This confidence interval suggests that the effect of homework on achievement was significantly different from zero. 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, meta-analysts usually carry out *homogeneity analyses.* Homogeneity analyses allow the meta-analyst to explore whether effect sizes vary from one study to the next. 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 and provides 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. If there is greater variation in effects than would be expected by chance, then the meta-analyst can begin the process of examining moderators of comparison outcomes. For example, in Cooper and colleagues meta-analysis on the effect of homework, the test of homogeneity revealed that the effect sizes were not significant, suggesting that the meta-analyst cannot reject the hypothesis that the effects from different studies are estimating the same underlying population value. In the case in which the observed variance is not significantly different from that expected by sampling error alone, many statisticians advise that meta-analysts stop the analysis there and not look for moderators. After all, chance is the most parsimonious explanation for the variation in effect sizes. Others suggest that meta-analysts may search for moderators in the absence of a statistically significant homogeneity analysis if there are good theoretical or practical reasons for doing so.

An alternative approach to examining if effect sizes vary across studies also compares the observed variation in obtained effect sizes with the variation expected due to sampling error, that is, the expected variance in effect sizes given that all observed effects are estimating the same underlying population value (Hunter & Schmidt, 2004). However, a formal statistical test of the difference between these two values is typically not carried out. Rather, meta-analysts

©2009 SAGE Publications, Ltd.. All Rights Reserved.

adopt a critical value for the ratio of observed-to-expected variance to use as a means for rejecting the null hypothesis. In this approach, meta-analysts might also adjust effect sizes to account for methodological artifacts such as sampling error, range restrictions, or unreliability of measurements. This method has been applied most often in the areas of industrial and organizational psychology.

Moderator Analyses

Another advantage of performing a statistical integration of research is that it allows synthesists to test hypotheses about why the outcomes of studies differ. To continue with the fear-arousing ad example, the synthesist might calculate average *d*-indexes for subsets of studies, deciding that he or she wants different estimates based on certain characteristics of the data. For example, the synthesist might want to compare separate estimates for studies that use different outcomes, distinguishing between those that measured likelihood of smoking and those that measured attitude toward smoking. Or, the synthesist might wish to compare the average effect sizes for different media formats, distinguishing print from video advertisements. Or, the synthesist might want to look at whether advertisements are differentially effective for males and females.

The ability to ask these questions about variables that moderate effects reveals one of the major contributions of research synthesis. Specifically, even if no individual study has compared different outcomes, media, or adolescent sexes, by comparing results across studies the synthesist can get a first hint about whether these variables would be important to look at in future research and/or as guides to policy.

Without the aid of statistics, the synthesist simply examines the differences in outcomes across studies, groups them informally by study features, and decides (based on an “interocular inference test”) whether the feature is a significant predictor of variation in outcomes. At best, this method is imprecise. At worst, it leads to incorrect inferences. In contrast, meta-analysis provides a formal means for testing whether different features of studies explain variation in their outcomes. After calculating the average effect sizes for different subgroups of studies, the synthesist can statistically test whether these factors are reliably associated with different magnitudes of effect also using homogeneity analyses. As previously suggested, homogeneity analysis allows meta-analysts to test whether sampling error alone accounts for variation in effect sizes or whether features of studies, samples, treatment designs, or outcome measures also play a role. This test is analogous to conducting an analysis of variance, in that a significant homogeneity statistic indicates that at least one group mean differs from the others. It is relatively simple to carry out a homogeneity analysis; formulas are described in Cooper

©2009 SAGE Publications, Ltd.. All Rights Reserved.

(1998), Cooper and Hedges (1994), Hedges and Olkin (1985), and Lipsey and Wilson (2001).

An alternative strategy for examining whether particular characteristics of studies are related to the sizes of the treatment effect is meta-regression. Unlike the strategy previously discussed, meta-regression allows the meta-analyst to explore the relationship between continuous, as well as categorical, characteristics and effect size, and allows the effects of multiple factors to be investigated simultaneously (Thompson & Higgins, 2002). In our example, imagine that our studies ranged in the duration of exposure to fear-arousing ads. One option would be to group studies into several distinct categories of duration of exposure to fear-arousing ads and continue with subgroup moderator analyses as previously discussed. However, an alternative would be to employ meta-regression, leaving this characteristic continuous. The interested reader may refer to Thompson and Higgins (2002) or Higgins and Thompson (2004) for a full discussion of this method.

In sum, a generic meta-analysis might contain three or four separate sets of statistics: (a) a frequency analysis of positive and negative results, (b) estimates of average effect sizes with confidence intervals, (c) homogeneity analyses to assess dispersion and examine study features that might influence study outcomes, and possibly, (d) regression coefficients if meta-regression is used to examine the relationship between continuous study characteristics and effect size. The need for vote counts diminish as the body of literature grows or if the synthesist provides confidence intervals around effect size estimates.

Difficult Decisions in Research Synthesis

When conducting primary research, investigators encounter decision points at which they have multiple choices about how to proceed. The same is true when conducting research syntheses. Some of these decisions will be easy to make, with choices being dictated by topic area considerations and the nature of the research base. Other decisions will be less clear. Six choice points have been generally perplexing for research synthesists. One occurs during data collection, two during data evaluation, and three during data analysis. These involve (a) how exhaustive the literature search should be, (b) what rules should be used for including or excluding studies from syntheses, (c) how to handle data missing from research reports, (d) how to determine whether separate tests of hypotheses are actually independent of one another, (e) how to decide what model of error underlies the generation of study outcomes, and (f) how to synthesize slopes from multiple regression.

Publish or Perish

Research synthesists disagree about how exhaustive a literature search needs to be. Some

©2009 SAGE Publications, Ltd.. All Rights Reserved.

synthesists go to great lengths to locate as much relevant material as possible; others are less thorough. Typically, disagreement centers on the importance of including unpublished research in syntheses.

Those in favor of limiting syntheses to only published material argue that publication is an important screening device for maintaining quality control. Because published research has been reviewed for quality, it provides the best evidence available. Also, the inclusion of unpublished material typically does not change the conclusions drawn by synthesists. Therefore, the studies found in unpublished sources do not warrant the additional time and effort needed to obtain them.

Those who argue that research should not be judged based on publication status give three rationales. First, they dispute the claim that published research and unpublished research yield similar results; statistically significant results are more likely to be published (Begg, 1994). That is, studies revealing smaller effects maybe systematically censured from the published literature, making relationships appear stronger than if all estimates were retrieved (Rothstein, Sutton, & Borenstein, 2005). Lipsey and Wilson (1993) compared the magnitudes of effects reported in published versus unpublished studies contained in 92 different research syntheses. They reported that the impacts of interventions in unpublished research were, on average, one third smaller than published effects.

Second, even if publication status does relate to the quality of research, there will still be much overlap in the quality of published and unpublished studies. Superior studies sometimes are not submitted or are turned down for publication for other reasons. Inferior studies sometimes find their way into print. Application of the “publish or perish” rule may lead to the omission of numerous high-quality studies and will not ensure that only high-quality studies are included in the synthesis.

And finally, in a meta-analysis, both the reliability of effect size estimates, expressed through the size of confidence intervals, and tests for effect size moderators will depend on the amount of available data. Therefore, synthesists may unnecessarily impede their ability to make confident statistical inferences by excluding unpublished studies (Rothstein et al., 2005).

Consequently, it is accepted practice that rigorous research syntheses should always access multiple channels to retrieve studies and operate with the goal of obtaining all relevant research (Cooper, 1998; Lipsey & Wilson, 2001), regardless of whether or where it was published. If the synthesis includes only published research it must be accompanied by a convincing justification.

©2009 SAGE Publications, Ltd.. All Rights Reserved.

Judging the Quality of Primary Studies

Another area of controversy in meta-analysis is related to the publication issue. All research synthesists agree that the quality of a study should dictate how heavily it is weighted when inferences are drawn about a research literature. However, there is disagreement about whether studies should be excluded from syntheses entirely if they are flawed.

Proponents of excluding flawed studies often employ the "garbage in, garbage out" axiom (Eysenck, 1978). They argue that amassing numerous flawed studies cannot replace the need for better-designed ones. Others argue that synthesists should employ the principle of best evidence used in law. This principle argues that "the same evidence that would be essential in one case might be disregarded in another because in the second case there is better evidence available" (Slavin, 1986, p. 6). Thus, a synthesist evaluates the entire literature and then bases decisions on only those studies that are highest in quality, even if these are not ideal.

Opponents of excluding studies contend that flawed studies can, in fact, accumulate to valid inferences. This might happen if the studies do not share the same design flaws but do come to the same result. Furthermore, global decisions about what makes a study good or bad are fraught with difficulty. There is ample evidence that even the most sophisticated researchers can disagree about the dimensions that define quality and how these dimensions apply to particular studies (see Valentine & Cooper, 2005). And finally, opponents of exclusion contend that the effect of research design on study outcomes is an empirical question. Rather than leaving studies out based on disputable, global judgments of rigor, synthesists can examine the operational details of studies empirically for their relation to outcomes. That is, synthesists code study features are known to vary with the strength of inferences they permit (e.g., research design, sampling frame, measurement reliability), and determine if those features covary with effect sizes uncovered by different studies (Berlin & Rennie, 1999; Jüni, Witschi, Bloch, & Egger, 1999). Then, if studies with more desirable features produce results different from other studies, inferences about the literature can be adjusted accordingly (Lipsey & Wilson, 2001).

How to Handle Missing Data

Missing data constitute one of the most frustrating problems faced by research synthesists. Missing data can take two forms. First, the synthesist may miss entire research reports that are pertinent to the topic or that he or she knows about but cannot retrieve. The above discussion of publication bias is relevant to this issue. Second, there may be data missing from the reports themselves. Within a report, missing information might include (a) the magnitude of the effect size (because it is not reported and not enough information is given for the meta-analysts to

©2009 SAGE Publications, Ltd.. All Rights Reserved.

calculate it) and/or (b) important study characteristics that might be tested as moderators of study outcomes. When data are missing, not only is the size of the sample gathered for the research synthesis reduced but the representativeness of the sample and the validity of the results are compromised, regardless of the quality of the meta-analysis in all other respects (Rothstein et al., 2005).

There are a number of strategies that meta-analysts can use to deal with missing data and data censoring. Rothstein et al. (2005) provide an in-depth treatment of numerous approaches. A number of graphical and statistical tests can be used to assess the possible presence of missing data and data censoring, and the implications of this threat to the validity of the conclusions drawn from the meta-analysis. Techniques include regression methods such as the rank correlation test (Begg & Mazumdar, 1994) and Egger's Test (Egger, Davey Smith, Schneider, & Minder, 1997), funnel plots (Light & Pillemer, 1984), as well as the Trim-and-Fill method (Duval & Tweedie, 2000a, 2000b). Furthermore, strategies for handling missing data within reports have been proposed. Some are simple. These include (a) omitting the cases with data missing from a given analysis or from the meta-analysis entirely, (b) assuming that missing values are equivalent to a very conservative estimate, such as zero, or (c) replacing missing values with the mean value calculated from available cases for that variable. Alternatively, missing data points can be estimated using single-value imputation procedures. More complex methods using multiple imputation procedures (Rubin, 1987) involve the employment of maximum likelihood models, though these methods are not widely used in meta-analysis. Details of these procedures are given by Pigott (1994). Regardless of which method is employed, meta-analysts are obligated to discuss the possibility and impact of missing entire reports and data censoring on the conclusion of the meta-analysis, how much data were missing within reports included in the synthesis, how they handled the situation, and why they chose the methods they did. Furthermore, it is becoming increasingly common practice for meta-analysts with much missing data to conduct their analyses using more than one strategy and determining whether their findings are robust across different missing data assumptions (see Greenhouse & Iyengar, 1994).

Finally, prospective registration and prospective meta-analysis have been recently proposed as two strategies which, if widely adopted, would decrease the occurrence of missing data and minimize publication bias (Berlin & Ghersi, 2005). Prospective registration entails registering a study on its inception when the researcher receives ethical or funding approval, allowing both the description of the study as well as eventual results to be publicly available. This would create an unbiased compilation of studies for subsequent meta-analyses and allow the synthesist to obtain information and results about studies regardless of the significance of their

©2009 SAGE Publications, Ltd.. All Rights Reserved.

findings or publication status. In prospective meta-analysis, studies are identified and determined to be eligible before the results of any of the studies are known. Prospective meta-analysis may be accomplished when multiple groups of investigators agree to combine their findings on completion. Furthermore, the comparability of research included in the meta-analysis is improved when investigators also decide prospectively to employ the same methods and assessment instruments across studies. Because the studies and specific analyses to be included in the meta-analysis are determined prior to any single study being conducted, missing data and data censoring is virtually eliminated.

Identifying Independent Hypothesis Tests

Meta-analysts must make decisions concerning how to handle multiple effect sizes coming from the same study. These effect sizes may share method variance that make them nonindependent data points. The problem this creates is that the assumption that effects are independent underlies the meta-analysis procedures described above.

Sometimes, a single study can contain multiple estimates of the same relation because (a) more than one measure of the same construct is used and the measures are analyzed separately or (b) results are reported separately for different samples of people. Taken a step further, synthesists also might conclude that the separate but related studies in the same report, or multiple reports from the same laboratory, are not independent.

Meta-analysts employ multiple approaches to handling nonindependent tests. Some treat each effect size as independent, regardless of the number that come from the same study. The strength of this technique is that it does not lose any of the within-study information regarding potential moderators. However, this strategy violates the assumption that the estimates are independent. Furthermore, the results of studies will not be weighted equally in any overall conclusion about results. Rather, studies will contribute to the overall effect in relation to the number of statistical tests contained in it.

Others use the study as the unit of analysis. In this strategy, a mean or median result is calculated to represent the study. This strategy ensures that the assumption of independence is not violated and that each study contributes equally to the overall effect. However, some within-study information may be lost in this approach.

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 still rarely found in practice.

©2009 SAGE Publications, Ltd.. All Rights Reserved.

Other meta-analysts suggest a shifting unit (Cooper, 1998). Here, each study is allowed to contribute as many effects as there are categories in the given analysis, but effects within any category are averaged. For example, if a study on whether fear-arousing advertisements promotes change in smoking behavior by adolescents used two different measures, one attitudinal and one behavioral, two separate *d*-indexes would be calculated. In the shifting unit of analysis approach, for estimating the overall relation between exposure to fear-arousing ads and smoking, statistical independence would be maintained by averaging these two *d*-indexes prior to entry into the analysis, so that the study only contributes one effect size. However, in an analysis that examined the effect of measurement characteristics, attitudinal or behavioral, on smoking outcomes, this sample would contribute one estimate to each category in the moderator analysis. This shifting unit of analysis approach retains as much data as possible from each study while holding to a minimum any violations of the assumption that data points are independent.

Models of Error

Another aspect of conducting a meta-analysis that has recently received considerable attention involves the decision about whether a fixed-effects or random-effects model of error underlies the generation of study outcomes. In a fixed-effects model, all studies are assumed to be drawn from a common population. As such, variance in effect sizes is assumed to reflect only sampling error, that is, error solely due to participant differences. However, sometimes other features of studies can be viewed as random influences. For example, studies that look at the impact of fear-arousing advertisements on smoking might vary in the length of exposure to ads or in how the ads are introduced to participants. In this case, it may be most appropriate to consider advertisements as randomly sampled from all fear-arousing advertisements. That is, in a random-effect analysis, study-level variance is assumed to be present as an additional source of random influence.

The question meta-analysts must ask is whether the effect sizes in their data set are affected by a large number of these study-level random influences. If it is the case that the meta-analysts suspect a large number of these additional sources of random error, then a random-effects model is most appropriate to take these sources of variance into account. If the meta-analysts suspects that the data are most likely little affected by other sources of random variance, then a fixed-effects model can be applied. Alternatively, 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 consideration is that in the search for moderators, fixed-

©2009 SAGE Publications, Ltd.. All Rights Reserved.

effect models may seriously underestimate error variance and random-effects models may seriously overestimate error variance when their assumptions are violated (Overton, 1998).

In view of these competing sets of concerns, the meta-analysts might consider applying both models (e.g., Cooper et al., 2006). 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.

Calculating random-effects estimates of the mean effect size, confidence intervals, and homogeneity statistics are complex and involve a two-stage process. As such, the interested reader should refer to Hedges and Olkin (1985), Raudenbush (1994), and Lipsey and Wilson (2001) for a full discussion of random-effects computation. In addition, several statistical packages have recently been developed specifically for meta-analysis that allow the meta-analysts to easily conduct analyses using both fixed-effects and random-effects assumptions (e.g., Borenstein, Hedges, Higgins, & Rothstein, 2005).

Combining Slopes from Multiple Regressions

Up to this point, the procedures for combining and comparing study results have generally assumed that the measure of effect is a mean difference, correlation, or odds ratio. However, regression analysis is a commonly used technique in the social sciences, particularly for nonexperimental studies. Like 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 used in meta-analyses because, like the d-index and r-index, it standardizes effect size estimates when different measures are used in different studies. β represents the standardized score change in a predictor variable, controlling for all other predictors, given one unit change in the criterion variable.

Syntheses of regression analyses are difficult to conduct for a variety of reasons. First, models using multiple regression generally differ from study to study. 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 (Wu & Becker, 2004). Second, the scale of the predictor of interest and outcome may vary across studies. This problem can be overcome by using β, the fully standardized estimate of the slope for a particular predictor. “Half-standardizing” is an alternative way to create similar slopes when only outcomes are dissimilar (Greenwald, Hedges, & Laine, 1996).

If slopes are independently and identically distributed, we can apply standard methods for

©2009 SAGE Publications, Ltd.. All Rights Reserved.

meta-analysis. Slopes will be identically distributed across studies when the outcome and predictor of interest are measured in a similar fashion, the other predictors in the model are the same across studies, and when predictor and outcome scores are similarly distributed (Becker, 2005). However, it is rare that data sets meet the assumption of being identically and independently distributed. Typically, measures differ across studies and regression models are diverse in terms of which additional variables are included in them. And, because few studies provide descriptive statistics on the variables measured and included in the regression model, it remains difficult to assess whether the assumption that scores are distributed similarly across studies has been met. Given the current limitations, a common method for summarizing the results of regression analyses has been to use a vote-count strategy (see, e.g., Hanushek, 1989 or Cooper et al., 2006). What remains clear is that techniques for synthesizing results from multiple regression analyses need to be more extensively developed and studied.

Judging the Quality of Research Syntheses and Meta-Analyses

Given the potential value and increased dependence on research syntheses for assisting the development of effective explanations for behavior and behavioral interventions, an important question concerns how to distinguish good from bad syntheses. Throughout this chapter, we have suggested points of contention at which decisions the synthesist makes may affect the validity of conclusions drawn from the synthesis. The model of integrative synthesis as scientific research presented in Table 11.1 provides general guidelines for judging the quality of research syntheses. At each stage, explicit questions about synthesis methods that relate to quality are posed: (a) Do the operations appearing in the literature fit the synthesists' abstract definition? (b) Is enough attention paid to the methodological details of the primary studies? (c) Was the literature search thorough? (d) Were primary studies evaluated using explicit and consistent rules? (e) Were valid procedures used to combine the results of independent studies? Matt and Cook (1994) have expanded on this approach to assessing the validity of research synthesis conclusions. For example, Matt and Cook (1994) also suggest that the possibility that the meta-analyst has used an invalid rule for inferring a characteristic of the target population is another threat to the validity of meta-analytic conclusions. In addition, the validity of results might be threatened because of the probabilistic 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. Shadish, Cook, and Campbell (2002) have expanded Matt and Cook's compendium of threats even further.

In sum, social research methodologists need to continue to identify and systematize criteria for

©2009 SAGE Publications, Ltd.. All Rights Reserved.

the evaluation of meta-analyses. This effort should guide and facilitate the generation of high-quality research syntheses in the future. As the role of syntheses in our acquisition of knowledge expands, the ability to distinguish good from bad syntheses becomes more critical.

Discussion Questions

1. What is the primary impetus for adoption of meta-analysis in the social sciences?
2. Name several channels by which to search for relevant literature. What are the strengths, weaknesses, and cost-effectiveness of each?
3. Briefly review the key components of a meta-analysis. Discuss any potential threats to validity that may occur as a result of decisions the synthesist makes at the data analysis stage.
4. What criteria are most crucial to consider when evaluating the quality of primary research?
5. What criteria are most crucial to consider when evaluating the quality of a research synthesis?

Exercises

1. Identify a conceptual variable and list the operational definitions associated with it that are known to you now.
2. List the keywords that you would use to search for articles relevant to your conceptual variable in electronic reference databases. Use them to identify other related terms in the thesauri of at least two reference databases. What did you learn about your concepts from the new keywords you discovered? Did the keywords differ for the different reference databases and if so, how?
3. Find several reports that describe research relevant to your topic. How many new operational definitions did you find? Evaluate these with regard to their correspondence to the conceptual variable.
4. Read two research syntheses. Outline what the authors report on each of the following: (a) how the literature search was conducted, (b) what rules were used to decide if studies were relevant to the hypothesis, and (c) what rules were used to decide if cumulative relations existed. Was there any information that the synthesists did not report that would be needed to fully evaluate the quality of the research

©2009 SAGE Publications, Ltd.. All Rights Reserved.

syntheses?

Authors' Note: Portions of this chapter appeared originally in H. M. Cooper, “Meta-analysis and the Integrative Research Synthesis,” in C. Hendrick and M. S. Clark (Eds.), *Research Methods in Personality and Social Psychology* (Sage, 1990); H. M. Cooper, J. C. Robinson, and N. Dorr, “Conducting a Meta-analysis,” in F. T. L. Leong and J. T. Austin (Eds.), *The Psychology Research Handbook: A Guide for Graduate Students and Research Assistants* (Sage, 2006); and E. A. Patall and H. Cooper, “Conducting a Meta-Analysis,” in P. Alasuutari, L. Bickman, and J. Brannen (Eds.), *The Handbook of Social Research Methods* (Sage, 2008).

References

http://dx.doi.org/10.4135/9781483348858.n11