1-1-2000

Quantitative Synthesis of Social Psychological Research

Blair T. Johnson
University of Connecticut, blair.t.johnson@uconn.edu

Alice H. Eagly
Northwestern University

CHAPTER NINETEEN

Quantitative Synthesis of Social Psychological Research

BLAIR T. JOHNSON AND ALICE H. EAGLY

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.

As in other scientific fields, the progress of social psychology has always hinged on investigators' abilities to cumulate empirical evidence about phenomena in an orderly and accurate fashion. This empirical evidence, consisting of multiple studies examining a phenomenon, exists as a literature on the topic. Although new studies rarely replicate earlier studies without adding or removing features, many studies are conceptual replications that use different stimulus materials and dependent measures to test the same hypothesis, and still others contain exact replications embedded within a larger design that adds new experimental conditions. In other instances, repeated tests of a relation accrue in a less systematic manner because researchers sometimes include in their studies tests of particular hypotheses in auxiliary or subsidiary analyses.

In order to reach conclusions about empirical support for a phenomenon, it is necessary to compare and contrast the findings of relevant studies. Therefore, accurate comparisons of study outcomes – reviews of research – are at the very center of the scientific enterprise. Until recently these comparisons were nearly always made using informal methods that are now known as *narrative reviewing*. Scholars drew overall conclusions from their impressions of the overall trend of the studies' findings, 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. For example, textbooks typically contain narrative reviews of many hypotheses, and introductions to journal articles reporting primary research usually include reviews conducted with narrative methods.

Despite the usefulness of narrative reviewing, the method has often proven to be inadequate for reaching definitive conclusions about the degree of empirical support for a phenomenon or a theory of the phenomenon. One symptom of this inadequacy is that independent narrative reviews of the same literature often reached differing conclusions. For example, two separate reviews (Brubaker & Powers, 1976; Green, 1981) concluded that younger adults are better liked than older adults, but another review (Lutsky, 1981) concluded that there is little difference. In such cases, there are no easy rules for deciding which review has reached the most accurate conclusions about the phenomenon in question.

Critics of the narrative reviewing strategy (e.g., Eagly, 1987; Glass, McGaw, & Smith, 1981; Rosenthal, 1991) have pointed to four general faults in narrative reviewing. Although these faults are not necessarily inherent in narrative reviewing, they typify narrative reviewing as it usually has been practiced.

The preparation of this chapter was facilitated by National Institutes of Health Grants K21 MH01377 and R01 MH58563 to Blair T. Johnson and R01 MH48972 to Alice H. Eagly.

We thank Robert F. Bornstein, Steven J. Karau, Mary E. Kite, Charles M. Judd, David A. Kenny, Kenneth D. Levin, Fulgencio Marín-Martínez, Julio Sánchez-Meca, Charles A. Pierce, Harry T. Reis, M. Blanche Şerban, and Lance S. Weinhardt for their helpful comments on previous drafts of this chapter.

Correspondence should be directed to either Blair T. Johnson, Department of Psychology, University of Connecticut, U-0020, 406 Babbidge Road, Storrs, CT 06269-1020 (e-mail: bjohnson@psych.psy.uconn.edu) or to Alice H. Eagly, Department of Psychology, Northwestern University, Swift Hall, 2029 Sheridan Road, Evanston, IL 60208-2710 (e-mail: eagly@nwu.edu).

1. Narrative reviewing generally involves the use of a convenience sample of studies, perhaps consisting of only those studies that happen to be known to the reviewer. Any criteria that a reviewer may have used to select these studies typically go unstated and may never have been formalized by the reviewer. Because the parameters of the reviewed literature are not explicit, it is not possible to evaluate the adequacy of the definition of the literature or the thoroughness of the search for studies. If the sample of studies was biased, the conclusions reached may also be biased.
2. Narrative reviewers generally do not publicly state the procedures they used for either cataloging studies' characteristics or evaluating the quality of the studies' methods. Moreover, the rules and procedures that are applied are often not applied uniformly to all of the studies in the sample, and checks on the reliability of the reviewers' judgments about studies are generally absent. Therefore, the review's claims about the characteristics of the studies and the quality of their methods are difficult to judge for their accuracy.
3. In cases in which study findings differed, narrative reviewing has difficulty reaching clear conclusions about whether differences in study methods explain differences in results. 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 reviewing typically relies much more heavily on statistical significance to judge studies' findings than on the magnitude of the findings. Statistical significance is a poor basis for comparing studies that have different sample sizes, because effects of identical magnitude can differ in statistical significance. Because of this problem, narrative reviewers often reach erroneous conclusions about the confirmation of a hypothesis in a series of studies, even in literatures as small as 10 studies (Cooper & Rosenthal, 1980).

Of course, these potential flaws in the review process are aggravated as the number of available studies cumulates. In contemporary psychology, large research literatures are not uncommon: For example, even as early as 1978, there were at least 345 studies examining interpersonal expectancy effects (Rosenthal & Rubin, 1978). Similarly, by 1983, there were over 1,000 studies evaluating whether birth order is related to personality (Ernst & Angst, 1983). As the number of studies grows, the conclusions reached by narrative reviewers become increasingly unreliable because of the informality of the methods they use to draw these conclusions.

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 and thereby preventing criticisms such as those listed above. The result of these efforts has been the emergence of scientific review techniques known as *quantitative research synthesis*, or *meta-analysis*, which statistically cumulate the results of independent empirical tests of a particular relation between variables. Although scientists have cumulated empirical data from independent studies since the early 1800s (see Stigler, 1986), relatively sophisticated techniques to synthesize study findings emerged only after the advent of such standardized indexes as *r*-, *d*-, and *p*-values (e.g., Birge, 1932; Cochran, 1937; Fisher, 1932; Glass, 1976; Glass et al., 1981; Hunter & Schmidt, 1990; Pearson, 1933; Rosenthal, 1984, 1991; Yates & Cochran, 1938). Reflecting the field's maturation, Hedges and Olkin (1985) presented a sophisticated version of the statistical bases of meta-analysis, and standards for meta-analysis have grown increasingly rigorous.

Social psychologists' first rudimentary applications of quantitative review techniques occurred in the 1960s (e.g., Rosenthal, 1968; Wicker, 1969), but it was not until the late 1970s and early 1980s that these techniques were applied to a wide range of social psychological phenomena (e.g., C. F. Bond & Titus, 1983; Cooper, 1979; Hall, 1978; N. Miller & Cooper, 1991). In many instances these reviews have overturned or enhanced prior narrative reviewers' conclusions. As one example, although Schooler's (1972) and Ernst and Angst's (1983) influential narrative reviews concluded that birth order had little or no relation to personality, Sulloway's (1996, pp. 72–75) subsequent meta-analysis revealed that birth order had significant associations with four of the big five traits (openness to experience, conscientiousness, agreeableness, and emotional instability). Within social psychology, as in many other sciences (Cooper & Hedges, 1994a; Hunt, 1997; Mann, 1994; Thacker, 1988; Wachter & Straf, 1990), quantitative research synthesis is now common and well-accepted because scholars realize that careful application of these techniques will yield the clearest conclusions about a research literature.

In order to provide a general introduction to the techniques of research synthesis, we (a) introduce and detail the steps involved in research synthesis, (b) consider some options that reviewers should consider as

they proceed through these steps, (c) discuss appropriate standards for conducting and evaluating quantitative reviews, and (d) evaluate the role of quantitative synthesis relative to primary research and other methods of testing hypotheses.

PROCEDURES FOR QUANTITATIVE SYNTHESES

An Overview of the Process of Quantitative Synthesis

The research process underlying quantitative synthesis can be broken into a number of discrete yet interrelated steps (see Cooper, 1982). Each stage contributes to the attainment of the next stage: Careful work in the early stages of the synthesis makes the later stages easier to accomplish. As a preview to a more detailed exposition, we list the stages and some of the questions that often accompany them:

1. *Conceptual analysis of the literature.* What independent and dependent variables define the phenomenon? How have these variables been operationalized in research? Have scholars debated different explanations for the relation demonstrated between these variables? Can the meta-analysis address these competing explanations? When, how much, and in what pattern should the variables relate? Should the size of the relation be relatively consistent or inconsistent across studies?
2. *Setting boundaries for the sample of studies.* What criteria should be used to select studies into the sample? Should considerations of study quality play a major role? What criteria should be used to *exclude* studies from the sample?
3. *Locating relevant studies.* What strategies will best locate the universe of studies? How can unpublished studies be obtained?
4. *Creating the meta-analytic database.* Which effect size metric should be used? Which study characteristics should be represented, and how can these characteristics be coded or otherwise assessed? How can the quality of studies' methods be assessed?
5. *Estimating effect sizes.* What are the best ways to convert study statistics into effect sizes? How can extraneous influences on effect size magnitude best be controlled?
6. *Analyzing the database.* How should the effect size data be analyzed statistically? Which of the available meta-analytic frameworks for statistical analysis is most appropriate? What sorts of statistical models can be used? How can the tests associated with these models be interpreted? How can statistical outliers among the effect sizes be located and treated?
7. *Presenting and interpreting the results.* What information about the studies should be presented? Which meta-analytic models should appear? What are the best techniques for displaying the meta-analytic results? What knowledge accrues from the synthesis? How do the meta-analytic results reflect on the theoretical analysis? Has the synthesis uncovered important areas of neglect in the literature that warrant future research?

Conceptual Analysis of the Literature

The initial conceptual exploration of a research literature is critical because ideas formulated at this early point can dramatically affect the methods that follow, such as the criteria the reviewer formulates for including and excluding studies. The first conceptual step leading to a successful meta-analysis is to specify with the greatest possible clarity the phenomenon under review by defining the variables whose relation is the focus of the review. Ordinarily a synthesis evaluates evidence relevant to a single hypothesis that is defined as a relation between two variables, often stated as the influence of an independent variable on a dependent variable (e.g., the influence of group cohesiveness on group performance, synthesized by Mullen & Copper, 1994). Moreover, a synthesis must take study quality into account at an early point to determine the kinds of operations that constitute acceptable operationalizations of these conceptual variables. Typically, studies testing a particular hypothesis differed in the operations used to establish the variables (e.g., different manipulations of the independent variable, different measures of the dependent variable), and it is therefore not surprising that these different operations were often associated with variability in studies' findings. If the differences in studies' operations can be appropriately judged or categorized, it is likely that an analyst can explain some of this variability in effect size magnitude (see sections on *Study characteristics,* and *Testing models of meta-analytic moderators,* below).

Essential to this conceptual analysis is careful study 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. Authors' theories or even their 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 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*, is analogous to the examination of interactions with primary-level data (see section *Estimating Effect Sizes*). However, instead of testing the interaction within one study's data, the meta-analysis tests whether the moderator affects the examined relation across the studies included in the sample. This moderator variable approach, advancing beyond the simple question of *whether* the independent variable is related to the dependent variable, addresses the question of *when* the magnitude or sign of the relationship varies. Using this strategy, Karau and Williams (1993) found that the tendency for people to expend less effort when working collectively than when working individually (known as *social loafing*) was larger for male participants than for female participants. Similarly, C. F. Bond and Titus's (1983) synthesis showed that the presence of others improves the performance of simple tasks but impairs the performance of complex tasks.

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. 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. Using such techniques, Driskell and Mullen (1990) reviewed seven studies that correlated status cues, expectations for performance, and performance; their results showed that the influence of status cues on performance was largely mediated by expectations for performance. Shadish (1996) provided a very helpful exposition of this mediator variable approach to analyzing meta-analytic data as well as other approaches.

Setting Boundaries for the Sample of Studies

In beginning a research synthesis, the reviewer should think about how a relation has been tested and consider whether all tests should be included in the synthesis. Decisions about the inclusion of studies are important because the inferential power of any meta-analysis is limited by the methods of the studies that are integrated. To the extent that all (or most) of the reviewed studies share a particular methodological limitation, any synthesis of these studies would be limited in this respect. For example, a synthesis of correlational studies will produce only correlational evidence about the association in question. However, if the critical hypothesis was tested with true experiments, defined by one or more manipulated independent variables and the random assignment of participants to conditions, the synthesis gauges the causal effect of the independent variables on the dependent variable across the studies reviewed. As a general rule, research syntheses profit by focusing on the studies that used stronger methods to test the meta-analytic hypotheses. Nonetheless, it is important to note that studies that have some strengths (e.g., manipulated independent variables) may have other weaknesses (e.g., deficiencies in ecological validity) (see Brewer, this volume, Ch. 1).

Even though research syntheses of experimental studies are better able to draw conclusions about cause-and-effect relations than are syntheses of correlational studies, meta-analyses of experimental studies generally have other limitations. Of particular note, such syntheses produce only correlational evidence concerning the relations of study characteristics to studies' findings when the effect sizes are compared across studies in so-called "between-studies moderator analyses." For example, Wood, Rhodes, and Whelan (1989) found that in studies published after 1978, men reported more positive well-being than women, whereas this pattern reversed in studies published before 1978. As Wood et al. indicated, effects of year of publication are difficult to interpret because they could be confounded with any number of other variables (see discussion by Knight, Fabes, & Higgins, 1996). In contrast, moderator tests can yield stronger causal claims if the moderator dimension reflects within-studies manipulations. In such cases, random assignment of participants in the primary research to levels of the moderator makes it less likely that confounds were associated with the moderator. In this strategy, the results of each study are divided to produce separate effect sizes within levels of the moderator. For example, Karau

and Williams (1993) showed that social loafing was more pronounced as groups increased in size, a dimension that was experimentally manipulated in many of the studies. If an analysis were limited to the studies that contained this manipulation, any moderation that would be demonstrated could be attributed to the manipulated variable, and interpretation could proceed with greater certainty. Thus, it may be advantageous for reviewers to pay special attention to those within-study comparisons that are available in the reviewed studies.

In deciding whether some studies may be insufficiently rigorous to include in the meta-analysis, a reviewer should be alert to methodological standards within the area reviewed. Although a large number of potential threats to methodological rigor have been identified (see Brewer, this volume; Campbell & Stanley, 1963; Cook & Campbell, 1979; S. Greenwald & Russell, 1991; Wortman, 1994), there are few absolute standards of study quality, and, in practice, the characteristics considered essential to ensure high study quality vary widely from literature to literature. In some research literatures, it might already be known that a certain method (e.g., a measure or a manipulation) yields seriously flawed results; if so, an analyst might decide to eliminate studies that used this method. Indeed, one possible strategy is to omit obviously flawed studies in order to restrict the synthesis to studies of high quality (S. Greenwald & Russell, 1991). As an alternative, an analyst might attempt to correct the effect sizes for certain methodological biases (e.g., unreliability, restriction of range), a subject that we consider further below (see *Correcting effect sizes for bias*). Retaining potentially flawed studies and representing their quality-relevant features in the coding scheme is another defensible strategy. For example, if an analyst suspects that a given independent variable was not established uniformly across the literature, he or she might be able to code the variable's strength and determine whether it is correlated with effect size magnitude. More generally, when there are questions concerning whether variant methods yield the same results, a meta-analysis can be designed to address this issue. Exemplifying this strategy, Heinsman and Shadish (1996) examined whether randomized and nonrandomized experiments reached the same conclusions within four different research literatures; their results suggested that this dimension of study quality did not dramatically impact results.

In addition to study quality, many other considerations enter into setting the boundaries of the research literature that will be synthesized. Although adequate conceptualization of the phenomenon that is the focus of the synthesis should help to define these boundaries, sometimes boundary-setting is in itself a time-consuming process that forces reviewers to weigh conceptual and practical issues. These issues are particularly acute when a phenomenon has been studied using a variety of methods. Sometimes analysts set boundaries so that the studies included are relatively homogeneous methodologically (e.g., only laboratory studies), and sometimes boundaries encompass different methodologies (e.g., field studies are also included). Surely 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). Practical considerations sometimes impinge on reviewers' boundary conditions because 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.

If the boundaries of a meta-analysis are too wide, researchers may be the targets of what has become known as the "apples and oranges" critique (Glass et al., 1981). Critics might thus argue that the reviewer has combined in a single analysis studies that use noncomparable methods and thus has mixed apples and oranges. 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. This demonstration is achieved by disaggregating studies into various categories, as we discuss in the section on *Analyzing the Meta-Analytic Database*. Of course, analysts who do not perform these analyses that separate studies into various types may be appropriately criticized as having given insufficient attention to the effects that diverse methods may have had on study outcomes.

Analysts often set the boundaries of the synthesis so that the methods of included studies differ dramatically only on critical moderator dimensions. If other, extraneous dimensions are thereby held relatively constant across the reviewed studies, moderator variable analyses can be more clearly interpreted. Nonetheless, an analyst should include in the sample all studies or portions

of studies that satisfy the selection criteria. 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. Developing selection criteria is to some extent a process that continues as meta-analysts examine more studies and thereby discover the full range of research designs that have been used to investigate a particular hypothesis.

One issue that generally arises when setting the boundaries for the research literature is whether to include unpublished studies. Although these studies are certainly more difficult to access than published studies, the omission of unpublished studies from a review can bias the review's findings, generally in favor of larger effects, a pattern that has been demonstrated in several studies (e.g., Cooper, DeNeve, & Charlton, 1997; Dickersin, 1997; Glass et al., 1981). In a discussion of unpublished studies, Rosenthal (1979) referred to them as producing a file-drawer problem because they may be buried in researchers' file drawers, especially if their results were nonsignificant. In the extreme case of so-called "prejudice against the null hypothesis" (A. G. Greenwald, 1975, p. 1), it might be that the "journals are filled with the 5% of the studies that show Type I errors, while the file drawers back at the lab are filled with the 95% of the studies that show nonsignificant (e.g., $p > .05$) results" (Rosenthal, 1979, p. 638). In partial support of this claim, surveys of researchers suggest that from 15% to 40% of the studies that are conducted are never published (Cooper et al., 1997; Rotton, Foos, Van Meek, & Levitt, 1995; Shadish, Doherty, & Montgomery, 1989; Sommer, 1987). Although it appears common for authors to decide not to pursue the publication of their studies if the results are nonsignificant (A. G. Greenwald, 1975; Rotton et al., 1995), a variety of other factors also affect the publication status of a study (e.g., author productivity; see Sommer, 1987). Moreover, the studies reported in dissertations and master's theses are less likely to be screened for statistical significance than studies reported in journals. Although dissertation and thesis studies are not considered to be published unless they appear in a journal or some other source, they do not languish in file drawers.

These considerations strongly suggest that if it is plausible that studies with stronger findings were more likely to be published (see Begg, 1994), every effort should be made to obtain unpublished studies. Moreover, there are several other reasons that the inclusion of unpublished studies should almost always facilitate the goals of meta-analysis:

1. Because the potential for conducting an informative synthesis increases with a larger number of studies, no source of obtaining studies should be ruled out. As more studies are included, mean effect size estimates stabilize, and the power to detect moderators of effect sizes increases (Johnson, Mullen, & Salas, 1995).
2. It is generally an explicit goal of meta-analysis to describe the *universe* of studies on a topic or at least an unbiased sample of that universe (White, 1994). If a meta-analysis includes all studies in the literature or an appropriate sample of them, the validity of its representation of the research literature is enhanced. The inadequacy of sampling that excludes unpublished studies can be especially acute when research literatures contain large numbers of unpublished studies. In extreme instances, omitting the unpublished studies would eliminate the majority of evidence on the hypothesis (e.g., comparisons of women's and men's leadership styles; Eagly & Johnson, 1990). Ironically, a synthesist would not even learn that this evidence exists unless the unpublished literature were searched. In the section *Locating Relevant Studies*, below, we detail techniques to find such studies.
3. Regardless of whether a study is published, the analyst should judge it against a set of inclusion and exclusion criteria. Uniform implementation of these criteria circumvents the criticism that unpublished studies generally have unacceptable quality due to the absence of peer review. Rather than merely assume (perhaps incorrectly) that unpublished studies have inadequate quality, a meta-analyst should remove all studies, published or unpublished, that do not meet the review's quality criteria and code the remaining studies on quality-relevant study characteristics (e.g., reliability of measures).

Thus, as a general rule, it is advisable to include unpublished studies. To test whether unpublished and published reports differ in their findings, analysts should examine whether this variable moderates the magnitude of the effect sizes.

A further decision that often arises is whether the sample of studies should be restricted to one country or culture. Of course, the reasons that encourage sampling unpublished studies also encourage sampling studies from all countries and cultures. Moreover, including such studies would increase the inclusiveness of the meta-analysis by representing a broader sample of populations of research participants and permitting an analyst to answer questions about the generality of

the studied effect across diverse cultures. For example, R. Bond and Smith (1996) found that research participants in collectivistic cultures were more likely to conform in the Asch-style line-judgment experiment than were participants in individualistic cultures, although the conformity effect was significant within both types of cultures. Yet, in many research literatures in which it is reasonable to suspect cross-national variability, it may not be possible to address this issue meta-analytically because only a very small number of studies are available from countries other than the one in which the research paradigm first appeared (e.g., Eagly, Makhijani, & Klonsky, 1992; Kolodziej & Johnson, 1996). Therefore, as a general rule, studies from multiple cultures should be included in the sample if they are available in at least modest numbers. Commercially available computer software often can help to overcome foreign language barriers (e.g., Wood, Lundgren, Ouellette, Busceme, & Blackstone, 1994, used software to translate French documents into English).

A final issue is the completeness with which a research literature is reviewed. Some literatures are so enormous that including all studies would be impractical. In these instances, a practical solution might be to take from the entire literature a reasonably-sized random sample. The most defensible way to accomplish this solution is to list all studies in the pertinent literature, decide how many would make a sufficient sample, and randomly select this number of studies, a procedure that Rosenthal and Rubin (1978) followed in their meta-analysis of the interpersonal expectancy effect literature.

Locating 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. When a literature consists at least in part of findings whose presence cannot be discerned from reading studies' titles and abstracts, a reviewer may have to retrieve all studies in the general research area, in order to identify the reports that contain the finding of interest. For example, Eagly and Johnson (1990) screened over 6,000 abstracts of studies on leadership and related topics. They obtained for closer scrutiny those studies that offered the potential, at least, for comparisons of women's and men's leadership styles, and, in the end, 162 studies fit their inclusion criteria. To insure that a sufficient sample of studies is located, reviewers are well advised to err in the direction of being extremely inclusive in their searching procedures. As described elsewhere (e.g., Cooper, 1998; Glass et al., 1981; Mullen, 1989; White, 1994), there are many ways to find relevant studies; ordinarily, analysis should use all of these techniques. Unfortunately, computer searches of databases such as PsycLIT seldom locate all of the available studies, although such searches are extremely useful. The most important retrieval techniques follow:

1. *Computer database searches* are used to find citations of (a) articles whose titles or abstracts contain terms a reviewer has defined or (b) articles that were cataloged in a database under keywords selected by the reviewer. There are many different databases useful for social psychologists, including PsycLIT and PsycINFO (which are made available by the American Psychological Association). *Dissertation Abstracts* (electronic reference products, made available by UMI, contains abstracts for dissertations and some Master's theses), *Social Sciences Citation Index* (compact disc edition, made available by the Institute for Scientific Information), SocioFile, MedLine, and ABI/Inform. ERIC (Educational Resources Information Center) is especially important to psychologists because it contains some entries for papers delivered at meetings of psychological associations as well as unpublished documents such as government reports. ERIC thus provides partial access to the fugitive literature of unpublished studies.

   One way for reviewers to generate search terms is to determine whether the studies that they initially have in hand (perhaps identified from prior reviews) are located by preliminary search terms and thus appear on the output from a trial search. This examination as well as an informal process of trial-and-error using various search terms are generally required to shape an appropriate search strategy, and different sets of search terms can be used in multiple passes in databases. It is important to realize that the yield of such searches can vary widely, depending on the terms used. For example, in separate reviews of the literature on attitudes toward homosexuality, Whitley and Kite (1995) obtained 50 more studies than Oliver and Hyde (1993). Such searches can generally obtain more studies if wildcards are used (e.g., *attitud** will match *attitude, attitudes,* and *attitudinal*), because it is often the case that researchers used a variety of related terms to describe the variables they studied. Because the reviewer can interact directly with these databases, preliminary search terms can be quickly tried and adjusted. Reviewers can generate search output easily through the resources

available in most university libraries. Therefore, there is little reason not to use inclusive search strategies, even though the yield of suitable studies may be small. Librarians expert on computer-based information searching often can provide helpful advice to novice searchers.

2. The *ancestry approach* involves examining the reference lists of existing reviews and of studies in the targeted literature to find likely candidates for the review. In fact, the reference list of every study that is located should be scrutinized. This method is important for locating studies whose publication dates precede the establishment of databases such as PsycINFO, although the American Psychological Association now provides a database of psychological literature extending to 1887 (PsycINFO Complete) and the dissertation abstracts database begins with 1861.
3. In the *descendancy approach*, a reference to a seminal article is specified in a database, and the articles that have cited it are listed. This type of search can be accomplished in Social SciSearch or (more arduously) in the print copies of the *Social Sciences Citation Index*. These searches are particularly useful in research literatures that were initiated by a particular study that was then typically cited by subsequent investigators.
4. The *invisible college technique* makes use of the network of researchers who work on a given topic. Because these researchers may be in frequent contact with one another, they may know who has new studies on which topics, or in some instances, they may have unpublished data sets that they are willing to make available. The reviewer can contact these individuals and may obtain some of the fugitive literature, even though he or she may not be regarded as a member of the invisible college. Helpful in identifying the invisible college are the programs of recent and upcoming meetings of psychological associations; these programs list titles (and often abstracts) of papers delivered at these events. With the advent of widespread access to the Internet and its discussion groups on various issues, yet another technique is to send electronic mail describing the type of study desired to the various "listservs" to which researchers subscribe. In this way, the need for studies is broadcast to hundreds or even thousands of researchers. Such technology has made the "invisible" college increasingly visible.
5. Hand searches of important journals consist of scrutinizing all articles published in key journals. The abstracts of articles with promising titles are read to determine if the text of the article should be examined. This method may turn up some reports that were overlooked by other techniques, and it certainly provides a good cross-check on the adequacy of searching conducted by the other methods.

Finally, to enable readers to judge the adequacy of search procedures and to enable other reviewers to evaluate and replicate these procedures, the review should describe in detail their methods of locating studies, including the names of databases that were searched, and for each database the time period covered and the keywords used. Reviewers should also describe their inclusion and exclusion criteria and provide a rationale for using these criteria.

Creating the Meta-Analytic Database

Once the sample of studies is in hand, the next step is to code them for their characteristics and to calculate effect sizes that estimate the relation being examined. In other words, the studies must be entered into a data set that includes each study's important characteristics and its effect size information.

STUDY CHARACTERISTICS. In conceptualizing the hypothesis that is under scrutiny and the conditions under which studies' results should vary in magnitude (or direction), reviewers develop ideas about the study characteristics that should be coded. The most important of these characteristics are potential moderator variables, which the analyst expects on an a priori basis to account for variation among the studies' effect sizes. It is also important to consider whether studies that differ along a critical moderator dimension also differ on other dimensions. As Lipsey (1994) suggested, it "is quite unlikely that study characteristics will be randomly and independently distributed over the studies in a given literature" (p. 117). Because such confounds could produce interpretational difficulties, it is important to code these additional characteristics so that their moderating influences can be examined, if only on an exploratory or post hoc basis. Finally, it is also important to code the studies for numerous other characteristics such as their date of publication and participant population, even if these characteristics are not expected to account for variation in studies' outcomes (see Lipsey, 1994). The central tendencies of these characteristics are ordinarily displayed, often in a table, to provide readers of the review with a description of the usual context of studies in the literature.

Study characteristics may be either continuous or *categorical*. Variables on a *categorical* metric consist of a

discrete number of values that reflect qualitative differences between those values. For example, among the categorical study characteristics that Bornstein (1989) coded in his synthesis of studies on the mere exposure effect was the type of measure used to assess affect, which had categories such as liking, goodness-of-meaning, and pleasantness. Variables on a *continuous* metric consist of values that exist along ratio, interval, or ordinal scales. For example, among the continuous study characteristics that Bornstein coded was stimulus exposure time. Typically, synthesists code variables by using a coding form that displays the classes comprising each categorical variable and provides blank spaces for entry of the values of continuous variable (see Stock, 1994, for examples of such forms).

Some important features of studies cannot be coded from their method sections but must be assessed from other sources. For example, Eagly and Johnson's (1990) synthesis of differences between men's and women's leadership style retrieved archival data from the U.S. Census in order to estimate the distribution of the sexes in various leadership roles. Similarly, in a review of the effects of tripling in college dormitory rooms, Mullen and Felleman (1989) learned what specific dormitories had been studied and then obtained blueprints from college administrators in order to gauge physical features that were relevant to crowding.

Other techniques may prove useful to gauge study features that are difficult to code accurately. For example, in a meta-analysis on sex-related differences in aggression, Eagly and Steffen (1986) wished to determine whether women and men differed in how unfavorably they perceived the aggressive acts that had been examined in the social psychological literature on aggression. Therefore, they had female and male students rate the extent to which each such act would produce harm to the target of aggression, guilt and anxiety in oneself as the aggressor, and danger to oneself. From these ratings, Eagly and Steffen estimated sex differences in these students' perceptions of the aggressive acts and related these scores to the effect sizes that represented sex differences in aggressive behavior. In other instances, experts' ratings could be obtained based on their reading of the method sections of the reports or of the actual stimulus materials used in the studies (e.g., Johnson & Eagly, 1989; N. Miller & Carlson, 1990). Yet, it is desirable to provide convergent evidence of the validity of the judges' ratings used by these methods, because these judges function only as observers of studies' methods. Observers may be biased, as suggested by the results of studies using roleplaying participants, which have sometimes deviated from the results of studies using actual participants (e.g., A. G. Miller, 1972). One way to determine the validity of judges' ratings of manipulation effectiveness is to compare them with effect sizes representing the manipulation checks present in the studies. To the extent that these values are highly correlated, the judges' ratings can be trusted (e.g., Bettencourt & Miller, 1996; N. Miller, Lee, & Carlson, 1991).

RELIABILITY OF CODING. Because each study outcome in a meta-analysis represents the data provided by many research participants, coding errors can be very consequential, perhaps even more critical than coding errors made in primary research. These errors are particularly consequential when they pertain to moderator variables. Illustrating this point, Wanous, Sullivan, and Malinak (1989) reviewed four literatures for which two independent meta-analyses had been performed. Their analysis of these syntheses suggested that some of the results differed because moderator variables had been coded differently and perhaps erroneously in some instances. Because accurate coding is crucial to the results of a meta-analysis, the coding of study characteristics should be carried out by two or more coders, and an appropriate index of interrater reliability should be calculated (such as the intraclass correlation and Cohen's, 1960, *kappa*: see Langenbucher, Labouview, & Morgenstern, 1996; Orwin, 1994). In most cases, disagreements between coders can be resolved by discussion, or perhaps by averaging. In order to improve the reliability of the coding of study characteristics, a small subset of the studies can be used as a trial run. If agreement is low on some study characteristics, they should be more carefully defined, and a further trial implemented. Trial runs may also reveal that the preliminary coding scheme is incomplete in its coverage of study attributes.

COMPUTATION OF EFFECT SIZES. To be included in the synthesis, a study must contain a report of a quantitative test of the hypothesis that is under scrutiny. In the best case these reports are quite precise (e.g., means and standard deviations, F-tests), but some reports may be ambiguous (e.g., "the groups did not differ"). Nonetheless, the goal is to convert summary statistics into effect sizes that can be statistically integrated. Most studies precisely report the examined relation by one or more of the following statistics: (a) means (M) and standard deviations (SD); (b) t-tests; (c) F-tests (ANOVAs); (d) r-values (e.g., Pearson, point-biserial, tetrachoric); (e) χ^2 values; (f) proportions or

frequencies; and (g) p-values. Ordinarily, each of these statistics can be converted with relative ease into an effect size, although some types of statistical information are more precise than others. In particular, p-values can be quite imprecise if reported only as a p-level (e.g., $p < .05$). Obviously, exact statistics are preferable to inexact statistics for calculating effect sizes. If only an inexact statistic is given, one possibility is to contact the authors of the study for more precise information. If the only information available is imprecise, it should nonetheless be used so that the study outcome can be included in the meta-analysis, as Rosenthal (1991) recommended. Even the information that the relevant finding was nonsignificant should be preserved, in the absence of a more precise statistical report (see *Dealing With Nonreported Results*, below). We list often-used effect size transformations of study information below.

EFFECT SIZE INDEXES. The most commonly used effect size indexes are the standardized difference and the correlation coefficient (see Rosenthal, 1991, 1994). The standardized difference, which expresses the finding in standard deviation units, was first proposed by Cohen (1969) in the following form:

$$g = \frac{M_A - M_B}{SD}, \tag{1}$$

where M_A and M_B are the sample means of two compared groups, and SD is the pooled standard deviation.[1] Because this formula overestimates population effect sizes to the extent that sample sizes are small, Hedges (1981) produced a correction for this bias, $d = J(m)g$, where d is an unbiased estimator of the population effect size, and $J(m)$ is the correction,

$$J(m) \approx 1 - \frac{3}{4m - 1}, \tag{2}$$

where m is $n_A + n_B - 2$, the degrees of freedom. To distinguish between corrected and uncorrected effect sizes, we follow Hedges and Olkin's (1985) convention of referring to Cohen's uncorrected effect size as g and Hedges's corrected effect size as d.

[1] As a supplement to examining differences between means, it is possible to examine differences in variances between two groups. This procedure is particularly useful when the variances are not homogeneous between the groups that are compared. Procedures that take variance differences as well as mean differences into account appear in articles by Feingold (1995) and Hedges and Friedman (1993)

A common equation for the correlation coefficient, r, is

$$r = \frac{\sum_{i=1}^{N} z_{X_i} z_{Y_i}}{N}, \tag{3}$$

where z_{X_i} and z_{Y_i} are the standardized forms of X and Y being related for each case i, and N is the number of observations. Like d-values, r-values have a bias, underestimating the population effect size especially for studies with small samples and for r-values near .60. Therefore, it is appropriate to implement the following correction for bias:

$$\tilde{G}_{(r)} \cong r + \frac{r(1 - r^2)}{2(n - 3)}, \tag{4}$$

where $\tilde{G}_{(r)}$ is the approximation of the population effect size and n is the sample size. Yet, because this bias correction is very small for sample sizes higher 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 (see Hays, 1988), it is conventional in meta-analysis to use Fisher's (1921) r-to-Z transform and to perform meta-analytic calculations on the Z-values,

$$Z_r = \frac{1}{2} \log_e \frac{1 + r}{1 - r}, \tag{5}$$

where $\log_e$ is a natural logarithm operation and r is corrected via Equation 4. Then, following operations on Z_r, Fisher's (1921) Z-to-r transform is used to convert a Z_r-value back into r,

$$r = \frac{e^{(2Z_r)} - 1}{e^{(2Z_r)} + 1}, \tag{6}$$

where e is the base of the natural logarithm.

Choice of Effect Size Index for Meta-Analysis

Because r can be transformed into d (in its g form),

$$g = \frac{2r}{\sqrt{1 - r^2}}, \tag{7}$$

and g into r,

$$r = \frac{g}{\sqrt{g^2 + 4}}, \tag{8}$$

the choice of an effect-size metric for meta-analysis may seem somewhat arbitrary. Some analysts who prefer r argue that it is more immediately interpretable than d (e.g., Mullen, 1989; Rosenthal, 1991, 1994). Others believe that ds are just as interpretable, because they are expressed in units of the standard deviation and are therefore a type of standard score. Despite these considerations of ease of interpretation, the convention

is to use r as the effect size if most of the studies that are integrated report correlations between two continuous variables. If most of the studies report ANOVAS, t-tests, and chi-squares for comparisons between two groups (e.g., experimental vs. control; women vs. men), analysts typically use d.

THE SIGN GIVEN TO EFFECT SIZES. The positive or negative sign of the effect sizes computed in a meta-analysis is defined so that studies with opposite outcomes have opposing signs. Ordinarily, in research literatures in which groups are compared, the positive sign is given to outcomes in the expected, hypothesized, or typical direction for the meta-analysis as a whole, and the negative sign is given to outcomes that reverse this direction. Only a relation that is exactly null would have no sign, because the effect size would be 0.00. Illustrating this practice is Kite and Whitley's (1996) meta-analysis of sex-related differences in attitudes toward homosexuals, in which the expected direction of the findings was that women would evaluate homosexuals more positively than do men. Therefore, the positive sign was given to effect sizes indicating that women's evaluations were more positive than men's, and the negative sign was given to effect sizes indicating that men's evaluations were more positive than women's. Alternatively, in research literatures in which experimental groups are compared with control groups, differences in favor of the experimental group might be given a positive sign, and differences in favor of the control group given a negative sign, regardless of the hypothesis or typical direction of the findings. Also, for meta-analyses of correlational studies, the positive sign is generally given to positive associations between the two focal variables, and the negative sign to negative or inverse associations.

MULTIPLE REPORTS FROM INDIVIDUAL STUDIES. 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, the analyst must decide whether to average the effect sizes that can be computed in order to represent the study with a single effect size estimate. To preserve the independence of the effect sizes in a meta-analysis, they must each come from a different study. That is, the participants whose data contribute to a given effect size must not contribute to any other effect sizes in the analysis. Therefore, the analyst would ordinarily average multiple effect sizes calculated from a single study, especially if the goal of the meta-analysis is to examine uncorrected effect sizes (see *Correcting effect sizes for biased methods*, below). Alternatively, Rosenthal and Rubin (1986) described a procedure to calculate a combined effect size, $g_{combined}$, for the study such that

$$g_{combined} = \frac{\sum_{i=1}^{N_r} g_i}{\sqrt{r_{dv} N_r^2 + N_r(1 - r_{dv})}}, \quad (9)$$

where N_r is the number of measures, g_i is the effect size for each measure i, and r_{dv} is either the correlation between the two measures or the average correlation between the measures, in the case of three or more measures. This equation relies on the Spearman-Brown prophecy equation, which dictates that augmenting the reliability of variables increases the magnitude of their observed associations (see Cronbach, 1990). Of course, the use of Equation 9 may be precluded in many instances because the requisite correlation or correlations between measures are not reported.[2]

Instead of or in addition to averaging effects within studies, the analyst may wish to investigate whether the results of the studies varied depending on the different operations by which their dependent variables were defined. For this purpose, the analyst should preserve the separate effect size estimates made within individual studies, in order to perform a subsequent analysis examining whether the operations produced differences in the effect sizes. For example, in a meta-analysis of sex differences in leaders' effectiveness, Eagly, Karau, and Makhijani (1995) analyzed effect sizes according to the identity of the raters who provided the effectiveness measure and the basic type of measure (e.g., objective vs. subjective). Although many individual studies contributed several effect sizes to these analyses, each study's effect sizes were subsequently aggregated into a single study-level effect size that was used in additional analyses that did satisfy the assumption that effect sizes are independent. Analyses using multiple effect sizes from single studies can certainly be informative, but they should be interpreted cautiously because they violate the assumption of independence of the effect sizes (see *Consequences of violations of the assumption of effect size independence*, below).

When a study examined the relation of interest within levels of another variable, effect sizes may be calculated within the levels of this variable as well as for the study as a whole. How seriously the use of such within-level effect sizes violates the independence assumption depends on whether these levels were created on a within-subjects basis or a between-subjects

[2] In the absence of the relevant correlation, analysts may wish to estimate the correlation from the subset of studies that provide this information.

basis. If the same participants took part at all levels of the variable (i.e., a within-subjects variable), the effect sizes would be highly dependent. The effect sizes would also be dependent if one control group serves as a comparison for more than one treatment group. Even if the participants at the different levels were not the same individuals, the effect sizes would be dependent because they came from the same study, which was carried out under conditions existing in a particular place at a particular point in time (Hedges, 1990). For example, effect sizes might be calculated separately for the male and female participants in experiments in order to examine sex-related differences in the relation (e.g., Eagly, Ashmore, Makhijani, & Longo, 1991; Karau & Williams, 1993), even though these effect sizes would not be independent.

Finally, reports may 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, and, as long as the effect sizes are similar, take a simple average of them. Yet, the analyst should 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.

Estimating Effect Sizes

A comprehensive treatment of the formulas to convert primary-level statistics to effect sizes is beyond the scope of this chapter (see Cooper & Hedges, 1994b; Glass et al., 1981; Johnson, 1993; Rosenthal, 1991). Here we offer only the most common transforms for deriving g. For producing r from various statistical reports, Glass et al. (1981) provided several useful formulas; alternatively, g may be calculated and transformed to r by Equation 7.

EFFECT SIZES FROM MEANS AND STANDARD DEVIATIONS. Equation 1, which appears in the subsection *Effect size indexes*, transforms two means and a standard deviation into an effect size. However, there are many possible forms for defining the standard deviation that appears in the dominator of the formula. To derive g from means and standard deviations in a between-subjects design, it is conventional to use the pooled standard deviation, SD, computed as follows:

$$SD = \sqrt{\frac{(n_A - 1)(SD_A)^2 + (n_B - 1)(SD_B)^2}{n_A + n_B - 2}}, \qquad (10)$$

where n_A and n_B are the number of observations in the two groups being compared, and SD_A and SD_B are their standard deviations (see Glass et al., 1981). Thus, SD represents the square root of a "pooling" of the variances of the two groups and is an identical variability estimate to that obtained when an F- or t-test evaluates the difference between the means of the two groups.

For within-subjects designs, SD can be replaced with SD_d, the standard deviation of the differences between paired observations,

$$SD_d = \sqrt{SD_A^2 + SD_B^2 - 2r_{EC}SD_ASD_B}, \qquad (11)$$

where r_{EC} is the correlation between the paired observations (e.g., Dunlap, Cortina, Vaslow, & Burke, 1996). Most often all of the components of this formula are not provided, and a paired-observation t-test or a within-subjects F is given instead. As we indicate in the next subsection, these statistics may be directly converted into the effect size that has the standard deviation of the differences in its dominator.[3]

As a rule, whenever possible, SD should be estimated only from the portion of each study's data entering into the effect size. For example, if the $M_A - M_B$ difference needs to be calculated within a level of another variable, SD should be estimated from the standard deviations given for participants within this level, if this information is available. Often, however, SD is available only pooled across all of the conditions of an experiment. If the SD pooled within the cells of the design is not available, but the report contains a standard deviation for the overall sample, it should be converted to the pooled SD by removing the variance due to the difference between M_A and M_B (e.g., Hedges & Becker, 1986; Johnson, 1993).

[3] An alternative is to use the within-groups pooled standard deviation as the denominator. Recommending this strategy, Dunlap et al. (1996) concluded that using the means and the pooled SD provides the least biased effect size estimate. However, this calculation is often not possible because the studies provide only a within-subjects summary statistic such as F or t. Converting such statistics into the pooled SD requires the correlation between the paired observations, which is typically not available in the studies' reports.

EFFECT SIZES FROM *t*- AND *F*-VALUES. Calculations of *g* can also be based on summary statistics. In the case of the *t*-test for independent groups,

$$t = \frac{M_A - M_B}{\sqrt{\frac{SD_A^2}{n_A} + \frac{SD_B^2}{n_B}}}. \quad (12)$$

Rearrangement of the terms of this equation produces the following formula for calculating *g*:

$$g = t\sqrt{\frac{n_A + n_B}{n_A n_B}}. \quad (13)$$

Or, if $n_A = n_B$,

$$g = t\sqrt{\frac{2}{n}} = \frac{2t}{\sqrt{2n}}. \quad (14)$$

Because $t = \sqrt{F}$ for a comparison of two groups, when the *F* results from a between-subjects design with unequal *n*,

$$g = \sqrt{F\frac{n_A + n_B}{n_A n_B}}. \quad (15)$$

Or, if $n_A = n_B$,

$$g = \sqrt{\frac{2F}{n}}, \quad (16)$$

where *n* is the within-cell *n* (not the total *N*). If a within-subjects *t* (i.e., for paired observations) is reported,

$$g = \frac{t}{\sqrt{n}}. \quad (17)$$

When a study reports an *F* for a two-groups within-subjects comparison,

$$g = \sqrt{\frac{F}{n}}. \quad (18)$$

F-values that derive from designs with three or more conditions require some special consideration. *F*-values that have more than one degree of freedom in the numerator cannot be directly converted into effect sizes because they do not directly gauge differences between individual means. Rather, a significant omnibus *F*-value implies that somewhere among the relevant means, one or more significant differences exist (see Judd, this volume, Ch. 14). Thus, for example, a significant *F*-value from a design that uses low, medium, and high levels of the independent variable must be decomposed in order to permit effect size derivations. If a linear contrast is reported, it will be equivalent to a comparison between the high and low levels. One could compare the means only for the high and low levels or also compare the medium level with the low and the high levels (e.g., Rhodes & Wood, 1992). Or, if the relation between the independent and dependent variables is expected to be linear, one could compute an *F* for the linear trend in the means and transform it into *g* (see Glass et al., 1981; Rosenthal & Rosnow, 1985). Of course, analysts should use the means in a particular study that would produce the most similar comparison to that used to represent the other studies in the sample. Treating studies' results in substantially different ways would introduce noise into the effect sizes in the database.

Similar issues arise in designs with two or more factors. In such instances, in order to make effect size comparisons more similar across the studies in a meta-analytic sample, some methodologists have recommended producing one-way designs by returning the effects of irrelevant factors to the error term of the ANOVA (Glass et al., 1981; Hedges & Becker, 1986; Morris & DeShon, 1997). This procedure should be seriously considered for individual-difference variables that were crossed with the crucial independent variable in only some of the studies, because this source of variability would not have been removed from the error term in studies that did not assess these individual differences. When these irrelevant variables were instead manipulated, the decision is less straightforward, to the extent that researchers have created extreme conditions atypical of natural settings by means of powerful experimental manipulations. Variability due to extreme or atypical conditions would not be in the error term of typical studies. Therefore, adding sums of squares for such manipulated variables to the sum of squares error could greatly inflate these error terms in at least some instances and thus decrease the absolute magnitude of effect sizes based on these error terms. As Morris and DeShon concluded, in deciding whether to return irrelevant factors to the error term, analysts should keep as their goal the production of error terms that are based on the same sources of variability across the studies in the sample.

To illustrate how to return irrelevant factors to the error term, Table 19.1 contains a hypothetical ANOVA for a two-factor design. The top panel contains the ANOVA summary for the two factors. Suppose that Factor *A* is the focal independent variable, and that Factor *B* is a meta-analytically irrelevant variable. To represent the impact of Factor *A* on the dependent variable, the variation due to Factor *B* can be returned to the error term. This operation is performed by (a) adding the sum-of-squares due to Factor *B* and its interaction with

Factor *A* to the error sum-of-squares and (b) adding the degrees of freedom due to Factor *B* and its interaction to the degrees of freedom for error. Once the sum-of-squares for error has been divided by its new degrees of freedom, the square root of the resulting mean-square for error would be interpretable as the standard deviation pooled within the two levels of *A*, or $SD = \sqrt{MS_e}$. The result of this reconstitution of the error term appears in Panel b. In this example, g may be derived by converting the F-value that resulted from the reconstitution procedure, or it may be derived by dividing the difference between the means of Factor *A* by *SD*. Morris and DeShon (1997) presented other equations and examples of this strategy; Nouri and Greenberg (1995) presented techniques for use with more complex ANOVA designs (e.g., those that mix between- and within-subjects factors).

TABLE 19.1. Hypothetical Analysis of Variance Summary Tables (a) Before Reconstitution and (b) After Returning Factor B's Sums of Squares to the Error Term

Source	Sum of Squares	Degrees of Freedom	Mean Squares	*F*
		(a) Before reconstituting		
A	430.33	1	430.33	15.22
B	200.12	1	200.12	7.08
A × B	43.55	1	43.55	1.54
Error	1244.29	44	28.28	
		(b) After reconstituting		
A	430.33	1	430.33	13.30
Error	1487.96	46	32.35	

If the effects of the focal independent variable on the dependent variable are expected to change within the levels of another independent variable, separate effect sizes can be calculated within levels of the second independent variable, as we already mentioned above (see subsection *Multiple reports from individual studies*). Specifically, as an alternative to representing the effect of the focal independent variable aggregated over this other variable (i.e., as a main effect), the analyst can partition each study on this other variable and represent the effect of interest within levels of this variable (i.e., as a simple main effect). When interactions are expected, simple main effects are the desired comparison, and the other, interacting variable can function as a moderator of the relation between the focal variables. As an example, Table 19.2 displays a 2 × 3 factorial design in which the focal independent variable (IV_{focal}) and a moderator variable ($IV_{moderator}$) serve as the factors. Suppose that we expect the effect of IV_{focal} on the dependent variable to change depending on the level of $IV_{moderator}$. To represent these contrasting expectations, a separate effect size must be derived for each level of $IV_{moderator}$. Thus, the first g would result from a comparison of the means from cells *a* and *b*, the second from cells *c* and *d*, and the third from cells *e* and *f*. In order to perform this calculation, it is necessary to obtain all cell means and either (a) the within-cell standard deviations, (b) the standard deviations for each relevant level of $IV_{moderator}$ (and transformed to SD_{pooled}), or (c) MS_e for the ANOVA. The MS_e can be recovered when all cells means are reported and at least one F-value is known for the dependent variable, even when the available F is not the most relevant to the analysts' focal comparison (Johnson, 1993; Morris & DeShon, 1997). These calculations are facilitated if the source report contains a complete ANOVA table, but the components of the table can be estimated if the means, cell sizes, and one or more F-values are known (Johnson, 1993). Then, $SD = \sqrt{MS_e}$. Once this value or the standard deviations are known, effect-size derivations continue as though each condition were a separate experiment.

Finally, F-values derived from multivariate analysis of variance (MANOVA), in which one or more independent variables were examined for their simultaneous influence on two or more dependent measures, should not be transformed into effect sizes if the dependent variable of interest was combined with other, irrelevant dependent variables (see Morrison, 1976; Timm,

TABLE 19.1. A Hypothetical Factorial Design in Which a Focal Independent Variable Is Crossed with a Moderator Independent Variable

		IV_{Focal}	
		Level 1	Level 2
$IV_{Moderator}$	Level 1	Cell *a*	Cell *b*
	Level 2	Cell *c*	Cell *d*
	Level 3	Cell *e*	Cell *f*

1975). If several measures of the same conceptual dependent variable were combined in a multivariate analysis, however, the analyst might derive an effect size by taking the square root of the proportion of variance that the independent variable accounts for in the best linear combination of the dependent variables and treating this value as an r (see Tabachnick & Fidell's, 1996, pp. 388–391, discussion of Wilk's Lambda), even if univariate F-values from ANOVAs are not available. However, because such effect sizes would be dependent on the exact set of dependent variables included in the multivariate analysis, some meta-analysts recommend against such procedures (Hunter & Schmidt, 1990).

This section about t- and F-values shows that complex statistical considerations can arise in translating source reports into effect sizes. Because of these potential complexities, a reviewer should never proceed to calculate effect sizes from an ANOVA without thoroughly understanding the design used for the data analysis. The reviewer would be well advised to diagram the design with the relevant ns. Because multiple error terms are common in the designs used in experimental social psychology, it is easy to use the wrong error term for calculating the effect size. To prevent such errors, advanced ANOVA texts are invaluable (e.g., Myers & Well, 1991; Winer, Brown, & Michels, 1991). For reference purposes, meta-analysts may find it convenient to produce a packet of the clearest textbook descriptions of designs that occur often in their literatures.

EFFECT SIZES FROM r-VALUES. Although r can be readily transformed to g by Equation 7, correlational reports often appear in a form other than r (see Carroll, 1961; Cohen & Cohen, 1983; Glass et al., 1981; Rosenthal, 1991, 1994). When r-values other than the product-moment variety are reported (e.g., biserial r, phi coefficient), they can usually be interpreted as product-moment rs, except when they are point-biserial rs. In this case, the meta-analyst would convert the point-biserial r into the biserial r, which approximates the product-moment r. If $n_A = n_B$, or when n_A is approximately n_B, $r_b = 1.253r_{pb}$, or, if $n_A \neq n_B$,

$$r_b = \frac{r_{pb}\sqrt{n_A n_B}}{\mu N}, \tag{19}$$

where N is the total sample size, and μ is the ordinate of the unit normal distribution (i.e., the height of normal curve with surface equal to 1.0 at the point of division between segments containing n_A and n_B cases). Similarly, if a study reports t calculated based on any r value, the t can be converted to a product-moment correlation using

$$r_b = \frac{r_{pb}\sqrt{n_A n_B}}{\mu N}. \tag{20}$$

Whereas standardized regression weights (β) deriving from simple linear regressions are r-values and can be so interpreted, βs deriving from regressions with more than one predictor *cannot* be directly interpreted as r-values. The β-value for a given predictor in a multiple regression equation is *adjusted* for the other independent variables present in the equation. In the case of suppressor variables (Cohen & Cohen, 1983), these adjustments can affect not only the value of β but also its sign, which could be reversed from the sign of the correlation between the two variables. Yet another problem with converting β-values to effect sizes is that under some circumstances β-values from multiple regression equations exceed $|1|$, whereas r-values never exceed $|1|$. For example, if Equation 7 is used with a β of 1.1, the denominator of the equation will be the square root of a negative number, -0.21, which is an irrational mathematical operation. Therefore, as a general rule, in meta-analyses for which multiple regression results are the exception and other studies in the sample report statistics unadjusted for the other variables in the equation, multiple regression results should not be converted to effect sizes (see Hunter & Schmidt, 1990). Of course, before discarding a study because its findings were reported in a multiple regression, one should see whether a correlation matrix or comparable statistics appear in the report or could be obtained from its authors.

If many of the studies in a literature contain multiple regression equations that use the same conceptual independent variables to predict the same conceptual dependent variable, syntheses could pursue two strategies. One alternative is to examine how much variance (estimated by multiple R^2) was explained in the criterion variable by the set of predictor variables. For example, an analyst might examine each study to determine how much variance in intentions to perform a behavior was explained by the simultaneous impact of attitudes toward performing the behavior and normative expectations about the behavior (see Sheppard, Hartwick, & Warshaw, 1988). Hedges and Olkin (1985, p. 239) provide an alternative strategy that relies directly on the βs and their sample sizes to produce an aggregate weighted beta-weight.

EFFECT SIZES FROM CHI-SQUARE VALUES. Chi-square (χ^2) values are sometimes used to test for the frequency with which groups meet some criterion or

to test for the association between two variables (Hays, 1988). If the χ^2 results from a 2 × 2 classification table linking a predictor (X) to the outcome (Y),

$$r_\phi = \sqrt{\frac{\chi^2}{n}}, \tag{21}$$

where r_ϕ is a phi coefficient, which approximates the product-moment r and can be converted to g by Equation 7. However, if there is more than 1 degree of freedom in the χ^2 value, it cannot be directly converted into an effect size because the χ^2 may describe a nonlinear pattern. It may be possible to compute χ^2 for an appropriate 2 × 2 table based on the proportions of the relevant groups that meet a criterion (see next subsection). If the data for these re-computations are not available, the study result cannot be used to derive an effect size.

EFFECT SIZES FROM PROPORTIONS MEETING A CRITERION. In some designs, the proportion of individuals in one group (p_E) who meet a given criterion is compared with the proportion in another group (p_C) who meet it. For example, the proportion of people who help another person in one experimental condition can be compared to the proportion of people who help in another condition (see Eagly & Crowley, 1986). These proportions can be transformed into an effect size by using a probit transformation (Glass et al., 1981) or by treating the proportions as means (Snedecor & Cochran, 1980) such that

$$g = \frac{p_A - p_B}{SD}, \tag{22}$$

where

$$SD = \sqrt{\frac{(n_A - 1)p_A q_A + (n_B - 1)p_B q_B}{n_A + n_B - 2}}, \tag{23}$$

where $q_E = 1 - p_E$ and $q_C = 1 - p_C$. Note that Equations 22 and 23 assume that the proportions are in relation to the study's unit of analysis, which is generally persons. The equations do not apply to proportions that represent values of dependent variables assessed for each unit of analysis. For example, if each participant's helping were assessed a self-report of the proportion of occasions on which he or she helped, these data would produce an effect size by equations that use the variability of these proportions (e.g., Equation 1) rather than Equation 23.

EFFECT SIZES FROM PROBABILITIES ASSOCIATED WITH INFERENTIAL STATISTICS. Source reports sometimes contain only a p-value associated with the critical effect (e.g., $p = .0439$), which can be used to calculate an effect size if the direction of the finding and the sample size (n) are known. To do so, the analyst would use a statistical package's (e.g., SAS, IMSL, SPSS, Stata) inverse probability distribution functions, which provide an exact solution of a test statistic from p. For example, SAS provides BETAINV, which yields F from p and df, after which the F can be converted to g using Equation 15, 16, or 18 (assuming that the F compares the means of only two groups). Obviously, an exact p allows an excellent estimate of a test statistic and therefore of g. However, a level p (e.g., $p < .05$) gives a poorer estimate, because it would ordinarily be treated as exactly the p level given (e.g., $p < .01$ would be understood as $p = .01$). The mere statement that a finding is "significant" can be treated as $p = .05$ in studies that apparently use the conventional $p < .05$ rule for determining significance and indicate the direction of the effect, but the effect sizes estimated on this basis may be quite inaccurate (Ray & Shadish, 1996). Finally, reports often differ in whether a one-tailed or two-tailed probability level is reported; if no information is provided, the convention is that the study authors have used a two-tailed test.

DERIVING EFFECT SIZES FROM STATISTICALLY IMPOVERISHED REPORTS. For many reasons, some source reports will contain less than desirable amounts of information for calculating effect sizes. We have tried to make readers aware of certain "tricks" for deriving effect sizes in such cases, but it is difficult to anticipate all possible problems. Some routes to effect sizes merely require a great deal of effort on the part of the analyst (e.g., reanalyzing raw data found in an appendix of a dissertation). In other instances, deriving an effect size may require the application of several nonroutine techniques in sequence. Each meta-analysis poses different statistical challenges that may call for novel solutions.

Whatever the problems that analysts encounter in trying to calculate effect sizes, we urge them to contact studies' authors, if possible, to acquire any essential information that is not included in a report. In our experience, cordial invitations for authors to provide the requisite information have produced moderate success rates. Obtaining such information allows the report to be represented to fullest advantage in the sample; failing to obtain the needed information renders the meta-analysis less comprehensive and potentially less representative and less valid. Finally, it is important to realize that a lack of statistical detail in reports does not necessarily reflect their authors' oversights, errors, or poor methods. Rather, omissions generally occur

because the authors' goals differed from those pursued in a subsequent meta-analysis. For example, to a study's authors, a small sex-of-participant effect on helping behavior might have been incidental and worthy of no more than a footnoted p-value or statement that the effect was nonsignificant; to a meta-analyst pursuing sex-related differences in helping behavior (e.g., Eagly & Crowley, 1986), such findings are crucial.

DEALING WITH NONREPORTED RESULTS. Reports that describe the effect of interest merely as "nonsignificant" are highly problematic in meta-analysis (Bushman & Wang, 1996). It is common to represent such effects by $d = 0.00$, but such estimates are obviously poor. If the N in the study was small, its actual effect size could be quite large and yet not be significant. Introducing such effect sizes into a meta-analysis would yield a mean effect size that underestimates the population value (Schmidt, 1996). Especially if many of these reports exist in a literature, it is advisable to estimate mean effect sizes with and without these 0.00 values. It also may be better to omit these values when attempting to fit models to the effect sizes.

At the synthesis stage of a meta-analysis, one way to incorporate imprecisely reported results, including those described as nonsignificant, is to use so-called "vote-counting procedures" to summarize findings (Bushman, 1994; Bushman & Wang, 1996). In these procedures, rather than using relatively exact effect size estimates to represent the studies' outcomes, an analyst examines how many studies obtained a result in the hypothesized direction or how many obtained a significant result in this direction. Because the strategy relies only on findings' directions or significance levels, it allows an analyst to include even the imprecisely reported nonsignificant results. More formally, calculating what is sometimes called the "sign test," which is based on the binomial distribution, entails determining the exact p of obtaining the observed distribution of positive and negative outcomes (or one more extreme), given that the probability of obtaining a positive result is .5, according to the null hypothesis, which specifies that half of the results should be positive and half negative. This probability can be calculated by standard statistics packages (e.g., SAS's PROBBNML function). An analyst can also use the binomial distribution to calculate a p-value for obtaining the observed distribution of significant positive findings versus other findings (nonsignificant and reversed), given that the probability of obtaining a significant result in one tail of the distribution is .025, according to the null hypothesis and assuming .05 for two-tailed significant testing. The p-values associated with the proportion of the studies that have a positive direction or that produced a significant positive result can be used to estimate a mean effect size (e.g., d) for a sample of studies. These estimated effect sizes can then be compared to the exact mean effect size based on the studies that permitted this calculation (see Bushman, 1994; Bushman & Wang, 1996). For example, Wood (1987) used these techniques to estimate the mean effect size for sex-related differences in group performance, because many of the studies did not permit an effect size to be estimated. Of course, it is much better to calculate the mean effect size by averaging effect sizes from individual studies when the majority of studies permit this strategy.

RELIABILITY OF EFFECT SIZE CALCULATIONS. We strongly recommend that at least two analysts compute effect sizes for each of the studies independently and then meet to compare solutions and resolve discrepancies. Given the complexity of many research designs and the ambiguity of some research reports, various errors of effect size estimation do occur. Moreover, sometimes one analyst may discover an indirect route to computing an effect size that is missed by a second analyst. When two or more analysts calculate effect sizes, these errors and omissions can be minimized (see also *Reliability of coding*, above).

CORRECTING EFFECT SIZES FOR BIASED METHODS. In addition to correcting the raw g and r because they are biased estimators of the population effect size (see prior subsection *Effect size indexes*), analysts may correct for many other biases that accrue from the methods used in each study. For example, as already noted, the Spearman-Brown prophecy equation (see Cronbach, 1990) dictates that as the reliability of a measure increases (and its measurement error therefore decreases), its relations with other variables will also increase. Increased measurement error decreases a measure's ability to predict another variable. Corrections for measurement unreliability and other forms of error or bias can be implemented in a meta-analysis, in order to estimate what the strength of a relation would be in the absence of such artifacts. In their presentations of such corrections, Hunter and Schmidt (1990, 1994; 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 population effect size – that is, its value had studies not been affected by these biases.

These corrections are quite popular in industrial and organizational psychology, particularly for meta-analyses on the validity generalization question of whether the validity of personnel selection tests varies across organizations (e.g., Hunter, Schmidt, & Hunter, 1979). Such corrections have seldom been used in social psychological meta-analyses because in most research areas relatively few studies include the information that would be required to perform all of the corrections. For example, the majority of the study reports generally do not include information on the reliability or validity of the independent or dependent variables. In research literatures that are more psychometrically developed in the sense that reliabilities and other relevant information are routinely provided, meta-analysts may be able to perform such corrections.

When meta-analysts do implement these corrections, the resultant corrected mean effect size yields an idealized estimate of the magnitude of the population effect rather than an estimate of the relation that would be obtained in a future study in which the corrections were not implemented. Even when it is possible to implement the corrections within a literature, problems may emerge; in particular, Rosenthal (1991) noted that the corrections can result in irrational effect sizes (e.g., correlations larger than 1.00). Therefore, we recommend that analysts consider their goals when deciding whether to use such corrections. 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 (perhaps the more typical goal for research literatures in which such corrections are seldom implemented), 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, effect size estimates are a ratio 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. 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 that the other study does not (e.g., by statistically controlling for an individual difference characteristic), the first study will have a smaller error term (standard deviation), and the effect size for this study will be larger than that for the second study. To minimize this type of variation in effect sizes, our recommendations regarding effect size derivation have emphasized equating as much as possible the comparisons that the studies yield, so that the effect sizes are not impacted by differing statistical operations. For example, one such recommendation was that in meta-analyses of experimental effects, analysts return irrelevant individual difference factors to the error term. Of course, reconstituting the error term in this way would not be necessary if the variable in question were controlled in all of the studies in the review. In such circumstances, the conclusions of the synthesis should take the presence of this controlled factor into account.

Additional problems can arise from the inclusion of studies that used within-subjects designs. For example, a researcher might have implemented a within-subjects design that required each participant to judge two objects along the same dimension. Such multiple assessments can produce many complications, including carryover, priming, and contrast effects (E. Smith, this volume). In analyzing such data, researchers nearly always use a repeated-measures inferential statistic that removes from the error term the variation due to the individual participants (e.g., Equation 11). Consequently, these tests are more statistically powerful than those produced by a comparable between-subjects design (Dunlap et al., 1996). If the meta-analyst uses these within-subjects error terms to calculate effect sizes, it is likely that these effect sizes will be larger than those based on standard deviations pooled from the cells of the design (e.g., Kite & Johnson, 1988; for an exception, see Symons & Johnson, 1997). Therefore, in such circumstances, analysts should use type of design as a moderator variable.

Although it is unrealistic for analysts to take into account all potential sources of bias in a meta-analysis, they should remain aware of biases that may be important within the context of their research literature. Some of these biases can be corrected in the process of computing the effect sizes. Others can be examined empirically for their influence on studies' results. Still others can be eliminated by narrowing the boundaries of the literature under investigation to exclude biased studies. When it is not possible to control a bias in some fashion, analysts should consider what influence a bias might have exerted on their findings and interpret the results accordingly.

Analyzing the Meta-Analytic Database

PRELIMINARY CONSIDERATIONS. The general steps involved in the analysis of effect sizes usually

are the following: (a) aggregate effect sizes across the studies to determine the overall strength of the relation between the examined variables; (b) analyze the consistency of the effect sizes across the studies; (c) diagnose statistical outliers among the effect sizes; (d) examine visual displays of the distribution of effect sizes to determine whether any irregularities exist; and (e) perform tests of whether study attributes moderate the magnitude of the effect sizes. The three principal meta-analytic frameworks for analyzing effect sizes that have developed over the past twenty years are those proposed by Hedges and Olkin (e.g., 1985), Rosenthal (e.g., 1991), and Hunter and Schmidt (e.g., 1990). Although each framework encompasses valuable recommendations for the conduct of meta-analysis (for reviews, see Aguinis & Pierce, 1998; Johnson et al., 1995; Sánchez-Meca & Marín-Martínez, 1997), we will focus on the Hedges and Olkin (1985) approach in this section, while occasionally adding elements of the other approaches.

Although our exposition considers only fixed effects models, random effects models or blends of random and fixed effects models may also be useful in some contexts (Hedges & Vevea, 1998; Overton, 1998; Raudenbush, 1994). This fixed versus random distinction may be familiar to readers from discussions of ANOVAs for primary research. In meta-analysis, as in primary research, the model used depends on the type of generalization that the analyst desires to make. Fixed effects models permit an analyst to generalize results only to groups of studies identical (or at least quite similar) to those in the meta-analytic sample, except for the particular participants who appear in the studies. In contrast to fixed effects models, random effects models assume that the studies in the meta-analysis are randomly sampled from some population of studies and permit an analyst to generalize to this population. As Hedges (1994) and Rosenthal (1995) explained, there are pros and cons of treating meta-analytic data by fixed effects or random effects models. Moreover, there are few conventions concerning when to use fixed or random effects models (Cooper, 1998). As a practical matter, nearly all meta-analyses to date have used fixed effects models, and the computational techniques for these models are simpler and have been worked out more completely. Nonetheless, some methodologists argue that, in comparison to random effects models, fixed effects models manifest greater Type I bias in significance tests for the mean effect size and moderator relations and thus are insufficiently conservative (see Hunter & Schmidt, 1997; Overton, 1998). These considerations suggest a greater use of random effects models (Hedges & Vevea, 1998).

The fixed effects model-testing procedures that we present are analogous to techniques used in data analysis in primary research. Categorical models are analogous to fixed effect analyses of variance, and continuous models are analogous to regression procedures. Yet, the procedures used in meta-analysis differ from those used in primary research in two main respects. The first difference pertains to the heterogeneity of the variances ordinarily associated with the individual effect sizes, which would likely violate the homoscedasticity assumption of conventional regressions and ANOVAs – that is, the assumption that the nonsystematic variance associated with every observation is equal. In regressions, this assumption is checked by evaluating the constancy of the residual variance around the regression line for all values of the predictor variable. In ANOVAs, the within-cell variances are checked to determine if they are similar in value across the cells of the design. In contrast, meta-analytic statistics were designed to take advantage of differing variances by calculating the nonsystematic variance of the effect sizes analytically (Hedges & Olkin, 1985). Because this nonsystematic variance of an effect size is inversely proportional to the sample size of the study and sample sizes vary widely across the studies, the error variances of the effect sizes are ordinarily quite heterogeneous. The second difference between the statistical procedures of meta-analysis and primary research is that meta-analytic statistics permit an analysis of the consistency (or homogeneity) of the effect sizes across the studies, a highly informative analysis not produced by conventional statistics. As the homogeneity calculation illustrates, analyzing effect sizes with specialized meta-analytic statistics rather than the ordinary inferential statistics used in primary research allows a reviewer to use a greater amount of the information available from the studies (Rosenthal, 1991, 1995).

COMPUTATION OF A COMPOSITE EFFECT SIZE. As a first step in a quantitative synthesis, the study outcomes are combined by averaging the d-values with each d weighted by the reciprocal of its variance. The weighted mean effect size d_+ is computed as a weighted average of the individual studies' effect sizes,

$$d_+ = \frac{\sum_{j=1}^{k} w_j d_j}{\sum_{j=1}^{k} w_j}, \tag{24}$$

where k is the number of effect sizes. The variance, v_+ of the weighted mean effect size d_+ is

$$v_+ = \frac{1}{\sum_{j=1}^{k} w_j}, \tag{25}$$

where the weights for each effect size j, w_j, are defined

$$w_j = \frac{1}{v_j} = \frac{2\left(n_j^A + n_j^B\right) n_j^A n_j^B}{2\left(n_j^A + n_j^B\right)^2 + n_j^A n_j^B d_j^2}. \quad (26)$$

Note that v is unrelated to the variance of the raw data that entered into the inferential statistics in the first place. Equation 24, which is a simple fixed effects meta-analytic model, gives greater weight to the more reliably estimated study outcomes, which are in general those with the larger sample sizes (Hedges & Olkin, 1985); other frameworks provide for weighting effect sizes on this and other quality-relevant bases (Hunter & Schmidt, 1990; Rosenthal, 1991). As a test for significance of this weighted mean effect size, a confidence interval is computed around this mean, based on its standard deviation, $d_+ \pm 1.96\sqrt{v_+}$, where 1.96 is the unit-normal value for a 95% CI (assuming a nondirectional hypothesis). 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 significant association between the independent and dependent variable (X and Y). Alternatively, a unit-normal z-value for a weighted mean effect size can be calculated directly, similar to the convention in the Rosenthal (1991) approach,

$$z_+ = \frac{d_+}{\sqrt{v_+}}, \quad (27)$$

and this z_+ can be evaluated by determining if its exact p-value is less than α or by comparing z_+ against the z equivalent to the chosen significance level (see Becker, 1987; Hedges, Cooper, & Bushman, 1992; Johnson et al., 1995).

When the weighted mean effect size and the CI are computed, the homogeneity of the ds should be examined in order to determine whether the studies can be adequately described by a single effect size (Hedges, 1981; Hunter & Schmidt, 1990; Rosenthal, 1991). If the effect sizes can be so described, they would differ only by unsystematic sampling error. The test statistic, Q, which is known as a test of the homogeneity (or heterogeneity) of the effect sizes, evaluates the hypothesis that the effect sizes are consistent,

$$Q = \sum_{j=1}^{k} w_j (d_j - d_+)^2, \quad (28)$$

where k is the number of effect sizes in the class. Q has an approximate χ^2 distribution with $k - 1$ degrees of freedom. If Q is significant, the 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 explanatory work would be merited, even when the composite effect size is significant. The magnitude of individual study outcomes would differ systematically, and these differences may include differences in the direction (or sign) of the relation. 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. Even if the homogeneity test is nonsignificant, significant moderators could be present, especially when Q is relatively large (for further discussions, see Johnson & Turco, 1992; Rosenthal, 1995). Also, Q could be significant even though the effect sizes are very close in value, especially if the sample sizes are very large. These complexities suggest that Q deserves careful interpretation, in conjunction with inspecting the values of the effect sizes. Nonetheless, in a meta-analysis that attempts to determine X's impact on Y, rejecting the hypothesis of homogeneity could be troublesome, because it implies that the association between these two variables likely is complicated by the presence of interacting conditions. However, 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 of the distribution of effect sizes.

TESTING MODELS OF META-ANALYTIC MODERATORS. To determine the relation between study characteristics and the magnitude of the effect sizes, both categorical models and continuous models can be tested (Hedges, 1982a, 1982b; Hedges & Olkin, 1985; Rosenthal, 1991). In the Hedges and Olkin (1985) approach, *categorical models*, which are analogous to ANOVAs, may show that effect sizes differ in magnitude between the subgroups established by dividing studies into classes based on study characteristics. For example, Stangor and McMillan's (1992) meta-analysis found that studies with expectancy-congruent stimuli increased memory relative to studies with expectancy-

incongruent stimuli when the processing goal was to form impressions and that this pattern reversed when the processing goal was to evaluate the presented information. 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 the systematic variability between the effect sizes. Following Hedges and Olkin's (1985) statistical procedures, categorical models provide a between-classes effect (analogous to a main effect in an analysis of variance) and a test of the homogeneity of the effect sizes within each class. The between-classes effect is estimated by Q_B,

$$Q_B = \sum_{i=1}^{p} w_{i+}(d_{i+} - d_{++})^2, \tag{29}$$

where w_{i+} is the reciprocal of variance of d_{i+} for class i, d_{i+} is its weighted mean effect size (using Equation 24 within the class), and d_{++} is the weighted grand mean effect size (Equation 24). Q_B has an approximate χ^2 distribution with $p - 1$ degrees of freedom, where p is the number of classes. To determine the fit of the model, the homogeneity of the effect sizes within each class i around d_{i+} is estimated by Q_{W_i}, which is calculated using Equation 28 with the effect sizes in class i. Q_{W_i} has an approximate χ^2 distribution with $m_i - 1$ degrees of freedom, where m_i is the number of effect sizes in class i. An alternative strategy to examine model specification is to sum the Q_{W_i} values of each class and determine the significance of this total Q value, with $k - p$ degrees of freedom, where k is the number of effect sizes and p is the number of classes. Each Q produced via these strategies is interpreted similarly to the overall Q value, as outlined above. A significant Q value is interpreted as evidence that variability exists in the effect sizes within the class. In other words, the mean effect size provides a poor description of the typical effect size within each class. Also calculated for these models are (a) the weighted mean effect size for each class, calculated with each effect size weighted by the reciprocal of its variance, and (b) a 95% confidence interval for each of these means.

When a categorical model with more than two classes yields a significant Q_B statistic, it is desirable to compute contrasts between the weighted mean effect sizes for these classes (Hedges, 1994; Hedges & Olkin, 1985). For example, Gordon's (1996) meta-analysis found that liking for an ingratiator varied depending on the type of ingratiation tactic used; a contrast showed that studies that used the other-enhancement tactic produced more liking than studies that used the tactic of giving favors. These contrasts, which are analogous to those used in the ANOVA procedure, are approximated by a χ^2 distribution with 1 degree of freedom for a priori tests. For post hoc tests, the same test is used but with more conservative degrees of freedom or a more conservative α. In the Scheffé method, the degrees of freedom are either $p - 1$, where p is the number of classes, or the number of contrasts, whichever value is smaller. In the Bonferroni method, the degrees of freedom term remains 1, but in order for a contrast to be considered significant at level α in the simultaneous analysis, it must be significant at level α/L, where L is the number of contrasts that could have been conducted, which are typically all pairwise contrasts (see Rosenthal & Rosnow, 1985, and Toothaker, 1991, for discussions of contrasts).

Similarly, *continuous models*, which are analogous to regression models, examine whether study characteristics that are assessed on a continuous 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. Continuous models are least squares regressions, calculated with each effect size weighted by the reciprocal of its variance (Hedges, 1982b; Hedges & Olkin, 1985; Hedges, 1994). Each such model yields a test of the significance of each moderator as well as a test of model specification, which evaluates whether variation remains unexplained by the moderators. The error sum of squares statistic, Q_E, which provides this test of model specification, has an approximate chi-square distribution with $k - p - 1$ degrees of freedom, where k is the number of effect sizes and p is the number of predictors (not including the intercept). Tests for the significance of the predictor's association with the effect sizes are given by the unstandardized regression (b) weight(s) in the model. For example, Gordon (1996) found that as the transparency of an ingratiation attempt increased, the success of the attempt decreased. Using Hedges and Olkin's (1985) statistical procedures, an analyst can also fit multiple-predictor models to effect sizes, and the predictors for these models can be either continuous or categorical, or both.[4]

[4] Regression procedures can represent categorical variables using dummy codes (e.g., Cohen & Cohen, 1983; Hardy, 1993). Although a categorical predictor specified by the regression procedure will explain the same amount of variation among the effect sizes as the corresponding categorical model, it would not be possible by regression procedures to determine which specific sub-classes have significant homogeneity statistics. Therefore, categorical models are usually calculated for categorical variables, which are entered into regression procedure (i.e., continuous models) only along with other predictors in a multiple regression.

Analysts may wish to determine the amount of variation that remains unexplained in the effect sizes after one or more moderators have been modeled. For this purpose, Hedges (1994) described the *Birge ratio*, which represents the amount of unexplained variation as a ratio of unexplained variation to degrees of freedom. Thus, for categorical models, the Birge ratio, R_B, is

$$R_B = \frac{Q_W}{k - p}, \tag{30}$$

where Q_W is the sum of the Q_{W_i} for each class in the model; for continuous models,

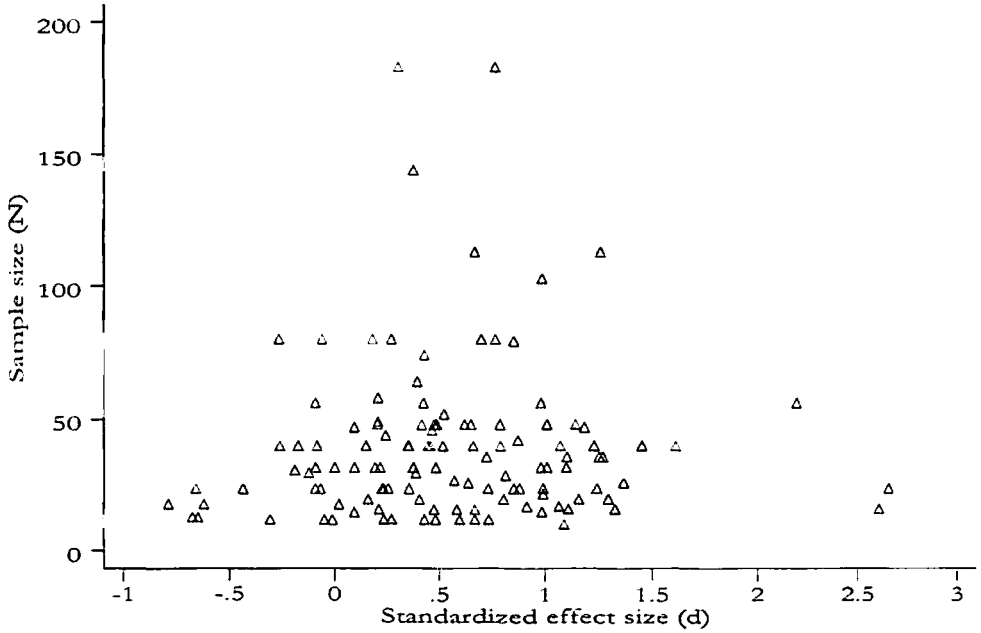
$$R_B = \frac{Q_E}{k - p - 1}. \tag{31}$$

In either case, when the model fits exactly, the Birge ratio has an expected value of 1. A Birge ratio of 1.75, in contrast, suggests that, given the amount of within-study sampling variance, there is 75 percent more between-study variation that is unexplained by the model.

OUTLIER DIAGNOSES. As an alternative analysis to predicting effect sizes using categorical and continuous models, an analyst can attain homogeneity by identifying outlying values among the effect sizes and sequentially removing those effect sizes that reduce the homogeneity statistic, Q, by the largest amount. Using such a procedure for several meta-analyses on psychological topics, Hedges (1987) found that the removal of 20% or fewer of the effect sizes from the heterogeneous sample included in the synthesis usually produced homogeneity. 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 (for e.g., see Stangor & McMillan, 1992; for a more comprehensive treatment of the topic of data outliers, see McClelland, this volume). Under such circumstances, these outliers should be removed from subsequent phases of the data analysis.

VISUAL DISPLAYS. Visual presentations can also assist in the interpretation of meta-analytic results (see Greenhouse & Iyengar, 1994; Light, Singer, & Willett, 1994). Visually examining study outcomes enhances the analyst's potential for finding anomalies in the meta-analytic data. For example, an analyst may determine that effect sizes are related to a continuous

Figure 19.1. Funnel plot when no publication bias is present.



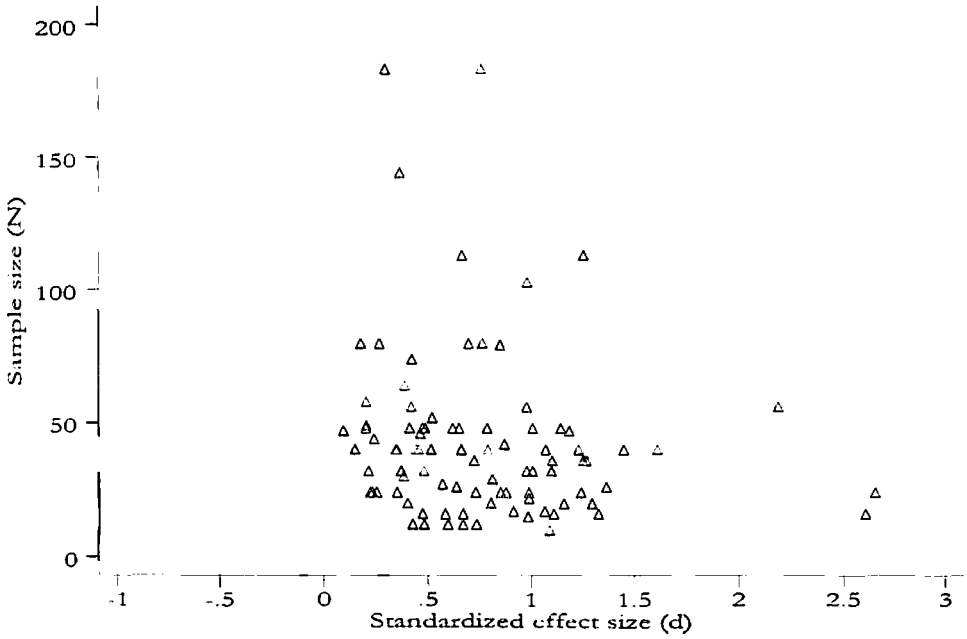


Figure 19.2. Funnel plot that shows a preponderance of significant effect sizes; in such cases, it is possible that there is a publication bias in the literature.

predictor in a nonmonotonic fashion, an outcome that would not be detected by the linear regressions that have been described.[5] Also useful is a *funnel plot* (Light & Pillemer, 1984; Mullen, 1989), which is a scatterplot of sample sizes versus effect sizes. Ordinarily, the scatterplot should take the shape of a funnel sitting on end in the sense that the effect sizes from smaller studies, which are unreliable, would show more scatter than the effect sizes from the larger studies, which would center on the best estimate of the population effect (see Figure 19.1). However, if there is a publication bias in the literature, a funnel plot should reveal few entries in the smaller effect size portion of the graph for smaller sample sizes (see Figure 19.2). Also commonly presented in meta-analytic contexts are some of the exploratory data analysis techniques introduced by Tukey (1977). For example, *stem-and-leaf displays* efficiently plot every effect size in a distribution and are useful for displaying the shape of the distribution. Each effect size appears as a leaf attached to a stem value. The possible stem values of the effect sizes appear as a scale placed to the left of a line and represent their first digit or first two digits (see Figure 19.3). The next digit is the leaf, which is placed to the right of the line. Because each leaf digit to the right of the line represents a separate effect size, the shape of the distribution is displayed. In addition, *schematic plots*, also known as box-and-whiskers plots, show the maximum and minimum effect sizes, the upper and lower quartile values, and the median.

CONSEQUENCES OF VIOLATING THE ASSUMPTION OF EFFECT SIZE INDEPENDENCE. The section on *Multiple reports from individual studies* introduced the meta-analytic assumption of independence among effect sizes and suggested that as a general rule, it is wise to represent studies' participants only once in effect size calculations. Following this recommendation, analysts should ordinarily combine effect sizes representing conceptually similar measures from any given study. If such effect sizes were not combined, the nonindependence that would result could have several effects on the findings of a meta-analysis, depending on the source of the nonindependence (see Gleser & Olkin, 1994). If the nonindependence results from producing more than one effect size from the same participants on conceptually similar measure, the meta-analysis will be likely to reach a liberal estimate of the significance of the weighted mean effect size: Its CI will grow tighter, and its z_+ larger. Similarly, including more

[5] Regression models may include tests of nonlinear associations (see Hedges & Olkin, 1985; Mullen, 1989). However, unless nonmonotonic associations are expected on an a priori basis, they are unlikely to be discovered except by the use of visual displays.

1.8	19
1.7	
1.6	5
1.5	59
1.4	13
1.3	
1.2	4
1.1	34
1.0	22359
0.9	0134689
0.8	2347
0.7	01247788
0.6	001111248
0.5	02333345799
0.4	012233344457889
0.3	011122222344456677889999
0.2	01223444566778899999
0.1	0001111222244556 89
0.0	000000000000000000000001222344566678889999
-0.0	88866443311
-0.1	88732
-0.2	9540
-0.3	421
-0.4	
-0.5	6
-0.6	
-0.7	
-0.8	9
-0.9	
-1.0	2

Figure 19.3. Stem-and-leaf plot of effect sizes rounded to two decimal places. Each effect size is grouped according to its membership in categories defined by the stem, which in this plot is the two digits to the left of the line, ordered from largest at the top to smallest (i.e., largest negative value) at the bottom. Along the stems, each effect size appears as a single leaf digit on the right side of the line, ordered from smallest to largest. In this plot, the first three effect sizes are 1.81, 1.89, and 1.65.

effect sizes from the same groups of participants increases the power of the overall homogeneity test, Q; therefore, the likelihood of rejecting the hypothesis of homogeneity would increase.

Despite these concerns, multiply representing studies may be defensible to address certain meta-analytic questions. For example, an analyst may be interested in whether an effect generalizes across various types of measures of a dependent variable. In such a case, the analyst could examine a model to determine if the effect sizes differed according to the type of measure used. If the synthesis forgoes this analysis to uphold the assumption that effect sizes are independent, potentially valuable information about a moderator would be lost. Therefore, one defensible strategy is to conduct a two-stage meta-analysis that shifts its units of analysis (Cooper, 1998). In the first stage, the meta-analysis would address the study-level effect sizes, which are calculated to represent the information from each study only once. A second stage would divide study outcomes into the various groupings specified by moderators and would permit information for a group of study participants to appear more than once, in order to examine the differences across the moderator (for examples of this strategy, see Gerrard, Gibbons, & Bushman, 1996; Kite & Johnson, 1988; Kolodziej & Johnson, 1996). This ordering of the stages assumes that analysts are usually interested in learning the overall, more general pattern in the literature first, prior to answering specific questions about moderators. This combination of approaches should help to allay concerns about nonindependence while still yielding the desired information. As yet another alternative, analysts might consider using multivariate procedures for the analysis of multiple effect sizes from each study (see Gleser & Olkin, 1994; Klaian & Raudenbush, 1996; for a general treatment of dependent data, see Kashy & Kenny, this volume, Ch. 17).

INTERPRETATIONS OF MEAN EFFECT SIZE INDEXES. Once a meta-analysis has derived a weighted mean effect size, it must be interpreted. 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 its 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. Cohen intended "that medium represent an effect of a size likely to be visible to the naked eye of a careful observer" (Cohen, 1992, p. 156). He intended 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 19.3 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. Although Cohen's guidelines for

TABLE 19.3. Cohen's (1969) Guidelines for Magnitude of *d* and *r*.

	Effect size metric		
Size	*d*	*r*	r^2
Small	0.20	.10	.01
Medium	0.50	.30	.09
Large	0.80	.50	.25

magnitude of effects are frequently cited, other ways of interpreting the magnitude of effects may prove more useful.

One 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 (see Table 19.3). Thus, a mean *d* of 0.50 produces an R^2 of .09. However, this value must be interpreted carefully because R^2, or variance explained, is a directionless effect size. Therefore, 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, as Hyde (1981) did in a meta-analysis of sex-related differences in cognitive performance.

A number of methodologists have discussed the magnitude of effects and argued that even quantitatively small effects can be quite consequential (e.g., Abelson, 1985; Prentice & Miller, 1992; Rosenthal, 1990, 1994; Ross & Nisbett, 1991). Especially useful in understanding the meaning of small effects is Rosenthal and Rubin's (1982) binomial effect size display (BESD), which is defined as a difference in "success" rates between treatment and control groups. In other words, this index consists of the difference between the proportion of cases (or research participants) who are successful (e.g., pass a test) in the treatment group and the proportion who are successful in the control group.

The BESD can be obtained from *r*, the correlation between the independent and dependent variables, by computing the success rate in the treatment condition as .50 plus *r*/2 and in the control condition as .50 minus *r*/2. For example, an *r* of .20 yields a success rate in the treatment group of .50 + .20/2 = .60 and a success rate in the control group of .50 − .20/2 = .40. Therefore, if the treatment and control groups each contained 100 individuals, 20 more people would survive in the treatment condition than in the control condition. Researchers who think about associations between variables in terms of the BESD are less likely to trivialize or dismiss small effects by saying, for example, that a correlation of .20 is "small" because it accounts for "only" 4% of the variance in the dependent variable. The BESD shows that, nonetheless, participants' "success rate" would be 20% higher in the treatment group.

The BESD is most easily implemented when researchers use status on a dichotomous independent variable (experimental vs. control group) to predict a dichotomous outcome, such as surviving versus not surviving or helping another person versus not helping. Nonetheless, this calculation can also be performed when the dependent variable is expressed in a continuous metric. In such circumstances, the researcher must dichotomize this variable at the median and thereby categorize participants as "below average" or "above average" on the dependent variable in question. For example, for a sex-related height difference of $d = 2.00$, the point-biserial *r* between sex and height would be .71. The probability of a man being above the "human" average (or "tall") would be roughly .85, and the probability of a woman being tall would be .14. The BESD would therefore be .61.

Offering another index for interpreting effect magnitude, McGraw and Wong (1992) defined their common language effect size statistic index (CL) as the probability that a score randomly sampled from one distribution will be larger than a score randomly sampled from another distribution. As McGraw and Wong explained, estimation of this index requires computation of the mean and standard deviation of the distribution of difference scores created by randomly comparing cases from the two distributions (e.g., male and female; treatment and control). CL is then the probability of obtaining a difference score greater than 0 in this distribution. This probability can be determined by converting the metric of raw difference scores to one of *z*- scores and consulting the normal curve (see Dunlap, 1994). For example, for a sex-related difference in height of $d = 2.00$, the CL, which is the probability that the man would be taller in any random pairing of a man and woman, is .92. As another example, the sex-related difference favoring males on the American College Test of math achievement, which can be expressed as a *d* of 0.48, translates into a CL of .63. That is, 63% of the time, a randomly sampled male will have higher achievement than a randomly sampled female.

In addition to the common language effect size statistic and the binomial effect size display, other helpful indexes for interpreting effect magnitude have been proposed (e.g., Cohen, 1988). These indexes include the *counternull* statistic, which is a nonnull effect size that is as probable as the null value, to which the effect size is ordinarily compared to establish its significance (see Rosenthal & Rubin, 1994).

To answer potential or actual critics' assertions that unpublished or unretrieved studies not present in the meta-analytic sample would invalidate or reverse the review's conclusions, analysts sometimes calculate the number of studies averaging a null effect that would be necessary to bring an overall meta-analytic mean to the point of nonsignificance (Rosenthal, 1979). If this fail safe $N(N_{fs})$ is large, then the meta-analytic result gains credibility; if it is small, then the result seems less trustworthy. For example, Rosenthal and Rubin (1978) calculated that it would take 65,123 studies averaging a null result in order to invalidate their observed experimenter expectancy effect. To calculate N_{fs},

$$N_{fs} = \frac{\left(\sum_{j=1}^{k} z_j\right)^2}{z_\alpha^2}, \tag{32}$$

where k is the number of studies, z_j is the unit normal value corresponding to a one-tailed test of significance, and z_α is the critical value (i.e., 1.645 for a one-tailed hypothesis). Orwin (1983) offered a variant of this equation that estimates N_{fs} directly from the mean weighted effect size. Although N_{fs} may have heuristic value in some instances, there are cautions worth noting. The equation for N_{fs} assumes that unretrieved studies would average null when in fact they may have the same pattern as the retrieved studies or even a reversed pattern (Becker, 1994; Begg, 1994). Also, it is difficult to evaluate the magnitude of N_{fs}, because it has no distribution theory (Rosenthal, 1979, 1984).

Another method of interpreting the magnitude of effect sizes is to compare them with effect sizes in similar domains in which magnitude is already known. For example, Eagly (1995) argued that claims that sex-related differences in behavior are necessarily small should be evaluated in relation to the magnitude of other known effects in psychology. Following this strategy, Bettencourt and Miller (1996) compared the magnitude of sex-related differences in aggression to the magnitude of the effect of provocation on aggression; this comparison was especially relevant because the mean effect sizes were derived from the same sample of studies. More generally, meta-analysts ought to compare the magnitude of a newly derived meta-analytic effect size to the magnitude of known effects in related research areas. For example, Lipsey and Wilson (1993) gathered 302 meta-analyses of psychological, educational, and behavioral interventions and determined the typical effect size obtained in such interventions. Similarly, Johnson, Carey, and Muellerleile (1997) gathered meta-analytic tests of the impact of behavioral interventions for behaviors relevant to various public health problems and used fixed and random effects meta-analytic model tests to compare the mean effect sizes obtained in these different literatures. This "meta-meta-analysis" inferentially compared the magnitude of effects across various domains of behavioral interventions. Alliger (1995) described and compared several small-sample techniques for performing these sorts of comparisons.

Many aspects of studies' methods can constrain effect magnitude, as noted above (see *Correcting effect sizes for bias*). Effects are larger or smaller depending on factors such as reliability of measures, heterogeneity of the participant population, and so on. Some of these factors lend themselves to bias corrections, and the magnitude of the effect size that represents a study depends on whether corrections have been applied for such problems. In addition, characteristics of the situation in which experiments are carried out can increase or reduce the impact that experimental manipulations and individual-difference variables have on dependent variables (Prentice & Miller, 1992). Analysts should code the studies in their databases for the presence of a wide range of such factors, in order to account for effect size variance that is produced by studies' nonequivalence on such factors.

CONDUCTING AND EVALUATING QUANTITATIVE SYNTHESES

Our treatment of quantitative synthesis has stressed the importance of high standards in conducting and evaluating these reviews. The hallmarks of a high-quality meta-analysis include success in locating studies, explicitness of criteria for selecting studies, thoroughness and accuracy in coding moderators variables and other study characteristics, accuracy in effect size computations, and adherence to the assumptions of meta-analytic statistics. When research syntheses meet 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, pre-quantitative manner" (p. 131). However, even a quantitative review that meets high standards does not necessarily constitute an important scientific contribution.

One factor affecting the 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 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. Similarly, 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.

Some reports of research syntheses may fail to tell a good story because they are overly complex. This complexity may arise from the fact that quantitative synthesis forces the reviewer to study the minute details of the studies' methods and findings. Although this close scrutiny can yield valuable insights, it may also foster a review that reflects too many complexities and thereby obscures its major findings. Sharing our concern about excessive complexity, Rosenthal (1995) stated, "I have never seen a meta-analysis that was 'too simple,' but I have often seen meta-analyses that were very fancy and very much in error" (p. 183). In short, even if a synthesis happens to solve a time-honored problem, it will have a poor reception if its message is mired in a forest of distracting minutiae.

Although in practice most critiques of meta-analyses take a narrative form by discussing the methods and findings of a published synthesis, the most informative critiques take a quantitative approach by empirically evaluating the findings and conclusions. A critique that may seem reasonable based on sheer logic may become overwhelming when supported by the data. For example, if a critic reasons that the selection criteria of a meta-analysis are faulty, showing that the presumably superior criteria yield different results makes the argument much more compelling. Similarly, if a critic claims that a particular moderator in a meta-analysis was confounded with another variable that is genuinely causal, he or she should demonstrate that the confound is consequential to the results obtained (e.g., by conducting model tests showing how the findings change when the confound is controlled). In this manner, scientific disputes can be arbitrated by empirical tests. In primary research, the most influential critiques take the form of replications with variations – often showing how an effect disappears once a confound is controlled. Similarly, criticism of quantitative syntheses proceeds most effectively in an empirical fashion. In our view, replications of meta-analytic reviews should become more frequent, so that faults that may be present in one review are evaluated or improved in later reviews.

With quantitative syntheses becoming commonplace, investigators should redouble their efforts to report the method and results of their primary-level studies as accurately and completely as possible. In particular, for experimental studies, a table of means and standard deviations for each primary dependent variable, reported for all cells of the design, should be conventional. It is very helpful if exact statistics are provided even for auxiliary effects that may be nonsignificant (e.g., the comparison of female and male participants). For correlational studies, a complete matrix of the variables' intercorrelations should be conventional.

ADDITIONAL RESOURCES ON RESEARCH SYNTHESIS

Essential reference works for quantitative synthesis are *The Handbook of Research Synthesis*, edited by Cooper and Hedges (1994b), as well as texts by Hedges and Olkin (1985), Cooper (1998), and Rosenthal (1991). Glass et al.'s (1981) book remains a good source on derivations of effect sizes. Hunt (1997) provides a compelling and highly readable history of the subject of research synthesis. Other works may be particularly valuable for other aspects of meta-analysis. In particular, Hunter and Schmidt's (1990) book provides an extensive treatment of effect size corrections. Mullen (1989) provided a well-rounded treatment of meta-analysis and includes software for many of the analyses he suggests. Schwarzer's (1989) software is in the public domain and contains many useful functions. Johnson's (1993) software is reasonably comprehensive and also includes much practical meta-analytic information (see Normand's, 1995, review of meta-analytic software).

THE FUTURE OF QUANTITATIVE SYNTHESES IN SOCIAL PSYCHOLOGY

The growing numbers of studies on social psychology's central phenomena dictate that, in the future, greater importance will be accorded to high-quality meta-analyses of these knowledge bases. Consumers of research, such as textbook authors, often express enthusiasm about meta-analytic contributions. The information-reduction benefits of quantitative reviews led one author of a social psychology textbook to write,

"I am not so much a critic or connoisseur of meta-analysis as an enthusiastic consumer" (D. G. Myers, 1991, p. 265). Yet, because meta-analysis is relatively new among scholars who practice it, the quality of published syntheses has been quite variable, and some have not been as informative as they might have been. In particular, some meta-analysts have not adequately searched for relevant studies and may have no greater claim to comprehensiveness than typical narrative reviewers. Also, meta-analyses that are confined to aggregating findings over studies fail to examine findings' homogeneity or to account for discrepancies between them. Such reviews inform their readers about the average direction and magnitude of an effect but not about its patterning. However, 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 N. 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. One misperception that psychologists 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, we assert that 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, to help plan additional, primary-level research (see Eagly & Wood, 1994). 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.

The advent of computerized and readily accessible databases of psychological research literatures (e.g., PsycINFO) has meant that less time and financial resources are necessary to conduct meta-analyses. Whereas past reviewers had to spend endless hours examining entries in print volumes of *Psychological Abstracts*, modern meta-analysts are able to quickly generate lists of studies that may be suitable for their reviews. Despite these gains, psychologists face severe limitations in obtaining access to the data underlying completed research. In contrast to some other scientific fields (e.g., sociology, political science), few raw data from primary research are archived in psychology, and this omission greatly limits the opportunity for reviewers to perform the secondary analyses that in some cases are necessary for calculating effect sizes for effects that have not been adequately reported. Primary researchers are often unable or unwilling to provide needed statistical information when they are contacted directly. Routine data archiving in a central location would remedy this unfortunate situation (see Cooper et al., 1997).

As time passes, psychologists and other 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, we believe that, on the contrary, meta-analysis will become a routine part of graduate training in many fields. With computer programs to aid calculations, most researchers should be able to integrate findings across studies as a normal and routine part of their research activities. Therefore, in the future, a substantial proportion of investigators in many fields will ply the art and science of research synthesis.

REFERENCES

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

Aguinis, H., & Pierce, C. A. (1998). Testing moderator variable hypotheses meta-analytically. *Journal of Management, 24,* 577–592.

Alliger, G. M. (1995). The small sample performance of four tests of the difference between pairs of meta-analytically derived effect sizes. *Journal of Management, 21,* 789–799.

Becker, B. J. (1987). Applying tests of combined significance in meta-analysis. *Psychological Bulletin, 102,* 164–171.

Becker, B. J. (1994). Combining significance levels. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 215–230). New York: Russell Sage.

Begg, C. B. (1994). Publication bias. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 399–409). New York: Russell Sage.

Bettencourt, B. A., & Miller, N. (1996). Gender differences in aggression as a function of provocation: A meta-analysis. *Psychological Bulletin, 119,* 422–447.

Birge, R. T. (1932). The calculation of errors by the method of least squares. *Physical Review, 40,* 207–227.

Bond, C. F., Jr., & Titus, L. J. (1983). Social facilitation: A meta-analysis of 241 studies. *Psychological Bulletin, 94,* 265–292.

Bond, R., & Smith, P. B. (1996). Culture and conformity: A meta-analysis of studies using Asch's (1952b, 1956) line judgment task. *Psychological Bulletin, 119,* 111–137.

Bornstein, R. F. (1989). Exposure and affect: Overview and meta-analysis of research, 1968–1987. *Psychological Bulletin, 106,* 265–289.

Brubaker, T. H., & Powers, E. A. (1976). The stereotype of "old": A review and alternative approach. *Journal of Gerontology, 31,* 441–447.

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

Bushman, B. J., & Wang, M. C. (1996). A procedure for combining sample standardized mean differences and vote counts to estimate the population standardized mean difference in fixed effects models. *Psychological Methods, 1,* 66–80.

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

Carroll, J. B. (1961). The nature of the data, or how to choose a correlation coefficient. *Psychometrica, 26,* 347–372.

Cochran, W. G. (1937). Problems arising in the analysis of a series of experiments. *Journal of the Royal Statistical Society* (Suppl.), *4,* 102–118.

Cohen, J. (1960). Coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

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: Erlbaum.

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

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

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

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

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

Cooper, H., DeNeve, K., & Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods, 2,* 447–452.

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

Cooper, H., & Hedges, L. V. (1994b). 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.

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

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper Collins.

Dickerson, K. (1997). How important is publication bias? A synthesis of available data. *AIDS Education and Prevention, 9* (Suppl. A), 15–21.

Driskell, J. E., & Mullen, B. (1990). Status, expectations, and behavior: A meta-analytic review and test of the theory. *Personality and Social Psychology Bulletin, 16,* 541–553.

Dunlap, W. P. (1994). Generalizing the common language effect size indicator to bivariate normal correlations. *Psychological Bulletin, 116,* 509–511.

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.

Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation.* Hillsdale, NJ: Erlbaum.

Eagly, A. H. (1995). The science and politics of comparing women and men. *American Psychologist, 50,* 145–158.

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good, but . . .: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin, 110,* 109–128.

Eagly, A. H., & Crowley, M. (1986). Gender and helping behavior: A meta-analytic review of the social psychological literature. *Psychological Bulletin, 100,* 283–308.

Eagly, A. H., & Johnson, B. T. (1990). Gender and leadership style: A meta-analysis. *Psychological Bulletin, 108,* 233–256.

Eagly, A. H., Karau, S., & Makhijani, M. G. (1995). Gender and the effectiveness of leaders: A meta-analysis. *Psychological Bulletin, 117,* 125–145.

Eagly, A. H., Makhijani, M. G., & Klonsky, B. G. (1992). Gender and the evaluation of leaders: A meta-analysis. *Psychological Bulletin, 111,* 3–22.

Eagly, A. H., & Steffen, V. J. (1986). Gender and aggressive behavior: A meta-analytic review of the social psychological literature. *Psychological Bulletin, 100,* 309–330.

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.

Ernst, C., & Angst, J. (1983). *Birth order: Its influence on personality.* New York: Springer-Verlag.

Feingold, A. (1995). The additive effects of differences in central tendency and variability are important in comparisons between groups. *American Psychologist, 50,* 5–13.

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

Fisher, R. A. (1932). *Statistical methods for research workers* (4th ed.). London: Oliver and Boyd.

Gerrard, M., Gibbons, F. X., & Bushman, B. J. (1996). Relation between perceived vulnerability to HIV and precautionary sexual behavior. *Psychological Bulletin, 119,* 390–409.

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

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

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.

Gordon, R. A. (1996). The impact of ingratiation on judgments and evaluations: A meta-analytic investigation. *Journal of Personality and Social Psychology, 71,* 54–70.

Green, S. K. (1981). Attitudes and perceptions about the elderly: Current and future perspectives. *International Journal of Aging and Human Development, 13,* 99–119.

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

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

Greenwald, S., & Russell, R. I. (1991). Assessing rationales for inclusiveness in meta-analytic samples. *Psychotherapy Research, 1,* 17–24.

Hall, J. A. (1978). Gender effects in decoding nonverbal cues. *Psychological Bulletin, 85,* 845–857.

Hardy, M. A. (1993). *Regression with dummy variables.* Newbury Park, CA: Sage.

Hays, W. L. (1988). *Statistics* (4th ed.). Fort Worth, TX: Harcourt Brace Jovanovich.

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. (1982a). Fitting categorical models to effect sizes from a series of experiments. *Journal of Educational Statistics, 7,* 119–137.

Hedges, L. V. (1982b). Fitting continuous models to effect size data. *Journal of Educational Statistics, 7,* 245–270.

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. (1990). Directions for future methodology. In K. W. Wachter & M. L. Straf (Eds.), *The future of meta-analysis* (pp. 11–26). New York: Russell Sage.

Hedges, L. V. (1994). Statistical considerations. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 29–38). New York: Russell Sage.

Hedges, L. V., & Becker, B. J. (1986). Statistical methods in the meta-analysis of research on gender differences. In J. S. Hyde & M. C. Linn (Eds.), *The psychology of gender. Advances through meta-analysis* (pp. 14–50). Baltimore, MD: Johns Hopkins University Press.

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., & Friedman, L. (1993). Computing gender differences in the tails of distributions: The consequences of differences in tail size, effect size, and variance ratio. *Review of Educational Research, 63,* 110–112.

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

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

Heinsman, D. T., & Shadish, W. R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate answers from randomized experiments? *Psychological Methods, 1,* 154–169.

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

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.

Hunter, J. E., & Schmidt, F. L. (1997). *Type I error in the fixed effects formulas for meta-analysis.* Manuscript submitted for publication.

Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin, 31,* 215–232.

Hyde, J. S. (1981). How large are cognitive gender differences? A meta-analysis using ω^2 and *d*. *American Psychologist, 36,* 892–901.

Johnson, B. T. (1993). *DSTAT 1.10: Software for the meta-analytic review of research literatures.* Hillsdale, NJ: Erlbaum.

Johnson, B. T., Carey, M. P., & Muellerleile, P. A. (1997). Large trials vs meta-analysis of smaller trials [Letter to the editor]. *Journal of the American Medical Association, 277,* 377.

Johnson, B. T., & Eagly, A. H. (1989). Effects of involvement on persuasion: A meta-analysis. *Psychological Bulletin, 104,* 290–314.

Johnson, B. T., Mullen, B., & Salas, E. (1995). Comparison of three major meta-analytic approaches. *Journal of Applied Psychology, 80,* 94–106.

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.

Kalaian, H. A., & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods, 1,* 227–235.

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology, 65,* 681–706.

Kite, M. E., & Johnson, B. T. (1988). Attitudes toward the elderly: A meta-analysis. *Psychology and Aging, 3,* 233–244.

Kite, M. E., & Whitley, B. E. (1996). Sex differences in attitudes toward homosexual persons, behaviors, and civil rights: A meta-analysis. *Personality and Social Psychology Bulletin, 22,* 336–353.

Knight, G. P., Fabes, R. A., & Higgins, D. A. (1996). Concerns about drawing causal inferences from meta-analyses: An example in the study of gender differences in aggression. *Psychological Bulletin, 119,* 410–421.

Kolodziej, M. E., & Johnson, B. T. (1996). Effects of interpersonal contact on acceptance of individuals diagnosed with mentally illness: A research synthesis. *Journal of Consulting and Clinical Psychology, 64,* 1387–1396.

Langenbucher, J., Labouvie, E., & Morgenstern, J. (1996). Measuring diagnostic agreement. *Journal of Consulting and Clinical Psychology, 64,* 1285–1289.

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

Light, R. J., Singer, J. D., & Willett, J. B. (1994). The visual presentation and interpretation of meta-analyses. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 439–454). New York: Russell Sage.

Lipsey, M. W. (1994). Identifying potentially interesting variables and analysis opportunities. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 111–138). New York: Russell Sage.

Lipsey, M. W., & Wilson, D. B. (1993) The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48,* 1181–1209.

Lutsky, N. (1981). Attitudes toward old age and elderly persons. In C. Eisdorfer (Ed.), *Annual Review of Gerontology and Geriatrics* (Vol. 1, pp. 287–336). New York: Springer.

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

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111,* 361–365.

Miller, A. G. (1972). Role playing: An alternative to deception? *American Psychologist, 27,* 623–636.

Miller, N., & Carlson, M. (1990). Valid theory-testing meta-analyses further question the negative state relief model of helping. *Psychological Bulletin, 107,* 215–225.

Miller, N., & Cooper, H. M. (Eds.). (1991). Special issue: Meta-analysis in personality and social psychology. *Personality and Social Psychology Bulletin, 17,* 243–349.

Miller, N., Lee, J. Y., & Carlson, M. (1991). The validity of inferential judgment when used in theory-testing meta-analyses. *Personality and Social Psychology Bulletin, 17,* 335–343.

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 Press.

Morris, S. B., & DeShon, R. P. (1997). Correcting effect sizes computed from factorial analysis of variance for use in meta-analysis. *Psychological Methods, 2,* 192–199.

Morrison, D. F. (1976). *Multivariate statistical methods.* New York: McGraw-Hill.

Mullen, B. (1989). *Advanced BASIC meta-analysis: Procedures and programs.* Hillsdale, NJ: Erlbaum.

Mullen, B., & Copper, C. (1994). The relation between group cohesiveness and performance: An integration. *Psychological Bulletin, 115,* 210–227.

Mullen, B., & Felleman, V. (1989). Tripling in the dorms: A meta-analytic integration. *Basic and Applied Social Psychology, 11,* 33–43.

Myers, D. G. (1991). Union is strength: A consumer's view of meta-analysis. *Personality and Social Psychology Bulletin, 17,* 265–266.

Myers, J. L., & Well, A. D. (1991). *Research design and statistical analysis.* New York: Harper Collins.

Normand, S. (1995) Meta-analysis software: A comparative review. *American Statistician, 49,* 298–309.

Nouri, H., & Greenberg, R. H. (1995). Meta-analytic procedures for estimation of effect sizes in experiments using complex analysis of variance. *Journal of Management, 21,* 801–812.

Oliver, M. B., & Hyde, J. S. (1993). Gender differences in sexuality: A meta-analysis. *Psychological Bulletin, 114,* 29–51.

Orwin, R. G. (1983). A fail-safe *N* for effect size in meta-analysis. *Journal of Educational Statistics, 8,* 157–159.

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

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods, 3,* 354–379.

Pearson, K. (1933). On a method of determining whether a sample size *n* supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika, 25,* 379–410.

Prentice, D., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112,* 160–164.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–321). New York: Russell Sage.

Ray, J. W., & Shadish, W. R. (1996). How interchangeable are different estimators of effect size? *Journal of Consulting and Clinical Psychology, 64,* 1316–1325.

Rhodes, N., & Wood, W. (1992). Self-esteem and intelligence affect influenceability: The mediating role of message reception. *Psychological Bulletin, 111,* 156–171.

Rosenthal, R. (1968). Experimenter expectancy and the reassuring nature of the null hypothesis decision procedure. *Psychological Bulletin, 70* (6, Pt. 2), 30–47.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638–641.

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

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist, 45*, 775–777.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (rev. ed.). Beverly Hills, 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.

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

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*. New York: Cambridge University Press.

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

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

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

Rosenthal, R., & Rubin, D. B. (1994). The counternull value of an effect size: A new statistic. *Psychological Science, 5*, 329–334.

Ross, L., & Nisbett, R. (1991). *The person and the situation: Perspectives of social psychology*. New York: McGraw-Hill.

Rotton, J., Foos, P. W., Van Meek, L., & Levitt, M. (1995). Publication practices and the file drawer problem: A survey of published authors. *Journal of Social Behavior and Personality, 10*, 1–13.

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.

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. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*, 115–129.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199–223.

Schooler, C. (1972). Birth order effects: Not here, not now! *Psychological Bulletin, 78*, 161–175.

Schwarzer, R. (1989). *Meta-analysis programs* [Computer software]. Institut für Psychologie (WE 7), Frei Universität Berlin, Berlin, Germany: Author.

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., Doherty, M., & Montgomery, L. M. (1989). How many studies are in the file drawer? An estimate from the family/martial psychotherapy literature. *Clinical Psychology Review, 9*, 589–603.

Sheppard, B. H., Hartwick, J., & Warshaw, P. R. (1988). The theory of reasoned action: A meta-analysis of past research with recommendations for modifications and future research. *Journal of Consumer Research, 15*, 325–343.

Snedecor, G. W., & Cochran, W. G. (1980). *Statistical methods* (7th ed.). Ames: Iowa State University Press.

Sommer, B. (1987). The file drawer effect and publication rates in menstrual cycle research. *Psychology of Women Quarterly, 11*, 233–242.

Stangor, C., & McMillan, D. (1992). Memory for expectancy-congruent and expectancy-incongruent information: A review of the social and social developmental literature. *Psychological Bulletin, 111*, 42–61.

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.

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

Sulloway, F. H. (1996). *Born to rebel: Birth order, family dynamics, and creative lives*. New York: Pantheon Books.

Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: A meta-analysis. *Psychological Bulletin, 121*, 371–394.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: Harper Collins.

Thacker, S. B. (1988). Meta-analysis: A quantitative approach to research integration. *Journal of the American Medical Association, 259*, 1685–1689.

Timm, N. H. (1975). *Multivariate analysis, with applications in education and psychology*. Belmont, CA: Brooks-Cole.

Toothaker, L. E. (1991). *Multiple comparisons for researchers*. Newbury Park, CA: Sage.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

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

Wanous, J. P., Sullivan, S. E., & Malinak, J. (1989). The role of judgment calls in meta-analysis. *Journal of Applied Psychology, 74*, 259–264.

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.

Whitley, B. E., Jr., & Kite, M. E. (1995). Gender differences in attitudes toward homosexuality: A comment on Oliver and Hyde. *Psychological Bulletin, 117*, 146–154.

Wicker, A. W. (1969). Attitude versus actions: The relationship of verbal and overt behavioral responses to attitude objects. *Journal of Social Issues, 25*(4), 41–78.

Winer, B. J., Brown, D. R., & Michels, K. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.

Wood, W. (1987). Meta-analytic review of sex differences in group performance. *Psychological Bulletin, 102,* 53–71.

Wood, W., Lundgren, S., Ouellette, J. A., Busceme, S., & Blackstone, T. (1994). Minority influence: A meta-analytic review of social influence processes. *Psychological Bulletin, 115,* 323–345.

Wood, W., Rhodes, N., & Whelan, M. (1989). Sex differences in positive well-being: A consideration of emotional style and marital status. *Psychological Bulletin, 106,* 249–264.

Wortman, P. M. (1994). Judging research quality. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 97–110). New York: Russell Sage.

Yates, F., & Cochran, W. G. (1938). The analysis of groups of experiments. *Journal of Agricultural Science, 28,* 556–580.