11 Contemporary Approaches to Meta-Analysis in Communication Research

Blair T. Johnson

Lori A. J. Scott-Sheldon

Leslie B. Snyder

Seth M. Noar

Tania B. Huedo-Medina

As in any modern science, progress in the field of communication hinges on having trustworthy generalizations from past research on a particular topic. The ever-growing mountain of evidence available about communication research on one hand is an amazing resource but on the other hand represents a considerable challenge to any scholar reviewing

Authors' Note: We express our gratitude for the comments on a previous draft of this chapter provided by Jessa LaCroix, Jennifer Ortiz, Karin Weis, and two anonymous reviewers. The preparation of this chapter was supported by U.S. Public Health Service grants R01-MH58563 to Blair T. Johnson and 1P01CD000237 to Leslie B. Snyder.

this evidence. Consequently, meta-analysis has become a nearly indispensable tool in order to statistically summarize empirical findings from different studies. Meta-analysis is also known as *research synthesis* or *quantitative reviewing,* slightly broader terms that incorporate not only statistical aspects but also the surrounding steps that constitute a review.

The first quantitative reviews of empirical data from independent studies appeared in the early 1800s (Stigler, 1986), but as Olkin (1990) summarized, relatively sophisticated techniques to synthesize study findings began to emerge around 1900, following the development of standardized effect-size indices such as *r*-, *d*-, and *p*-values. Two high-profile reviews on education and psychotherapy (Smith & Glass, 1977; Smith, Glass, & Miller, 1980) helped to popularize the technique and its new name, "meta-analysis," and scholars in communication sciences as well as other disciplines were quick to realize its potential. Simultaneously, increasingly sophisticated statistical techniques emerged to support such efforts (e.g., Hedges & Olkin, 1985; Rosenthal & Rubin, 1978; Schmidt & Hunter, 1977). Standards for meta-analysis have grown increasingly rigorous in the past 20 years, and more "how to" books have appeared (e.g., Lipsey & Wilson, 2001).

Despite early controversy regarding the methods used by meta-analysts (for a review, see Hunt, 1997), meta-analysis has become quite common and well accepted because scholars realize that careful application of these techniques often will yield the clearest conclusions about a research literature (Cooper & Hedges, 1994a; Hunt, 1997). Even the most casual reader of scientific journals can easily witness the widespread acceptance of meta-analysis. For example, a title keyword search for "meta-analysis," ignoring its synonyms, retrieved 5,942 hits in *PsycINFO* and 24,829 in *PubMed;* more broadly, a search for "meta-analysis" in Google retrieved more than 1,600,000 Web hits (January 7, 2007). The story is the same in the field of communication research. Notably, two early proponents of meta-analysis, Alice H. Eagly and John E. Hunter, trained numerous doctoral students who focused on communication research. Several volumes compile meta-analyses on broad areas of communication research, including persuasion (Allen & Preiss, 1998), interpersonal communication (Allen, Preiss, Gayle, & Burrell, 2002), and mass media (Priess, Gayle, Burrell, Allen, & Bryant, 2006). Noar's (2006) recent review documented the growing application of meta-analysis to one of the communication discipline's fast-growing subdisciplines—health communication. Along with numerous other outlets, *Communication Yearbook* specifically welcomes meta-analytic reviews. The International Communication Association annually presents the John E. Hunter Memorial Award for the best meta-analysis in communication. In addition, a recent keyword search of "meta-analysis" within *Communication Abstracts,* which catalogs approximately 50 communication journals, revealed that the number of published meta-analyses among communication journals has increased steadily since 1984 (Noar, 2006).

Those interested in synthesizing communication research have asked and answered many questions through the use of meta-analysis, and in a variety of domains. Noar's (2006) *Communication Abstracts* review noted above found some of the earliest communication meta-analyses to be focused on persuasion and social influence, such as Dillard, Hunter, and Burgoon's (1984) meta-analysis of foot-in-the-door and door-in-the-face techniques and Buller's (1986) meta-analysis of distraction during persuasive communication. In the 1980s and 1990s, communication scholars applied meta-analysis to a variety of communication literatures within mass and interpersonal communication. More recent applications of the technique have included areas as diverse as organizational (Rains, 2005), instructional (Allen et al., 2004), political (Benoit, Hansen, & Verser, 2003), and health communication (Noar, Carlyle, & Cole, 2006; Snyder et al., 2004). Communication scholars have also contributed to discussion of issues surrounding the technique of meta-analysis itself. For instance, a special section of the December 1991 issue of *Communication Monographs* was dedicated to "Issues in Meta-Analysis" (i.e., Hale & Dillard, 1991; Hall & Rosenthal, 1991), and other work on meta-analysis has appeared in the literature both before (Morley, 1988) and after (Hullett & Levine, 2003) this special issue was published.

Historically, scholars used informal methods known as *narrative reviewing*—a summary of the results of individual primary studies sometimes guided by a count of the number of studies that had either produced or failed to produce statistically significant findings in the hypothesized direction. Narrative reviews have appeared in many different contexts and still serve a useful purpose in writing that does not have a comprehensive literature review as its goal (e.g., textbook summaries, introductions to journal articles reporting primary research). Nonetheless, narrative reviews can also prove inadequate for reaching definitive conclusions about the degree of empirical support for a phenomenon or for a theory about the phenomenon.

One indication of this inadequacy is that independent narrative reviews of the same literature often have reached different conclusions. For example, conclusions from the narrative reviews in the Surgeon General's 1972 report on the effects of violent television viewing on aggressive behavior and subsequent major narrative reviews (e.g., Comstock, Chaffee, Katzman, McCombs, & Roberts, 1978; Comstock & Strasburger, 1990; Huston et al., 1992; National Institute of Mental Health, 1982) were contradicted by other reviews (e.g., Friedman, 1988), enabling the controversy over violent television to continue. With the growing popularity of meta-analysis, some of the controversy diminished, at least among scholars; a meta-analysis of 200-plus studies found that after they viewed violent television, children acted more aggressively (Paik & Comstock, 1994). Comparisons between narrative and meta-analytic reviews in other domains (e.g., delinquency prevention and job training) have found similar results with narrative reviews underestimating treatment effects

(Mann, 1994). The reasons for such inaccurate conclusions hinge on at least four problems that have received much past attention (e.g., Glass, McGaw, & Smith, 1981; Rosenthal, 1991; Rosenthal & DiMatteo, 2001):

1. Narrative reviews generally gather only a convenience sample of studies, perhaps consisting only of those studies that the reviewer happens to know. Because the review typically does not state how the studies were gathered or selected for inclusion, it is difficult to evaluate whether the correct literature was gathered or whether the search for studies was thorough. If the sample of studies was biased, the conclusions reached may also be biased.
2. Narrative reviews generally lack statements about which study characteristics were considered or about how the quality of the studies' methods was evaluated, with the result that the accuracy of the reviewers' claims about the characteristics of the studies and the quality of their methods is difficult to judge.
3. When study findings in a literature vary widely, narrative reviews generally have difficulty reaching clear conclusions about what differences in study methods best explain disparate findings. Because narrative reviewers usually do not systematically code studies' methods, these reviewing procedures are not well suited to accounting for inconsistencies in findings.
4. Narrative reviews typically rely much more heavily on statistical significance than on effect-size magnitude to judge study findings. Statistical significance is a poor basis for comparing studies that differ in sample size because effects of identical magnitude can differ widely in statistical significance. As a result, narrative reviewers often reach erroneous conclusions about a pattern in a series of studies, even in literatures as small as 10 studies (Cooper & Rosenthal, 1980).

These problems are compounded by the increasing number of studies available to review—and large literatures are more and more the norm. For example, meta-analyses obtained 138 studies on attitude-behavior relations (Kim & Hunter, 1993), 114 on the persuasive impact of various message sources on attitudes and behaviors (Wilson & Sherrell, 1993), 94 examining disclosure and liking (Dindia, 2002), and 67 on disclosure and reciprocity (Dindia, 2002). Beyond a certain number of studies, note taking quickly becomes an ineffective means of gathering information. In contrast, meta-analytic procedures used to gather, code, and analyze study outcomes provide an improved alternative method for synthesizing information gathered from a large number of studies. Indeed, meta-analysis is the best available tool to conduct these empirical histories of a phenomenon, to show how researchers have addressed the phenomenon, and to

show how results may have changed over time. Meta-analysis has become critical in our understanding and contextualizing of new research findings. Acknowledging scholars' scientific, ethical, and financial responsibility to demonstrate how new research is related to existing knowledge, the British medical journal *The Lancet* now requires authors to reference an existing meta-analysis, conduct their own meta-analysis, or describe the quantitative findings that have appeared since a prior meta-analysis (Young & Horton, 2005).

Because of the importance of comparing study findings accurately, scholars have dedicated considerable effort to making the review process as reliable and valid as possible in an effort to circumvent the criticisms listed above. These efforts highlight the fact that research synthesis is a scientific endeavor with identifiable and replicable methods that are necessary in order to produce reliable and valid reviews (Cooper & Hedges, 1994a).

In spite of the advance it presents, meta-analysis is not without criticism (e.g., Sharpe, 1997). Six common criticisms (see Bangert-Drowns, 1997; Rosenthal & DiMatteo, 2001) are (1) bias in sampling the findings, (2) papers included may vary in quality, (3) nonindependence of effect sizes, (4) overemphasis on differences between individual effects (e.g., differences between means), (5) unpublished studies are underrepresented and published studies are overrepresented, and (6) the "apples and oranges" problem (i.e., summarizing studies with varying methodologies). Although these criticisms bear some resemblance to the criticisms of narrative reviews that we listed above, most of them have arisen out of a misunderstanding of meta-analytic methodology. We will address these criticisms throughout the remainder of this chapter, which provides a general introduction to the methodology of meta-analysis and emphasizes current advances in the technique. We (a) introduce and detail the basic steps involved in conducting a meta-analysis, (b) consider some options that meta-analysts should consider as they conduct such a review, (c) discuss appropriate standards for conducting and evaluating reviews, and (d) conclude with recent developments in meta-analytic methodology.

Meta-Analytic Procedures

Conducting a meta-analysis generally involves seven steps: (1) determining the theoretical domain of the literature under consideration—defining the question, (2) setting boundaries for the sample of studies, (3) locating relevant studies, (4) coding studies for their distinctive characteristics, (5) estimating the size of each study's effect on a standardized metric, (6) analyzing the database, and (7) interpreting and presenting the results. The details and success of each step heavily depend on those

preceding steps. For example, it is easier to set boundaries for studies (Step 2) and to find them (Step 3) if the analyst has first done a good job of defining the meta-analytic question and reviewing relevant theoretical domains (Step 1). In symmetric fashion, even the earlier steps should be accomplished with an eye to the steps that follow. For example, defining a problem too broadly (Step 1) may result in ambiguities in the following methods (Steps 2 through 6) as well as interpretation (Step 7). Some of the steps are similar to conducting a content analysis, a procedure that is familiar to many in communication research (Berelson, 1952; Holsti, 1969; Krippendorf, 1980). In this section, we discuss each step in turn.

DEFINING THE QUESTION

The first conceptual step is to specify with great clarity the phenomenon under review. Ordinarily, a synthesis evaluates evidence relevant to a single hypothesis, defined in terms of the variables that underlie the phenomenon. To select the variables on which to focus, the analyst studies the history of the research problem and of typical studies in the literature. Typically, the research problem will be defined as a relation between two variables, such as the influence of an independent variable on a dependent variable as in Casey et al.'s (2003) investigation of the impact of the public announcement about Earvin "Magic" Johnson's positive HIV status on HIV testing (Casey et al., 2003). Another example is the impact that communication with a sexual partner has on subsequent condom use (Noar et al., 2006).

A synthesis must take study quality into account at an early point to determine the kinds of operations that constitute acceptable operationalizations of the conceptual variables. Because the measures in the studies testing a particular hypothesis often differ, it is no surprise that different operationalizations are often linked with variability in studies' findings. If the differences in studies' measures and other operations can be appropriately judged or categorized, it is likely that an analyst can explain some of this variability in effect-size magnitude.

Essential to this conceptual analysis is a careful examination of the history of the research problem and of typical studies in the literature. Theoretical articles, earlier reviews, and empirical articles should be examined for the interpretations they provide of the phenomenon under investigation. Theories or even scholars' more informal and less-developed insights may suggest moderators of the effect that could potentially be coded in the studies and examined for their explanatory power. When scholars have debated different explanations for the relation, the synthesis should be designed to address these competing explanations.

The most common way to test competing explanations is to examine how the findings pattern across studies. Specifically, a theory might imply that a third variable should influence the relation between the independent

and dependent variables: The relation should be larger or smaller with a higher level of this third variable. Treating this third variable as a potential moderator of the effect, the analyst would code all of the studies for their status on the moderator. This meta-analytic strategy, known as the *moderator variable approach* (or *effect modification approach*), tests whether the moderator affects the examined relation across the studies included in the sample. This approach, advancing beyond the simple question of *whether* the independent variable is related to the dependent variable, addresses the question of *when,* or under what circumstances, the magnitude or sign of the association varies. This strategy aligns well with efforts to build communication theory by focusing on contingent conditions for communication effects (McLeod & Reeves, 1980).

In addition to this moderator variable approach to synthesizing studies' findings, other strategies have proven to be useful. In particular, a theory might suggest that a third variable serves as a mediator of the critical relation because it conveys the causal impact of the independent variable on the dependent variable (Baron & Kenny, 1986; McLeod & Reeves, 1980; also see Chapter 2 in this volume). If at least some of the primary studies within a literature have evaluated this mediating process, mediator relations can be tested within a meta-analytic framework by performing correlational analyses that are an extension of path analysis with primary-level data (Shadish, 1996). We discuss these options further in the sixth step, below; for now, note that there must be sufficient numbers of studies in order for the more sophisticated styles of meta-analysis to proceed.

It is also important to define a priori what constitutes "one study" for inclusion in the meta-analysis. Multiple publications may report on the same study. For example, a meta-analysis of mediated health campaigns chose the campaign as the unit of analysis, often drawing descriptive information about the campaign from one publication and information about campaign effects from another (Snyder et al., 2004). Alternatively, in the experimental literature, one publication often reports on several studies and each may be entered into the meta-analysis. Similarly, each study may be divided into substudies that, for the purpose of the review, are treated as independent studies: As an example, Johnson and Eagly's (2000) meta-analysis examining the role of participant involvement on persuasion treated as separate studies the strong and weak argument conditions of studies that manipulated this variable. In part, their results showed that outcome-relevant involvement increased persuasion for strong arguments and reduced it for weak arguments.

SETTING BOUNDARIES FOR THE SAMPLE OF STUDIES

Clearly, only some studies will be relevant to the conceptual relation that is the focus of the meta-analysis, so analysts must define boundaries

for the sample of studies. This step is similar conceptually to defining the universe of content to be included in a content analysis. Decisions about the inclusion of studies are important because the inferential power of any meta-analysis is limited by the number of studies that are reviewed. Boundary setting is often a time-consuming process that forces reviewers to weigh conceptual and practical issues. The sample of studies is routinely defined by such criteria as the presence of the key variables and acceptable measures, the study quality, and the type of methodology used.

Presence of key variables. The starting point for establishing boundaries is typically conceptualization of the phenomenon that is to be the focus of the synthesis, including identification of key variables. The key variables need to be present and adequately measured for a study to be included in the meta-analysis. The study must report key effects in quantitative terms.

Study quality. As a general rule, research syntheses profit by including those studies that used stronger methods. To the extent that all (or most) of the reviewed studies share a particular methodological limitation, any synthesis of these studies would share the same limitation. It is important to note a key trade-off: Studies that have some strengths (e.g., independent variables with random assignment, laboratory controls) may have other weaknesses (e.g., deficiencies in ecological validity, lack of random sampling).

In deciding whether some studies may lack sufficient rigor to include in the meta-analysis, it is important to adhere to methodological standards within the area reviewed, and these vary widely from discipline to discipline as well as within subdomains. For instance, whereas a meta-analysis conducted within medicine might include only those studies that used a double-blind, random-assignment experimental design, a meta-analytic study in the communication discipline would likely *not* apply such a standard. Rather, a meta-analysis in communication would likely focus more on other methodological aspects such as research design and measurement of variables. For example, Witte and Allen's (2000) meta-analysis of the effects of fear appeals on attitude and behavior change included only those studies that manipulated fear or threat within an experimental or quasi-experimental research design. Studies were excluded if they (a) were cross-sectional, correlating fear or threat with attitude/behavior change but not manipulating them; (b) did not measure the key dependent outcomes under examination; and (c) had a failed fear/threat manipulation check, because the project focused on reactions to various fear-based conditions, such as high and low fear/threat. This meta-analysis provided the most comprehensive synthesis of the fear appeal literature to date, answering a sometimes controversial question regarding whether fear appeals are effective. Witte and Allen (2000) found that fear appeals did, in fact, elicit small but consistent positive effects on attitudes ($r = .14$), behavioral intentions ($r = .11$), and behavior change ($r = .15$). They also found, however, that fear appeals tended to elicit defensive responses such as reactance among

participants (r = .20). Further analysis suggested that those fear appeals that included high-response and self-efficacy messages might have the greatest opportunity of being effective while minimizing the chances of defensive responses.

Although a large number of potential threats to methodological rigor have been identified (Campbell & Stanley, 1963; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002), there are few absolute standards of study quality that can be applied uniformly in every meta-analysis. As a case in point, we have observed that scholars typically think that published studies have higher quality than unpublished studies. Yet many unpublished studies (e.g., dissertations) have high quality and many studies published in reputable sources do not. Obviously, unpublished studies may be unpublished for many reasons, only one of which is low quality. Similarly, many studies may have passed peer review to be published despite the presence of what some may call serious flaws in their methodology. Scholars conducting their first meta-analyses often express amazement that there are so many published studies of low quality. These considerations make it incumbent on the analyst to define the features of a high-quality study and to apply this definition to all obtained studies, regardless of such considerations as the reputation of the journal or whether the study had survived peer review.

Research design. The boundaries of a research literature to be synthesized often include research design specifications. Sometimes analysts set boundaries so that the studies included are relatively homogeneous methodologically. For example, a study of the effects of family planning interventions wanted to control for self-selection into condition as an alternative hypothesis for the effects of the interventions, so the selection criteria included random assignment to conditions (Bauman, 1997).

Sometimes boundaries encompass a variety of methodologies. A meta-analysis of the effect of violent video games selected studies with different methodologies—experimental, correlational, and longitudinal—and then treated methodology as a potential moderator (Anderson, 2004). The results revealed that results in experimental studies paralleled those in correlational studies, providing a better demonstration of causality. In addition, the meta-analysis found larger effect sizes in more methodologically rigorous studies (i.e., those with better sampling), suggesting that earlier pooled estimates of the effects of playing video games on affect, cognition, and behavior were likely underestimates, as they included many methodologically weaker studies. In the past, critics have argued that the synthesizers have combined, in a single analysis, studies that use noncomparable methods, a practice that came to be known as the "apples and oranges" critique (Glass et al., 1981). Nonetheless, methodologists have been generally unsympathetic to this line of argument because they regard it as the task of the meta-analyst to show empirically that differences in methods produce consequential differences in study outcomes (e.g., Hall, Tickle-Degnen, Rosenthal, & Mosteller, 1994; Rosenthal & DiMatteo, 2001). By

treating the methodological differences as moderator variables—as in the video game meta-analysis—the model is fitted for type of fruit, to continue the metaphor. In short, do the results of "apple" studies differ from the results of "orange" studies? Of course, if the effects of methodological differences are known but ignored, analysts may be criticized appropriately as having given insufficient attention to the effects that diverse methods may have had on study outcomes.

Practical considerations sometimes impinge on reviewers' boundary conditions. In many domains, including a wide range of methods would make the project too large and complex to carry out in a reasonable time frame. In such instances, reviewers may divide a literature into two or more research syntheses, each addressing a different aspect of a broad research question. Keeping in mind the phenomena under study, the boundaries should be wide enough that interesting hypotheses about moderator variables can be tested within the synthesis. Yet if very diverse methods are included, the reviewer may need to define some moderator variables that can be implemented only within particular methodologies (e.g., participants' organizational status exists only within studies conducted in organizations).

Critical moderators. Analysts often set the boundaries of the synthesis so that the methods of included studies differ widely only on critical moderator dimensions. The moderators are intended to delineate the literature or expand upon the theory of interest. If other extraneous dimensions are held relatively constant across the reviewed studies by carefully defining the selection criteria, the moderator variable results ought to be more clearly and easily interpreted. An example of a situation suggesting the need for a moderator analysis was in a meta-analysis of studies evaluating HIV prevention interventions for adolescents. These programs varied in the degree to which they increased condom use for adolescents in the intervention compared to the control condition (Johnson et al., 2003). In such circumstances, the odds grow that different mean effects exist within different groups of studies. Indeed, subsequent moderator analyses showed, in part, that interventions were more successful the more condom-skills training was provided.

Cultural factors. For some questions, it may be appropriate to use geographic setting, culture, or study population as a limiting factor, such as when examining the effects of a culturally determined form of nonverbal communication. If the phenomenon under investigation is group specific, then including the studies covering other groups may only obscure the phenomenon. Alternatively, an analyst may choose to treat the setting, culture, or population as a moderating variable and test for differences when the literature includes enough studies for each group. Including reports from diverse settings, cultures, and populations also increases the degree to which the results can be generalized. In addition, to the extent that including such studies increases the ranges that moderator variables take, including studies

from diverse settings, cultures, and populations increases the ability of the meta-analysis to detect moderator variable effects.

Developing selection criteria is often a process that continues as meta-analysts examine more studies and thereby uncover the full range of research designs that have been used to investigate a particular hypothesis. If some studies meeting preliminary criteria established conditions that are judged to be extremely atypical or flawed, the selection criteria may need to be modified to exclude them. The dimensions above highlight the intricate nature of the process. Errors in selection, coding, effect-size calculation, and analyses are more serious than is the case with primary-level research. In primary-level research, such errors typically apply to the unit of analysis, individual observations; in meta-analysis, the errors apply to the entire study. In meta-analysis, errors ought to be envisioned as multiplied by the number of observations in the report for which the error occurred. A mistake in coding for a study of 400 participants is 10 times worse than a study of 40 participants. In the case of communication literatures that bear on public policy issues, one can imagine that meta-analytic errors could alter the conclusions of a review, making the translation of the research results into public policy more prone to error. Even if lives are not at stake, scientific reliability and validity are. For these reasons, we strongly encourage the team concept to meta-analysis, which at least permits ongoing checks and balances against errors. Even the most expert of analysts is subject to human error.

LOCATING THE LITERATURE OF RELEVANT STUDIES

Because including a large number of studies generally increases the value of a quantitative synthesis, it is important to locate as many studies as possible that might be suitable for inclusion, the third step of a meta-analysis. It is conventionally the tacit goal of meta-analyses to obtain *all* of the relevant studies. The very best sample is a complete census of the relevant studies. Indeed, when meta-analyses omit significant numbers of studies, they are often roundly criticized. Because the ideal in meta-analysis is a census of all the relevant studies, meta-analysis is different from the typical content analysis, for which content is systematically *sampled* from the population of relevant content.

To ensure that most if not all studies are located, reviewers are well advised to err in the direction of being overly inclusive in their search procedures. As described elsewhere (e.g., Cooper, 1998; Lipsey & Wilson, 2001; White, 1994), there are many ways to find relevant studies, and analysts are almost always well advised to use them all. Because computer searches of publication databases seldom locate all of the available studies, it is important to supplement them by (a) examining the reference lists of existing reviews (or consulting systematic reviews in specific databases, such as the Cochrane Library Plus and the Campbell Library, which regularly do updates of existing reviews) and of studies in the targeted literature, (b) obtaining published sources that

have cited seminal articles within the literature (using *Social Sciences Citation Index*), (c) contacting the extant network of researchers who work on a given topic to ask for new studies or unpublished studies, and (d) manually searching important journals to find some reports that might have been overlooked by other techniques. The last strategy is especially important for more recent papers that might not yet be included in the electronic databases. Although such a comprehensive search may seem overwhelming, it is imperative if the goal is to retrieve all studies relevant to the topic of interest. Indeed, researchers who have compared searches retrieved from several databases have found that database searching is an insufficient means of literature retrieval and even find differences among electronic reference databases (e.g., Glass et al., 1981; Lemeshow, Blum, Berlin, Stoto, & Colditz, 2005). The review team should carefully record their methods of locating studies, including the names and databases that were searched, and for each database the time period covered and the keywords used. The details of the search procedure should be included in the methods section of the meta-analysis report, to enable readers to make adequate judgments about the adequacy of the procedures used and to permit other analysts to replicate the search.

An important consideration at this stage is whether to include non-English reports, which typically have international samples as well. Decisions about how to deal with the language of the report, on the one hand, and setting, culture, and study populations, on the other hand, should be made separately. Assuming that the decision about whether to include studies from diverse settings, cultures, and populations was made in Step 2, there may be studies reported in foreign languages that otherwise meet the sample selection criteria. To include non-English reports at minimum has the advantage of increasing the sample size in the meta-analysis and thereby systematically increasing the statistical power available in all analyses. If non-English reports in fact comprise the majority of studies, then excluding them would bias the results as well as be an injustice to the excluded reports. Note that a decision to limit the search by setting, culture, or study population may seem to imply the exclusion of non-English-language reports, but it is still possible that studies published in another language sampled the target population. Decisions to exclude on the basis of the language of the publication need to be carefully justified based on the phenomena under study and the nature of the literature in that domain. Note that in meta-analysis, multilanguage ability often is a plus, and even when the analyst team cannot interpret a report on their own, there are software products available to assist in the process, and colleagues with the needed language can perform favors.

CODING STUDIES FOR THEIR DISTINCTIVE CHARACTERISTICS

Once the sample of studies is retrieved, the fourth step in the process is to code them. Coding a meta-analysis is very similar to coding a content

analysis. A coding sheet or an electronic database worksheet needs to be created, pretested, and revised. The variables to be coded and the possible values need to be operationalized precisely. Study characteristics may be either quantitative variables with values existing along ratio, interval, or ordinal scales or categorical variables having discrete numbers of values that reflect qualitative differences between those values. There may be a master codebook that explains the details for each category, or the information can be included in the database worksheet.

To the extent that the analyst team codes many features of the study, they should distinguish between study features that they expect on an a priori basis to account for variation among the studies' effect sizes, on the one hand, and those that provide merely descriptive information about the usual context of studies in the literature, on the other hand. A meta-analysis may be criticized for "fishing" for significant findings if it appears that too many study dimensions were tested as moderators of the magnitude of effects. Separating the study dimensions has the advantage of keeping the review as theory driven as possible (testing a few moderator variables), while at the same time being appropriately descriptive of the literature in question (including many descriptive variables).

To increase the reliability and accuracy of the coding, (a) coding should be carried out by two or more coders, (b) coders should be carefully trained, (c) the coding instructions should contain sufficient detail so that a new coder could apply the scheme and get similar results, and (d) disagreements between coders should be resolved through discussion or with a third coder. Good supervision is critical, including spot checks, trial runs, and easy access by coders for inevitable problems and questions. An appropriate index of intercoder reliability (e.g., Krippendorff's α, Cohen's k, etc.; see Hayes & Krippendorff, in press; Krippendorff, 1980, 2004) should be calculated and reported in the report of the meta-analysis.

Some variables may necessitate additional coders. For example, meta-analysts may consider recruiting outside judges to provide qualitative ratings of methods used in studies. Meta-analyses often use either groups of experts or novices similar to those participating in the studies in order to judge stimuli from study reports. The mean judgments are then put into the database as potential moderator variables.

ESTIMATING THE MAGNITUDE OF EFFECT IN EACH STUDY

The fifth step in a meta-analysis is to estimate the standardized effect size for each study, which quantitatively captures the phenomenon under scrutiny. The problem is that the studies almost always vary widely in terms of choice of statistic as well as sample size, rendering a comparison across the studies complicated. The solution is to impose an effect-size metric on all of the studies. Fortunately, nearly all inferential statistics (e.g., *t*-tests,

F tests) and many descriptive statistics (e.g., means and standard deviations) can be converted into an effect size (for specifics, see Cooper & Hedges, 1994b; Glass et al., 1981; Johnson & Eagly, 2000; Lipsey & Wilson, 2001; Rosenthal, 1991). In consulting and using such guides, it is important to make sure that the best formulas are employed. Failing to do so could result in effect-size estimates that are biased in liberal or conservative directions. As an example, *t*-values and *F* values can derive from both within- and between-subjects designs and formulas exist for both types of designs (see Johnson & Eagly, 2000; Morris & DeShon, 2002). Applying the formulas for the between-subjects cases to the within-subjects cases overestimates their effect size considerably (Dunlap, Cortina, Vaslow, & Burke, 1996; see Morris & DeShon, 2002, for discussion, and Seignourel & Albarracín, 2002, for relevant calculations). Clearly, analysts must carefully consider how the designs of the studies may affect the calculated effect size. If there are enough studies, it may be fruitful to consider conducting parallel, separate meta-analyses for studies with differing designs. Nonetheless, the goal is to convert summary statistics into effect sizes that can be statistically integrated.

*Effect sizes of association (*r, d, *and* OR*).* Effect-size indices usually gauge the association between two variables; an exception to this rule is the arithmetic mean, to which we will turn at the end of this section. Among indices of association, the most commonly used are the standardized mean difference and the correlation coefficient, although odds ratios are popular in some fields, such as medicine and public health (Lipsey & Wilson, 2001). The standardized mean difference, which expresses the difference between two means in standard deviation units, was first proposed by Cohen (1969; Table 11.1, Equation 1). Hedges (1981) showed that Cohen's *d,* which is now often labeled *g,* overestimates population effect sizes to the extent that sample sizes are small and provided a correction for this bias (Equations 2 and 3); with the bias corrected, this effect estimate is conventionally known as *d* (McGrath & Meyer, 2006). Another common effect size is the correlation coefficient, *r,* which gauges the association between two variables (Equation 4). Table 11.2 provides other conventional equations to convert some commonly encountered inferential statistics into *g* (for others, see Lipsey & Wilson, 2001).

Like *d*-values, *r*-values have a bias, in this case, underestimating the population effect sizes, especially for studies with small samples and for *r*-values near .60 (Table 11.1, Equation 5); yet because this bias correction is very small for sample sizes larger than 20, it is often omitted. Because the sampling distribution of a sample correlation coefficient tends to be skewed to the extent that the population correlation is large, many analysts use Fisher's (1921) *r*-to-*Z* logarithmic transform (Equation 6) when conducting analyses (see also Hays, 1988), and then use Fisher's *Z*-to-*r* transform (Equation 7) to return the output to the *r* metric. Although all agree that the distribution of *r* is skewed, Hunter and Schmidt (1990, 2004) have

Table 11.1 Conventional Equations for the Standardized Mean Difference and the Correlation Coefficient, Which Are Effect Sizes of Association Between Two Variables

Equation	*Description*	*Formula*	*Notes and Definitions of Terms*
		The Standardized Mean Difference	
1	Cohen's *d* (now usually labeled *g*)	$g = \frac{M_A - M_B}{SD}$	M_A and M_B are the sample means of two compared groups, and *SD* is the standard deviation, pooled from the two observations. Cohen's *d,* or *g,* is a raw, uncorrected index of association.
2	Hedges's *d*	$d = c(m) \times g$	*d* is the unbiased approximation of the population effect size; $c(m)$ appears as Equation 3.
3	Correction factor	$c(m) \approx 1 - \frac{3}{4m - 1}$	*m* is $n_A + n_B - 2$, the degrees of freedom, where the *n*s are the sample sizes associated with the two compared groups.
		The Correlation Coefficient	
4	Pearson's *r*	$r = \frac{\sum_{i=1}^{N} Z_{Xi} Z_{Yi}}{N}$	z_{Xi} and z_{Yi} are the standardized forms of *X* and *Y* being related for each case *i*, and *N* is the number of observations. Pearson's *r* is a raw, uncorrected index of association.
5	Correction to *r*	$\widetilde{G}_{(r)} \cong r + \frac{r(1 - r^2)}{2(N - 3)}$	$\widetilde{G}_{(r)}$ is the unbiased estimate of the population effect size.
6	Fisher's *r*-to-*Z* transform	$Z_r = \frac{1}{2} \log_e \frac{1 + r}{1 - r}$	$\log_e$ is a natural logarithm operation and *r* is corrected via Equation 5.
7	Fisher's *Z*-to-*r* transform	$r = \frac{e^{(2Z_r)} - 1}{e^{(2Z_r)} + 1}$	*e* is the base of the natural logarithm, approximately 2.718.

argued against the use of the *Z* transformations; Law (1995) provided an excellent review of this issue.

Because *r* can be transformed into *d* (in its *g* form), and vice versa, the choice of an effect-size metric for meta-analysis may seem somewhat arbitrary. Nonetheless, *d* was designed and is quantitatively appropriate for group comparisons of quantitative variables. Other advantages of using the standardized mean effect size are that *d* is well known with formulas for a

Table 11.2 A Selection of Conventional Equations to Translate Inferential Statistics Into the Standardized Mean Difference Effect Size (*g*)

Equation	*Source Statistic*	*Formula*	*Notes and Definitions of Terms*
8	Between-groups, Student's *t*-test	$g = t\sqrt{\frac{n_A + n_B}{n_A n_B}}$	n_A and n_B refer to the sample sizes of the compared groups.
9	Within-participants, Student's *t*-test	$g = \frac{t}{\sqrt{n}}$	This equation gauges change between two observations; *n* is the within-cell *n*, not the total number of observations.
10	Between-groups, *F* test	$g = \sqrt{F\frac{n_A + n_B}{n_A n_B}}$	n_A and n_B refer to the sample sizes of the two compared groups compared by the *F* test. (This equation is not for use with *F* tests comparing more than 2 groups.)
11	Within-participants, *F* test	$g = \sqrt{\frac{F}{n}}$	This equation gauges change between two observations; *F* compares only two groups; *n* is the within-cell *n*, not the total number of observations. (This equation is not for use with *F* tests comparing more than 2 groups.)
12	Correlation coefficient	$g = \frac{2r_{pb}}{\sqrt{1 - r_{pb}^2}}$	r_{pb} is the point-biserial correlation (comparing two groups).

wide array of statistical outcomes available for conversions into *d*, there are forms of *d* that take into account baseline differences (see Becker, 1988), and *d* is easily interpreted (see Van Den Noortgata & Onghena, 2003, for a discussion). Similarly, in its Pearson form, *r* was designed for associations between two quantitative variables. A variant of the family of *r* values, the point-biserial correlation, r_{pb}, is also appropriate for group comparisons on quantitative variables. If two groups are compared on a dichotomous outcome, then the effect size of choice is the odds ratio. Again, a variant of the *r* family can be used, in this case the φ (phi) coefficient. If *r* is used with any categorical variable, then the analyst should use the appropriate version (r_{pb} or r_{φ}) and interpret the results accordingly (McGrath & Meyer, 2006). Finally, just as primary researchers are extolled not to "dumb down" continuous variables into categorical variables, meta-analysts should also avoid this practice (Sánchez-Meca, Marín-Martínez, & Chacón-Moscoso, 2003).

In sum, the convention is to use *r* as the effect size if most of the studies that are integrated report correlations between two quantitative variables. If most of the studies report ANOVAs, *t*-tests, and chi-squares for comparisons between two groups (e.g., experimental vs. control), analysts typically select *d*. If both variables are dichotomous, then they typically select the *OR* (Haddock, Rindskopf, & Shadish, 1998). The positive or negative sign of *r* or *d* is defined so that studies with opposite outcomes have opposing signs; instances with exactly 0 have exactly no association or no difference, respectively. Further, those with values less than 0 have results opposite to those with values more than 0. In the case of the *OR*, instances with exactly 1 show exactly no difference; values less than 1 are opposed from those more than 1. Analyses of the *OR* use the logged form and transform output values for interpretation in the raw *OR* form.

If a report provides only an inexact statistic, or if the report merely states "the difference was nonsignificant," a meta-analytic team might contact the authors of the study for more precise information. If the only information available is imprecise and there is no feasible way to make it more precise, meta-analytic convention is to maintain the imprecise information so that the study is not lost to the review (Rosenthal, 1991). For example, a nonsignificant difference might be represented as $d = 0.00$, $r = .00$, or $OR = 1.00$. An effect described as "$p < .05$" can be converted as an exact *p*-value ($p = .05$) to an effect size. These estimates are conservatively biased (i.e., closer to zero than they are likely to be in reality) but have the advantage of keeping the report in the sample.

When one or both of the variables that are related in the meta-analysis were operationalized in more than one way in a given report or in two or more reports of the same study, the analyst must decide whether to average the effect sizes in order to represent the study with a single effect-size estimate. It is desirable to pool the estimates, rather than treat them as separate studies, in order to ensure that the participants' data contribute to only one effect size and preserve the independence of each effect size in the meta-analysis. Pooling is also a more systematic way to treat the data than arbitrarily choosing to include one effect size from a study rather than another. When pooling data, there are more accurate averaging procedures than using the mean or median of the effect sizes (see Gleser & Olkin, 1994; Rosenthal, 1991). Several scholars have described procedures to combine effect sizes within studies, taking into account the magnitude of their observed associations (Rosenthal & Rubin, 1986) or independent groups and repeated measures (Morris & DeShon, 2002).

Reports may also contain more than one form of statistical information that could be used to calculate a given effect size. For example, a report might contain an *F* test as well as means and standard deviations. The analyst should compute the effect size from both such sources, which, in the end, are all fundamentally interrelated forms of information, and, as long as the effect sizes are similar, take a simple average of them. Yet keep in

mind that more accurate statistics typically have more decimal places and that rounding errors can produce discrepancies in calculated effect sizes. If the effect-size estimates are highly dissimilar, there may be errors in the information reported or the analyst's calculations. In the absence of obvious errors, the analyst must judge which value to enter into the data set, if any. Sometimes an inspection of the report's quantitative information for its internal consistency suggests that one form of the information is more accurate. If the discrepancy is serious and not readily resolved, one possibility is to contact the authors of the report. Only as a final resort should the study be discarded as too ambiguous.

Finally, studies sometimes examine the relation of interest within levels of another independent variable. In such instances, effect sizes may be calculated within the levels of this variable as well as for the study as a whole. This procedure was followed in the example cited earlier for Johnson and Eagly's (1989) meta-analysis of involvement and persuasion. Overall, the effects of involvement on persuasion were uninteresting. By separating the effects of involvement separately for experimentally induced levels of argument strength, the results revealed that different forms of involvement had distinctively different effects on persuasion.

Artifact corrections of indices of association. No matter how reliable or valid, scientific measures are always subject to error. Consequently, any estimate of effect size is just that—an estimate. Corrections for measurement unreliability and other forms of error or bias can be implemented in a meta-analysis in order to estimate what the magnitude of a relation would be in the absence of such artifacts. Hunter and Schmidt (1990, 1994, 2004; Schmidt & Hunter, 1996) explained how to implement corrections in the independent and dependent variables for measurement error, artificial dichotomization of a continuous variable, imperfect construct validity, and range restriction. In theory, correcting for such errors permits a more accurate estimation of the true effect size—that is, what its value would take had studies not been affected by these biases. Even when it is possible to implement fully the corrections within a literature, problems may emerge. Rosenthal (1991) noted that corrected effect sizes can take on irrational values (e.g., correlations larger than 1.00); Schmidt and Hunter (1996) concluded that such observations are due to sampling error and thus more likely to occur with small samples.[1] In considering whether to use such corrections, we recommend that analysts consider their goals. If the goal is to estimate the effect size that would exist if there were no contamination by any artifacts of measurement, then the corrections would be desirable. In contrast, if the goal is to show how large a relation is in practice, then the corrections would be less useful.

Regardless of whether these corrections are implemented, it is wise for analysts to be aware of potential biases that might enter into their studies' effect sizes. In particular, the effect-size indices that we have considered are

ratios of signal to noise, like all inferential statistics. For example, in a between-groups design, the signal is the difference in means, and the noise is the pooled standard deviation (see Tables 11.1 and 11.2). Methodological factors can influence the effect size through their impact on signal, noise, or both factors. If two identical studies are conducted and one controls for noise and the other study does not (e.g., by statistically controlling for an individual difference characteristic), the first study will have a smaller error term than the second and its effect size will be larger. We recommend equating as much as possible how the comparisons are made across studies, so that the effect sizes are not impacted by differing statistical operations. Once again, analysts are wise to keep in mind that their effect-size indices as well as other measured features are estimates.

The arithmetic mean as an effect size. The strategies we have presented above pertain to effect sizes that relate one variable to another, whether in *r, d,* or *OR* forms. In the past decade, reviewers have begun to conceptualize arithmetic means as effect sizes, which gauge the magnitude of a dimension present in a sample rather than how much two variables are associated (Lipsey & Wilson, 2001). For example, Twenge (2000) used meta-analytic techniques to show that levels of anxiety steadily increased from the early 1940s to the 1980s among children and college students in the United States and that the increases were associated with cultural trends. To use such a meta-analytic strategy, the studies must express the phenomenon of interest on the same scale or else the analyst team must convert the scales to a common metric, along with their variability estimates (e.g., standard deviation, variance, or standard error).

Meta-analyses using means are rare in communication research, at least as of this date, but the potential may be enormous. Analyst teams might well examine changes in attitudes, beliefs, knowledge, or behavior defined as change against a baseline, as a mean rather than as a standardized mean effect size. In such a fashion, the team could examine, for example, whether resistance to political persuasion or apathy is becoming more the norm across time. Or research teams might examine change in a key mass communication variable such as average hours of television watched, or a health communication variable such as average amount of time a doctor spends with a patient, or measures of relational or work satisfaction in order to see if these variables are increasing, decreasing, or stable over time. Or the concern may be how the mean changes in response to other factors of interest. Researchers who currently use archival data in time series analyses—an approach commonly used in political communication, for example—may benefit from applying lessons from meta-analysis to better combine studies that have varying sample sizes and operationalizations of key variables (see Chapter 4 in this volume for a discussion of time series analysis in communication).

The disadvantage of invoking means as effect sizes is that their observed levels are likely to be more inconsistent than one typically observes with

indices of association. The increased variability reflects the impact of practically every conceivable factor (e.g., personality and cultural changes, biological factors, and temporal news events). When the effect size is instead, for example, a comparison of two groups in response to the same stimulus, then all these alternative causes are controlled (at least in experimental designs, less so in nonexperimental designs). The remaining difference presumably reflects factors related directly to group membership. Consequently, analysts who conduct reviews using the mean as an effect size should expect to find considerable unexplained variability.

Regression slopes as effect sizes. Similar to means, regression slopes defined as unstandardized regression coefficients also have been used as effect sizes in meta-analysis. The advantage to this strategy is in maintaining the units of the original scales so that inferences can maintain a clear application to some phenomenon. For example, an analyst may wish to see how increases in advertising relate to use of self-help Web sites. Keeping the effect size in real terms would permit a generalization about how much Web site usage increases as advertising increases. Such techniques have been used with different applications and in different contexts including validity generalization (economics, tourism, policy, psychology), dose-response models (epidemiology), and descriptive analysis (education, psychology, economics). Their use has been relatively rare in meta-analysis because their values depend on the scales used to measure the relevant variables (Hunter & Schmidt, 2004). Nonetheless, there are meta-analytic approximations for combining the slopes in meta-analysis (Raudenbush, Becker, & Kalaian 1988; Wu, 2006). When the same scale is used across studies, meta-analysis can be used to synthesize them.

ANALYZING THE META-ANALYTIC DATABASES

Once the effect sizes are calculated, the sixth phase in the process is to analyze the data. In this section, we will assume that the goal is to use quantitative techniques to gauge differences between or across clusters of studies; those who wish to use artifact corrections of effect sizes or to avoid significance testing may be wise to pursue other techniques (see Hall & Brannick, 2002; Schmidt & Hunter, 1996). An exhaustive survey of general analytic approaches to meta-analysis is beyond the scope of the current chapter, but further discussions and comparisons are available elsewhere (e.g., Field, 2001, 2005; Hall & Brannick, 2002; Hunter & Schmidt, 2004; Sánchez-Meca & Marín-Martínez, 1997). The general steps involved in the analysis of effect sizes usually are (a) to aggregate effect sizes across the studies to determine the overall strength of the relation between the examined variables, (b) to analyze the consistency of the effect sizes across the studies, (c) to diagnose outliers among the effect

sizes, and (d) to perform tests of whether study attributes moderate the magnitude of the effect sizes.

Averaging effect sizes. As a first step in a quantitative synthesis, the study outcomes are combined by averaging the effect sizes with the effect for study *i* weighted by the inverse of its variance (v_i), which typically rests heavily on sample size (Hedges & Olkin, 1985); some approaches advocate weighting each effect size by *N* (e.g., Hunter & Schmidt, 2004). Such procedures give greater weight to the more reliably estimated study outcomes, which are in general those with the larger samples (e.g., Hedges, Cooper, & Bushman, 1992). An indirect test for significance of this weighted mean effect size (T_+) is typically conducted using a confidence interval based on its standard deviation in the data, $T_+ \pm 1.96\sqrt{v}$, where 1.96 is the unit-normal value for a 95% *CI* (assuming a nondirectional hypothesis) and *v* is the variance of the estimates across all studies. If the confidence interval (*CI*) includes zero (0.00), the value indicating exactly no difference, it may be concluded that aggregated across all studies there is no association between the independent and dependent variable (*X* and *Y*). For example, Benoit et al. (2003) found that, across 13 studies, debate viewing increased issue knowledge. In a different literature, Sherry (2001) found that, across 25 studies, children's and adolescents' violent–video game playing had a small effect on aggression.

Calculating the heterogeneity of the effect sizes. The next concern is whether the studies can be adequately described by a single effect size, which is assessed by calculating the heterogeneity of the effect sizes across studies, which gauges the amount of variability in the effect sizes around the mean (Cochran, 1954; Hedges, 1981; Hunter & Schmidt, 2004; Rosenthal, 1991). If the effect sizes share a common, underlying population effect size, then they would differ only by unsystematic sampling error. The test statistic *Q* evaluates the hypothesis that the effect sizes are consistent and has an approximate χ^2 distribution with $k - 1$ degrees of freedom, where *k* is the number of studies. If *Q* is significant, the null hypothesis of the homogeneity (or consistency) of the effect sizes is rejected. In this event, the weighted mean effect size may not adequately describe the outcomes of the set of studies because it is likely that quite different mean effects exist in different groups of studies. Further analysis is warranted to test potential moderating variables responsible for different mean effects. *Q* deserves careful interpretation, in conjunction with inspecting the values of the effect sizes. Even if the homogeneity test is nonsignificant, significant moderators could be present, especially when *Q* is relatively large (Johnson & Turco, 1992, and Rosenthal, 1995, provide further discussion). Also, *Q* could be significant even though the effect sizes are very close in value, especially if the sample sizes are very large. Finally, if the number of studies is small, tests of homogeneity are known to have low power to

detect the null hypothesis of homogeneity (Hardy & Thompson, 1998; Harwell, 1997). Higgins and Thompson (2002) introduced a homogeneity index, I^2, whose values range from 0 to 100, where high values indicate more variability among the effect sizes. The I^2 index is subject to the same conditions and qualifications as is *Q* (Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006). The primary benefit of I^2 is that its use would allow for standardized comparisons between meta-analyses while providing the same information as *Q*.

As an example, imagine a meta-analysis that attempts to determine *X*'s impact on *Y.* Deciding not to accept the hypothesis of homogeneity implies that the association between these two variables likely is complicated by the presence of interacting conditions. In some studies, *X* might have had a large positive effect on *Y,* and in other studies, it might have had a smaller positive effect or even a negative effect on *Y.* The next task is to uncover the source of the variation in effect sizes. Because analysts usually anticipate the presence of one or more moderators of effect-size magnitude, establishing that effect sizes are not homogeneous is ordinarily neither surprising nor troublesome.

Finally, analysts often present other measures of central tendency in addition to the weighted mean effect size. For example, the unweighted mean effect size shows the typical effect without weighting studies with larger sample sizes more heavily. A substantial difference in the values of the unweighted and weighted mean effect sizes suggests that one or more studies with large sample sizes may deviate from the rest of the sample. Also, the median effect size describes a typical effect size but would be less affected than a mean effect size by outliers and other anomalies in the distribution of effect sizes.

Analysis of outliers. An analyst can attain homogeneity by identifying outlying values among the effect sizes and sequentially removing those effect sizes that reduce the homogeneity statistic by the largest amount (e.g., Hedges, 1987). Studies yielding effect sizes identified as outliers can then be examined to determine if they appear to differ methodologically from the other studies. Also, inspection of the percentage of effect sizes removed to attain homogeneity allows one to determine whether the effect sizes are homogeneous aside from the presence of relatively few aberrant values. Under such circumstances, the mean attained after removal of such outliers may better represent the distribution of effect sizes than the mean based on all of the effect sizes. In general, the diagnosis of outliers should occur prior to calculating moderator analyses; this diagnosis may locate a value or two that are so discrepant from the other effect sizes that they would dramatically alter any models fitted to effect sizes. Under such circumstances, these outliers should be removed from subsequent phases of the data analysis. More normally, outliers can be examined by analyzing potential moderators of effect sizes, as discussed in the next section. That

is, effect sizes that are apparently outliers may in fact be associated with the coded features of the studies.

Analysis of potential moderators of effect sizes. Ordinarily, analyst teams want to test a priori hypotheses about what explains variations in effect sizes across studies. To determine the relation between study characteristics and the magnitude of the effect sizes, both categorical factors and quantitative factors can be tested. Instead of using such familiar primary-level statistics as *t, F,* or *r* to evaluate whether study dimensions relate to the magnitude of effect sizes, it is best to use statistics that take full advantage of the information in each study's effect size (for discussion, see Hedges & Olkin, 1985; Johnson & Turco, 1992). In *categorical models,* which are analogous to the analysis of variance, analyses may show that weighted mean effect sizes differ in magnitude between the subgroups established by dividing studies into classes based on study characteristics. In such cases, it is as though the meta-analysis is broken into sub-meta-analyses based on their methodological features. For example, Albarracín et al.'s (2003) meta-analysis found that face-to-face or video communications promoted condom use better than those presented in print format (i.e., brochures, posters, or other print). If effect sizes that were found to be heterogeneous become homogeneous within the classes of a categorical model, the relevant study characteristic has accounted for systematic variability between the effect sizes.

Similarly, *continuous models,* which are analogous to regression models, examine whether study characteristics that are assessed on a quantitative scale are related to the effect sizes. As with categorical models, some continuous models may be completely specified in the sense that the systematic variability in the effect sizes is explained by the study characteristic that is used as a predictor. For example, Albarracín et al. (2003) found that exposure to condom-related persuasive communications resulted in greater condom use to the extent that the sample contained more male participants. Goodness-of-fit statistics enable analysts to determine the extent to which categorical, continuous, or mixtures of these models provide correct depictions of study outcomes. Finally, multiple moderators may appear in these models, provided sufficient numbers of studies exist.

Fixed-effects models. The preceding subsection assumed the most basic form of meta-analytic statistic, models based on fixed-effects assumptions, which are the most popular and generally match the assumptions of primary-level research. Fixed-effects models assume that the underlying effect sizes are fixed either as a single group or else along the range of a set of moderator values. In the case of a fixed-effects model specifying a simple weighted mean effect size, the assumption made is that there is one underlying but unknown effect size and that study estimates of this effect size vary only in sampling error. In this case, the test of model specification is the Q or I^2 statistic; a large or significant test implies that the model

is more complex than the model that the analyst assessed and that this simple model is inadequate as a description of the effect sizes.

In the case of a fixed-effects model assessing categorical, quantitative, or multiple predictors, large or significant Q_W or $Q_{Residual}$ values imply that the model is not correctly specified. To say that the effect sizes are fixed is to say that the differences are invariant save for sampling error either as a mean or along a range of moderator dimension(s). In other words, fixing effect sizes to the levels of the moderators has not explained enough of their variation in order for it to be plausible that only variation due to sampling error remains.

To the extent that they have sufficient numbers of studies and available moderators, analysts often add moderators in an effort to achieve a correctly specified model. They may very well do exploratory analyses using the descriptive features of the studies. An alternative is to pursue models with different assumptions, which we address next.

Random-effects models assume that each effect size is unique and that the study is drawn at random from a universe of related but separate effects (for discussions, see Hedges & Vevea, 1998; Hunter & Schmidt, 2000; Lipsey & Wilson, 2001). In addition to sampling error, such models assume that the variation due to the characteristics of studies estimates the between-studies variance present in the universe of effects. In essence, the random-effects model provides an estimate of the population effect size ignoring moderator dimensions, so it should be understood as such. Fitting a random-effects model to extremely heterogeneous sets of effect sizes may erroneously disguise distinct subpopulations of effect sizes. In contrast, when homogeneity tests are nonsignificant and therefore there is no population variance, random-effects models reduce to fixed-effects models: They produce exactly the same mean and confidence interval.

Reviewers of meta-analyses commonly demand random-effects models instead of fixed-effects models when the overall homogeneity statistic is significant. Yet such a criticism is unfounded when the goal of the review is to assess models with moderator dimensions; many reviewers do not realize that random-effects meta-analytic models provide only an estimate of mean effect size without moderators. Random-effects models do not provide estimates of moderators because the presence of moderators implies a dimension along which the effects are fixed.

Mixed-effects models. Models that attempt to maintain the overall random-effects assumption but also fix the effect sizes along certain moderator dimensions are called *mixed-effects models.* Such models assume that the variability in the effect-size distribution is attributed to some systematic between-study differences and an additional unmeasured random component. Strictly speaking, what is fixed is the coefficient of the moderator dimension, or coefficients in the case of multiple-predictor models,

and what is random is the constant of the underlying general linear model. If the constant is of no interest to the analyst team and if the only interest is fixing the effect sizes according to levels of a moderator, then there would seem to be little reason to pursue such models. As in simple regression, the constant in either fixed-effects or mixed-effects models is defined as the point at which the line crosses the *y*-axis. The constant can be of great interest when it reflects meaningful levels at one end of the moderator dimension. Thus, the constant assesses the value of the effect size at level zero of the moderator or moderators. A last consideration is model fit in the mixed-effects case: Because variation in the effect sizes is effectively used to estimate the random constant, there is correspondingly less available to explain when fixing the effect sizes to any moderators. The consequence is that mixed-effects models tend to appear far better fitting than their fixed-effects counterparts, particularly when the distribution of effect sizes is heterogeneous (see Overton, 1998). The risk, as with random-effects models, is that an apparently well-fitting mixed-effects model erroneously disguises subpopulations of effect sizes.

Statistical power. Statistical power assumptions underlie all of the analyses that we have discussed, and power will vary according to the studies' sample sizes, the numbers of studies, and other features. Even tests of model specification are subject to these considerations: If there are few studies, then there is likely to be low power to assess the assumption that the effect sizes are consistent (Hedges & Pigott, 2001). Conducting power analyses is particularly important for interpreting moderator tests, and the failure to do so may result in misleading information (Hedges & Pigott, 2004). If power is found to be low, Hedges and Pigott suggest not conducting moderator analyses or including the power analysis so that readers may be able to correctly interpret the outcomes of the study.

Publication bias. Our discussion of published versus unpublished studies raises the issue of *publication bias,* defined as a bias by authors, reviewers, and editors against null reports or, worse, bias against reports whose data actually oppose a popular hypothesis. Although scholars commonly consider it a bias by the "establishment" against publishing null or reversed effects, in fact, even study authors may exhibit a bias about reporting data that fail to support a pet theory, leaving these findings in the proverbial file drawer, probably not even written up for publication (e.g., Greenwald, 1975). Of course, to the extent that a meta-analysis team has located and retrieved unpublished studies, it is possible to test for publication bias directly by using publication status as a moderator of effect sizes; in such cases, analyst teams should be alert to the possibility that the "unpublished" studies they have obtained are in fact those likely in the passage of time to become published. Yet even when only published studies are included, it is still possible to test for publication bias through the use of a

growing number of techniques (for a review, see Thornton & Lee, 2000). The simplest way is to inspect a *funnel plot* of the distribution of effect sizes; these plots graph effect sizes and their sample sizes (or the inverse of their variance) and ought to reveal a normal distribution if publication bias is not present. Gaps or asymmetries in the graph therefore reveal potential publication bias. More sophisticated techniques attempt to quantify these gaps, as in the trim-and-fill method (Duval & Tweedie, 2000), or to estimate what the mean effect size would be if theoretically missing effect sizes were included (Hedges & Vevea, 1996; Vevea & Hedges, 1995). Another popular technique is to calculate the fail-safe *N,* which is the number of null-result studies necessary to reduce the mean effect size to nonsignificance (Rosenthal, 1991); an implausibly high number would suggest that publication bias is trivial. Despite the popularity of the technique, critics have noted that the index lacks a distribution theory, and therefore it is not known how likely a particular fail-safe *N* value would be to occur based on chance (Begg, 1994).

Even when publication bias seems obvious, analysts are wise to consider alternative reasons why the pattern may have occurred: It may be that the methods of larger studies differed systematically from those of smaller studies. In particular, publication bias is less of an issue when effect sizes lack homogeneity and when moderators can be identified. Publication bias should be considered in light of both the degree of homogeneity and of how effect sizes pattern according to features of the studies. Indeed, under such circumstances, publication bias often becomes a trivial or nonexistent concern.

Vote-counting techniques. In our introduction to this chapter, we mentioned that narrative reviewing has often relied on intuitive counts of the number of studies that had either produced or failed to produce statistically significant findings in the hypothesized direction. Although precision may be enhanced by relying on effect-size indices, statistical models actually exist for doing "vote counting" in a rather sophisticated manner (Darlington & Hayes, 2000). First, note that by sampling error and a conventional alpha level of .05, 1 in 20 studies should produce a significant result. Thus, one method for summarizing a literature would be to note the proportion of studies that obtained the predicted finding and to assess whether this outcome differs from that expected merely on sampling error (Wilkinson, 1951). Darlington and Hayes (2000) showed that such binomial analyses (and several extensions of them) can reduce or eliminate the criticisms that simple vote-counting techniques usually engender. Indeed, these techniques may prove an important adjunct to analyses of effect sizes in that they can provide refined estimates of the likely numbers of omitted reports (see also Bushman & Wang, 1996). Finally, such techniques may prove especially valuable for use in literatures for which many vague statistical reports appear.

INTERPRETING AND PRESENTING THE META-ANALYTIC RESULTS

Science offers no gauges of the truth, only tools with which to divine it. Meta-analysis is thus a tool whose "gauges," or output, must be interpreted in order to present them, which is the seventh step of the process. If the mean effect is nonsignificant and the homogeneity statistic is small and nonsignificant, an analyst might conclude that there is no relation between the variables under consideration. However, in such cases, it is wise to consider the amount of statistical power that was available; if the total number of research participants in the studies integrated was small, it is possible that additional data would support the existence of the effect. Even if the mean effect is significant and the homogeneity statistic is small and nonsignificant, concerns about the mean effect's magnitude arise.

To address this issue, Cohen (1969, 1988) proposed some guidelines for judging effect magnitude, based on his informal analysis of the magnitude of effects commonly yielded by psychological research. In doing so, he intended that a medium effect size would be "of a size likely to be visible to the naked eye of a careful observer" (Cohen, 1992, p. 156), that small effect sizes be "noticeably smaller yet not trivial" (p. 156), and that large effect sizes "be the same distance above medium as small is below it" (p. 156). As Table 11.3 shows, a "medium" effect turned out to be about $d = 0.50$ and $r = .30$, equivalent to the difference in intelligence scores between clerical and semiskilled workers. A "small" effect size was about $d = 0.20$ and $r = .10$, equivalent to the difference in height between 15- and 16-year-old girls. Finally, a large effect was about $d = 0.80$ and $r = .50$, equivalent to the difference in intelligence scores between college professors and college freshmen.[2]

In the field of mass communication, for example, meta-analyses have found small average effect sizes for the effect of health communication

Table 11.3 Cohen's (1969) Guidelines for Magnitude of *d* and *r*

	Effect Size Metric		
Size	*d*	*r*	R^2
Small	0.20	.100	.010
Medium	0.50	.243	.059
Large	0.80	.371	.138

Note: *r* appears in its biserial form.

campaigns on behavior (Snyder & Hamilton, 2002), the cultivation effect of television on beliefs (Shanahan & Morgan, 1999), and the association between television and video game use and body fat (Marshall, Biddle, Gorely, Cameron, & Murdey, 2004). The effect size for the impact of playing violent video games on violent behavior is slightly larger (roughly $r = .20$). It is valuable to be able to compare the magnitude of effects across phenomena, which over time will reveal new patterns across literatures of media effect studies.

Another popular way to interpret mean effect sizes is to derive the equivalent r and square it. This procedure shows how much variability would be explained by an effect of the magnitude of the mean effect size. Thus, under ideal circumstances (McGrath & Meyer, 2006), a mean of $d = 0.50$ and $r = .25$ produces an $R^2 = .09$. However, this value must be interpreted carefully because R^2, or variance explained, is a directionless effect size. If the individual effect sizes that produced the mean effect size varied in their signs (i.e., the effect sizes were not all negative or all positive), the variance in Y explained by the predictor X, calculated for each study and averaged, would be larger than this simple transform of the mean effect size. Thus, another possible procedure consists of computing R^2 for each individual study and averaging these values.

Trends in the Practice of Meta-Analysis

Although the quality of meta-analyses has been quite variable, it is possible to state the features that comprise a high-quality meta-analysis, including success in locating studies, explicitness of criteria for selecting studies, thoroughness and accuracy in coding moderator variables and other study characteristics, accuracy in effect-size computations, and adherence to the assumptions of meta-analytic statistics. When meta-analyses satisfy such standards, it is difficult to disagree with Rosenthal's (1994) conclusion that it is "hardly justified to review a quantitative literature in the pre-meta-analytic, prequantitative manner" (p. 131). Yet merely meeting these high standards does not necessarily make a meta-analysis an important scientific contribution.

One factor affecting scientific contribution is that the conclusions that a research synthesis is able to reach are limited by the quality of the data that are synthesized. Serious methodological faults that are endemic in a research literature may very well handicap a synthesis, unless it is designed to shed light on the influence of these faults. Also, to be regarded as important, the review must address an interesting question.

Moreover, unless the paper reporting a meta-analysis "tells a good story," its full value may go unappreciated by readers. Although there are many paths to a good story, Sternberg's (1991) recommendations to

authors of reviews are instructive: Pick interesting questions, challenge conventional understandings if at all possible, take a unified perspective on the phenomenon, offer a clear take-home message, and write well. Thus, the practice of meta-analysis should not preclude incorporating aspects of narrative reviewing, but instead should strive to incorporate and document the richness of the literature.

One reason that the quality of published syntheses has been quite variable is that it is a relatively new tool among scholars who practice it. Yet as the methods of quantitative synthesis have become more sophisticated and widely disseminated, typical published meta-analyses have improved. At their best, meta-analyses advance knowledge about a phenomenon by explicating its typical patterns and showing when it is larger or smaller, negative or positive, and test theories about the phenomenon (see Miller & Pollock, 1994). Meta-analysis should foster a healthy interaction between primary research and research synthesis, at once summarizing old research and suggesting promising directions for new research. It is valuable if the meta-analytic team includes scholars who are intimately familiar with the literature, to help frame the most interesting research questions and assist in study design.

Another reason that published syntheses have varied widely in quality is the simple reason that meta-analysis can be difficult. The nuances that we have covered in this chapter bear witness to the many nuances and cautions that analyst teams should bear in mind in accomplishing a good research synthesis. Research synthesis is simultaneously a teleological as well as a historical process, qualitative as well as quantitative. Because of the clear advantages of meta-analysis, communication scholars may be more and more expected to conduct meta-analyses rather than narrative reviews. Editors and reviewers are well advised to consider that meta-analysis may usually be preferable to narrative reviewing, but that it is also much more taxing.

One misperception that scholars sometimes express is that a meta-analysis represents a dead end for a literature, a point beyond which nothing more needs to be known. In contrast, carefully conducted meta-analyses can often be the best medicine for a literature, by documenting the robustness with which certain associations are attained, resulting in a sturdier foundation on which future theories may rest. In addition, meta-analyses can show where knowledge is at its thinnest, thus helping plan additional, primary-level research (Eagly & Wood, 1994). For example, the meta-analysis of violent video games by Anderson (2004) found a dearth of studies examining the longitudinal effects of violent video games and called for more primary research to address the gap. As a consequence of a carefully conducted meta-analysis, primary-level studies can be designed with the complete literature in mind and therefore have a better chance of contributing new knowledge. In this fashion, scientific resources can be directed most efficiently toward gains in knowledge.

As time passes and new studies continue to accrue rapidly, it is likely that social scientists will rely more on quantitative syntheses to inform them about the knowledge that has accumulated in their research. Although it is possible that meta-analysis will become the purview of an elite class of researchers who specialize in research integration, as Schmidt (1992) argued, it seems more likely that meta-analysis is becoming a routine part of graduate training in many fields, developing the skills necessary for plying the art and science of meta-analysis to integrate findings across studies as a normal and routine part of their research activities.

Resources

Some general resources on meta-analysis:

1. http://www.psychwiki.com/wiki/Meta-analysis lists many resources on meta-analysis.
2. Dr. William R. Shadish's Web site offers extensive lists related to meta-analysis (see http://faculty.ucmerced.edu/wshadish/Meta-Analysis%20Software.htm).

Some resources for calculating effect sizes:

1. Lipsey and Wilson's (2001) *Practical Meta-Analysis* offers a wide range of equations and an associated Web site with a spreadsheet calculator (at time of publication, posted on Dr. David B. Wilson's Web site http://mason.gmu.edu/~dwilsonb/ma.html).
2. Biostat, Incorporated's, Comprehensive Meta-Analysis software offers many routines to calculate effect sizes (see http://www.meta-analysis.com/).
3. Glass et al. (1981) provided many routines that still do not appear elsewhere.

Some resources for modeling effect sizes:

1. Comprehensive Meta-Analysis will perform nearly all of the analyses that have been described in this chapter.
2. SAS (http://www.sas.com), SPSS (http://www.spss.com), and STATA (http://www.stata.com) will perform the analyses described in this chapter, but users are well advised to invoke the macros provided on Dr. Wilson's Web site, listed above. Wang and Bushman (1999) also provided extensive techniques for meta-analysis using SAS.

3. Analysts who wish to apply the Hunter and Schmidt (2004) artifact corrections can use software available with this book, or another from Dr. Ralf Schwarzer (http://web.fu-berlin.de/gesund/gesu_engl/meta_e.htm).

Notes

1. Charles's (2005) fine review lists historical examples of correlations larger than |1.0| and provides other possible explanations of such instances.

2. McGrath and Meyer (2006) discuss the assumptions involved in these effect-size benchmarks. For example, they point out that Cohen's (1969, 1988) r_{pb} standards assume that the compared groups are equivalent in size. To the extent that the sizes differ, the r_{pb} benchmarks for size will drop. For example, if one group has 99% and the other 1% of the observations, a "large" r_{pb} drops by 78%, from .37 to .08! This change of benchmark does not occur for *d*, which is insensitive to base rates.

References

Albarracín, D., McNatt, P. S., Klein, C. T. F., Ho, R. M., Mitchell, A. L., & Kumkale, G. T. (2003). Persuasive communications to change actions: An analysis of behavioral and cognitive impact in HIV prevention. *Health Psychology, 22,* 166–177.

Allen, M., Mabry, E., Mattrey, M., Bourhis, J., Titsworth, S., & Burrell, N. (2004). Evaluating the effectiveness of distance learning: A comparison using meta-analysis. *Journal of Communication, 54,* 402–420.

Allen, M., & Preiss, R. W. (1998). *Persuasion: Advances through meta-analysis.* Cresskill, NJ: Hampton Press.

Allen, M., Preiss, R. W., Gayle, B. M., & Burrell, N. A. (2002). *Interpersonal communication research: Advances through meta-analysis.* Mahwah, NJ: Lawrence Erlbaum Associates.

Anderson, C. A. (2004). An update on the effects of playing violent video games. *Journal of Adolescence, 27,* 113–122.

Bangert-Drowns, R. L. (1997). Some limiting factors in meta-analysis. In W. J. Bukoski (Ed.), *Meta-analysis of drug abuse prevention programs.* National Institute on Drug Abuse Research Monograph 170, pp. 234–252. Rockville, MD: U.S. Department of Health and Human Services.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173–1182.

Bauman, K. E. (1997). The effectiveness of family planning programs evaluated with true experimental designs. *Public Health Briefs, 87,* 666–669.

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology, 41,* 257–278.

Begg, C. B. 1994. Publication bias. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 400–408). New York: Russell Sage Foundation.

Bem, D. J., & Honorton, C. (1994). Does psi exist? Replicable evidence for an anomalous process of information transfer. *Psychological Bulletin, 115,* 4–18.

Benoit, W., Hansen, G. J., & Verser, R. M. (2003). A meta-analysis of the effects of viewing U.S. presidential debates. *Communication Monographs, 70,* 335–350.

Berelson, B. (1952). *Content analysis in communication research.* New York: Free Press.

Buller, D. B. (1986). Distraction during persuasive communication. A meta-analytic review. *Communication Monographs, 53,* 91–114.

Bushman, B. J., & Wang, M. C. (1996). A procedure for combining sample standardized mean differences and vote counts to obtain an estimate and a confidence interval for the population standardized mean difference. *Psychological Methods, 1,* 66–80.

Campbell, D. T., & Stanley, J. T. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand-McNally.

Casey, M. K., Allen, M., Emmers-Sommer, T., Sahlstein, E., Degooyer, D., Winters, A. M., et al. (2003). When a celebrity contracts a disease: The example of Earvin "Magic" Johnson's announcement that he was HIV positive. *Journal of Health Communication, 8,* 249–265.

Charles, E. P. (2005). The correction for attenuation due to measurement error: Clarifying concepts and creating confidence sets. *Psychological Methods, 10,* 206–226.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10,* 101–129.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

Comstock, G., Chaffee, S., Katzman, N., McCombs, M., & Roberts, D. (1978). *Television and human behavior.* New York: Columbia University Press.

Comstock, G., & Strasburger, V. C. (1990). Deceptive appearances: Television violence and aggressive behavior. *Journal of Adolescent Health Care, 11,* 31–44.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design & analysis issues for field settings. Boston: Houghton Mifflin.

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

Cooper, H., & Hedges, L. V. (1994a). Research synthesis as a scientific enterprise. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 3–14). New York: Russell Sage Foundation.

Cooper, H., & Hedges, L. V. (Eds.). (1994b). *The handbook of research synthesis.* New York: Russell Sage Foundation.

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

Darlington, R. B., & Hayes, A. F. (2000). Combining independent *p* values: Extensions of the Stouffer and binomial methods. *Psychological Methods, 5,* 496–515.

Dillard, J. P., Hunter, J. E., & Burgoon, M. (1984). Sequential-request persuasive strategies: Meta-analysis of foot-in-the-door and door-in-the-face. *Human Communication Research, 10,* 461–488.

Dindia, K. (2002). Self-disclosure research: Knowledge through meta-analysis. In M. Allen, R. W. Preiss, B. M. Gayle, & N. A. Burrell (Eds.), *Interpersonal communication research: Advances through meta-analysis* (pp. 169–185). Mahwah, NJ: Lawrence Erlbaum Associates.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1,* 170–177.

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

Eagly, A. H., & Wood, W. (1994). Using research syntheses to plan future research. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 485–500). New York: Russell Sage Foundation.

Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods, 6,* 161–180.

Field, A. P. (2005). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychological Methods, 10,* 444–467.

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron, 1,* 1–32.

Friedman, J. L. (1988). Television violence and aggression: What the evidence shows. *Applied Social Psychology Annual, 8,* 144–162.

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

Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 339–355). New York: Russell Sage Foundation.

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

Haddock, C. K., Rindskopf, D., & Shadish, W. R. (1998). Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. *Psychological Methods, 3,* 339–353.

Hale, J. L., & Dillard, J. P. (1991). The uses of meta-analysis: Making knowledge claims and setting research agendas. *Communication Monographs, 58,* 464–471.

Hall, J. A., & Rosenthal, R. (1991). Testing for moderator variables in meta-analysis: Issues and methods. *Communication Monographs, 58,* 437–448.

Hall, J. A., Tickle-Degnen, L., Rosenthal, R., & Mosteller, F. (1994). Hypotheses and problems in research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 17–28). New York: Russell Sage Foundation.

Hall, S. M., & Brannick, M. T. (2002). Comparison of two random-effects methods of meta-analysis. *Journal of Applied Psychology, 87,* 377–389.

Hardy, R. J., & Thompson, S. G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine, 17,* 841–856.

Harwell, M. (1997). An empirical study of Hedge's homogeneity test. *Psychological Methods, 2,* 219–231.

Hayes, A. F., & Krippendorff, K. (in press). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures.*

Hays, W. L. (1988). *Statistics* (4th ed.). Fort Worth, TX: Holt, Rinehart & Winston Inc.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6,* 107–128.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? The empirical cumulativeness of research. *American Psychologist, 42,* 443–455.

Hedges, L. V., Cooper, H., & Bushman, B. J. (1992). Testing the null hypothesis in meta-analysis: A comparison of combined probability and confidence interval procedures. *Psychological Bulletin, 111,* 188–194.

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

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods, 6,* 203–217.

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods, 9,* 426–445.

Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics, 21,* 299–333.

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

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21,* 1539–1558.

Holsti, O. (1969). *Content analysis.* Reading, MA: Addison-Wesley.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: *Q* statistic or I^2 index? *Psychological Methods, 11,* 193–206.

Hullett, C. R., & Levine, T. R. (2003). The overestimation of effect sizes from *F* values in meta-analysis: The cause and a solution. *Communication Monographs, 70,* 52–67.

Hunt, M. (1997). *How science takes stock: The story of meta-analysis.* New York: Russell Sage Foundation.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Hunter, J. E., & Schmidt, F. L. (1994). Correcting for sources of artificial variation across studies. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 323–336). New York: Russell Sage Foundation.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment, 8,* 275–292.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Huston, A. C., Donnerstein, E., Fairchild, H., Feshbach, N. D., Katz, P. A., Murray, J. P., et al. (1992). *Big world, small screen: The role of television in American society.* Lincoln: University of Nebraska Press.

Johnson, B. T., Carey, M. P., Marsh, K. L., Levin, K. D., & Scott-Sheldon, L. A. J. (2003). Interventions to reduce sexual risk for the Human Immunodeficiency Virus in adolescents, 1985–2000: A research synthesis. *Archives of Pediatrics & Adolescent Medicine, 157,* 381–388.

Johnson, B. T., & Eagly, A. H. (2000). Quantitative synthesis of social psychological research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 496–528). London: Cambridge University Press.

Johnson, B. T., & Turco, R. (1992). The value of goodness-of-fit indices in meta-analysis: A comment on Hall and Rosenthal. *Communication Monographs, 59,* 388–396.

Kim, M., & Hunter, J. E. (1993). Attitude-behavior relations: A meta-analysis of attitudinal relevance and topic. *Journal of Communication, 43,* 101–142.

Krippendorff, K. (1980). *Content analysis.* Beverly Hills, CA: Sage.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30,* 411–433.

Law, K. S. (1995). The use of Fisher's *Z* in Schmidt-Hunter-type meta-analyses. *Journal of Educational and Behavioral Statistics, 20,* 287–306.

Lemeshow, A. R., Blum, R. E., Berlin, J. A., Stoto, M. A., & Colditz, G. A. (2005). Searching one or two databases was insufficient for meta-analysis of observational studies. *Journal of Clinical Epidemiology, 58,* 867–873.

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

Mann, C. (1994). Can meta-analysis make policy? *Science, 266,* 960–962.

Marshall, S. J., Biddle, S. J. H., Gorely, T., Cameron, N., & Murdey, I. (2004). Relationships between media use, body fatness and physical activity in children and youth: A meta-analysis. *International Journal of Obesity, 28,* 1238–1246.

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of *r* and *d. Psychological Methods, 11,* 386–401.

McLeod, J. M., & Reeves, B. (1980). On the nature of mass media effects. In S. Withey & R. Abeles (Eds.), *Television and social behavior: Beyond violence and children* (pp. 17–54). Hillsdale, NJ: Lawrence Erlbaum Associates.

Miller, N., & Pollock, V. E. (1994). Meta-analysis and some science-compromising problems of social psychology. In W. R. Shadish & S. Fuller (Eds.), *The social psychology of science* (pp. 230–261). New York: Guilford.

Morley, D. D. (1988). Meta-analytic techniques: When generalizing to message populations is not possible. *Human Communication Research, 15,* 112–126.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7,* 105–125.

National Institute of Mental Health. (1982). *Television and behavior: Ten years of scientific progress and implications for the eighties.* Washington, DC: U.S. Government Printing Office.

Noar, S. M. (2006). In pursuit of cumulative knowledge in health communication: The role of meta-analysis. *Health Communication, 20,* 169–175.

Noar, S. M., Carlyle, K., & Cole, C. (2006). Why communication is crucial: Meta-analysis of the relationship between safer sexual communication and condom use. *Journal of Health Communication, 11,* 365–390.

Olkin, I. (1990). History and goals. In K. W. Wachter & M. L. Straf (Eds.), *The future of meta-analysis* (pp. 3–10). New York: Russell Sage Foundation.

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.

Paik, H., & Comstock, G. (1994). The effects of television violence on antisocial behavior: A meta-analysis. *Communication Research, 21,* 516–546.

Preiss, R. W. Gayle, B. M., Burrell, N. A., Allen, M., & Bryant, J. (2006) *Mass media effects research: Advances through meta-analysis.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Rains, S. A. (2005). Leveling the organizational playing field—virtually: A meta-analysis of experimental research assessing the impact of group support system use on member influence behaviors. *Communication Research, 32,* 193–234.

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

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (rev. ed.) Newbury Park, CA: Sage.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

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

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

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

Rosenthal, R., & Rubin, D. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin, 99,* 400–406.

Sánchez-Meca, J., & Marín-Martínez, F. (1997). Homogeneity tests in meta-analysis: A Monte-Carlo comparison of statistical power and Type I error. *Quality & Quantity, 31,* 385–399.

Sánchez-Meca, J., Marín-Martínez, F., & Chacón-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods, 8,* 448–467.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47,* 1173–1181.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529–540.

Seignourel, P., & Albarracín, D. (2002). Calculating effect sizes for designs with between-subjects and within-subjects factors: Methods for partially reported statistics in meta-analysis. *Metodologia de las Ciencias del Comportamiento, 4,* 273–289.

Shadish, W. R. (1996). Meta-analysis and the exploration of causal mediating processes: A primer of examples, methods, and issues. *Psychological Methods, 1,* 47–65.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Shanahan, J., & Morgan, M. (1999). *Television and its viewers.* Cambridge, UK: Cambridge University Press.

Sharpe, D. (1997). Of apples and oranges, file drawers and garbage: Why validity issues in meta-analysis will not go away. *Clinical Psychology Review, 17,* 881–901.

Sherry, J. L. (2001). The effect of violent video games on aggression: A meta-analysis. *Human Communication Research, 27,* 409–431.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752–760.

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy.* Baltimore, MD: Johns Hopkins University Press.

Snyder, L. B., & Hamilton, M. A. (2002). A meta-analysis of U.S. health campaign effects on behavior: Emphasize enforcement, exposure, and new information, and beware the secular trend. In R. C. Hornik (Ed.), *Public health communications: Evidence for behavior change* (pp. 357–384). Mahwah, NJ: Lawrence Erlbaum Associates.

Snyder, L. B., Hamilton, M. A., Mitchell, E. W., Kiwanuka-Tondo, J., Fleming-Milici, F., & Proctor, D. (2004). A meta-analysis of the effect of mediated health communication campaigns on behavior change in the United States. *Journal of Health Communication, 9,* 71–96.

Sternberg, R. J. (1991). Editorial. *Psychological Bulletin, 109,* 3–4.

Stigler, S. M. (1986). *History of statistics: The measurement of uncertainty before 1900.* Cambridge, MA: Harvard University Press.

Surgeon General's Scientific Advisory Committee on Television and Social Behavior. (1972). *Television and growing up: The impact of televised violence.* Report to the Surgeon General, United States Public Health Service. Washington, DC: U.S. Government Printing Office.

Thornton, A., & Lee, P. (2000). Publication bias in meta-analysis: Its causes and consequences. *Journal of Clinical Epidemiology, 53,* 207–216.

Twenge, J. M. (2000). The age of anxiety? The birth cohort change in anxiety and neuroticism, 1952–1993. *Journal of Personality and Social Psychology, 79,* 1007–1021.

Van Den Noortgata, W., & Onghena, P. (2003). Estimating the mean effect size in meta-analysis: Bias, precision, and mean squared error of different weighting methods. *Behavior Research Methods, Instruments, & Computers, 35,* 504–511.

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika, 60,* 419–435.

Wang, M. C., & Bushman, B. J. (1999). *Integrating results through meta-analytic review using SAS® software.* Cary, NC: SAS Institute Inc.

White, H. D. (1994). Scientific communication and literature retrieval. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 41–55). New York: Russell Sage Foundation.

Wilkinson, B. (1951). A statistical consideration in psychological research. *Psychological Bulletin, 48,* 156–158.

Wilson, E. J., & Sherrell, D. L. (1993). Source effects in communication and persuasion research: A meta-analysis of effect size. *Journal of the Academy of Marketing Science, 21*(2), 101–112.

Witte, K., & Allen, M. (2000). A meta-analysis of fear appeals: Implications for effective public health campaigns. *Health Education & Behavior, 27,* 591–615.

Wu, M. J. (2006, April). *Applications of generalized least squares and factored likelihood in synthesizing regression studies.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Young, C., & Horton, R. (2005). Putting clinical trials into context. *The Lancet, 366,* 107–108.