Guidelines for Meta-Analyses of Counseling Psychology Research
Stephen M. Quintana and Takuya Minami
The Counseling Psychologist 2006; 34; 839
DOI: 10.1177/0011000006286991

The online version of this article can be found at:
http://tcp.sagepub.com/cgi/content/abstract/34/6/839

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Division of Counseling Psychology of the American Psychological Association

Additional services and information for The Counseling Psychologist can be found at:

Email Alerts: http://tcp.sagepub.com/cgi/alerts

Subscriptions: http://tcp.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

Citations http://tcp.sagepub.com/cgi/content/refs/34/6/839
Guidelines for Meta-Analyses of Counseling Psychology Research

Stephen M. Quintana
University of Wisconsin–Madison

Takuya Minami
University of Utah

This article conceptually describes the steps in conducting quantitative meta-analyses of counseling psychology research with minimal reliance on statistical formulas. The authors identify sources that describe necessary statistical formula for various meta-analytic calculations and describe recent developments in meta-analytic techniques. The authors illustrate meta-analytic procedures with recently published meta-analyses in counseling psychology, and last, they provide a guide for writing up meta-analytic studies for publication.

The basis of scientific knowledge is the accumulation of research. To make sense of a body of research, some form of synthesis is needed. It would be unthinkable to submit a manuscript, thesis, dissertation, or grant proposal without some synthesis of extant research. Moreover, it would seem equally naïve to reach some conclusions about a research area without a review. At present, there are two main options available to review a research literature: traditional narrative reviews and quantitative meta-analyses. In traditional narrative reviews, researchers review extant research and summarize the findings in a verbal format; in contrast, quantitative meta-analyses summarize review findings using statistical procedures. Hedges and Pigott’s (2001) definition follows:

Meta-analysis involves describing the results of each study using a numerical index . . . and then combining these estimates across studies to obtain a summary. (p. 203)

Researchers developed meta-analytic techniques to redress problems with narrative reviews by providing systematic procedures for reviewing
available research (Smith & Glass, 1977). In short, meta-analyses involve the application of statistical procedures to literature reviews, replacing somewhat subjective decisions about research trends, such as magnitude and consistency of research trends, with statistically informed decisions. This is not to suggest, however, that all subjectivity has been removed from meta-analyses; rather, meta-analyses allow for some judgments about research literatures to be based on statistical results. Hence, meta-analyses share some of basic features of narrative reviews but replace the somewhat informal process of narrative reviews with formal statistical formulation.

A further understanding of meta-analyses can be gained by illustrating how meta-analytic procedures systematize some of the procedures typically used in narrative reviews. Specifically, both narrative reviews and meta-analyses are based on procedures for aggregating research findings, weighting studies, and selecting studies to include in the review, but meta-analyses involve more systematic processes for aggregating, weighting, and reviewing the research—that is, aggregating research findings in narrative reviews is typically based on comparing the number of studies demonstrating statistical significance with the number of studies revealing nonsignificant findings to support some conclusion about the research. This is considered vote counting (Cohen, 1994; Hedges & Olkin, 1980; Lipsey & Wilson, 1993; Schmidt & Hunter, 1996; Vacha-Haase & Thompson, 2004) in which the reviewer counts the number of studies demonstrating significance versus the number failing to demonstrate significance in the predicted direction. As is well known, drawing conclusions based solely on statistical significance is problematic because statistical significance is a reflection of sample size: The larger the sample size, the more likely the findings will be significant, and with a large enough sample size, even effects with trivial magnitude will be statistically significant despite little or no practical significance (Cohen, 1994). Consequently, the failure to detect statistical significance is partly a function of the studies’ statistical power (i.e., probability of making a Type II error; Fagley, 1985). Because many studies in counseling and other areas of psychology are underpowered, crude procedures such as counting the instances of statistical significance can be misleading.

In contrast, meta-analyses do not count votes of significance but instead focus on effect sizes and confidence intervals (CIs) for the effect size (for more on CIs, see Kahn, 2006 [TCP, special issue, part 1]; Weston & Gore, 2006 [TCP, special issue, part 1]; Worthington & Whittaker, 2006 [this issue]). Most commonly, an effect-size parameter is the magnitude of the true relationship between two variables in a population (please see Step 4 for different indices of effect sizes), which is estimated using the sample data (see Henson, 2006 [TCP, special issue, part 1]). CI is the range in which the true effect-size parameter is expected to be found with a speci-
fied level of confidence; a 95% CI is the range of values in which the true effect size is expected to fall 95% of the time. If the CI includes 0, then the effect size is not statistically significant from 0. This emphasis on effect sizes and on the CIs around them is consistent with the growing consensus that effect sizes are more meaningful for making conclusions about research findings. In addition, meta-analyses not only provide procedures for describing the overall effect for a specific set of studies but also allow researchers to estimate the overall effect size for the sample population. Researchers tend to be more interested in reaching conclusions attributable to populations of interest than in simply describing the characteristics of a set of studies.

In addition, both narrative reviews and meta-analyses weight individual studies in their reviews—they differ in that meta-analyses involve systematic procedures for determining the weights given to studies. Narrative reviews typically identify a few key studies and give them more weight in drawing conclusions about the body of research. However, the problem with this latter procedure is that the weightings are rarely made blind to the direction and strength (e.g., significance) of the findings—hence, the researchers’ investment in a particular conclusion about the body of research can taint decisions about weighting studies. Moreover, researchers are not, as once thought, neutral but are susceptible to inferential errors (Spengler, Strohmer, Dixon, & Shivy, 1995). For example, a researcher who believes that an authoritative counselor is clinically superior to those who are not may have a bias toward those studies supporting the researcher’s belief. These errors may influence judgments during narrative reviews, which may help explain why different reviewers of the same research literature may reach different conclusions (Lipsey & Wilson, 1993; Wampold, 2001). In contrast, well-conducted meta-analyses decide a priori on the weighting strategies. In meta-analyses, researchers base weighting decisions on relatively objective criteria (e.g., sample size of studies) and apply them independently of the direction and strength of the findings.

Some defenders of narrative reviews have pointed out some serious problems of meta-analyses but may fail to realize that quantitative as well as narrative reviews share these problems (Sharpe, 1997). For example, the so-called file-drawer problem for meta-analyses is serious (Lipsey & Wilson, 1993; Rosenthal, 1979; Sharpe, 1997) but is an equally, if not more, pernicious problem for narrative reviews. In short, the file-drawer problem occurs when the researcher basis the synthesis on a biased sample of studies, specifically biased in favor of those that reveal publishable results and against those languishing in researchers’ file drawers or computer hard drives, which were unable to reject null hypotheses. (Null hypotheses are statistical hypotheses [e.g., \( r = 0 \)] that are evaluated with sample data, and if the null hypothesis can
be rejected, the results are interpreted as supporting the investigator’s predictions.) The concern with the file-drawer problem is that the results of the meta-analysis may be invalid or misleading because there may be a large number of unpublished studies in file drawers that the researcher did not include in the analysis. It is important to note that the file-drawer problem also undermines the conclusions of narrative reviews, which typically share the same bias given the nearly impossible challenge of collecting research studies unbiased by the significance of the findings (Sharpe, 1997). In addition, although researchers have developed specific procedures to address the file-drawer problem in meta-analyses, there is no systematic redress of the problem in narrative reviews. For example, researchers conducting a meta-analysis could calculate how many studies are necessary to refute their conclusion (e.g., Rosenthal, 1979, described later in more detail), whereas such procedures are unavailable for narrative reviews.

More generally, Sharpe (1997) indicated that narrative analyses are not immune to common criticisms of meta-analyses. The increasing acceptance and growth of meta-analysis (Cheung & Chan, 2004), despite criticisms from some (for a review of the criticism, see Sharpe, 1997), reflect appreciation for the value of applying principles of research methodology—such as random sampling, careful selection procedures, and statistical evaluation—to reviews of extant research.

The two important points from this brief comparison of narrative and quantitative syntheses of research (i.e., meta-analyses) are as follows: (a) They share many fundamental processes and (b) they differ in that quantitative meta-analyses have systematized many of the subjective decisions made in narrative syntheses of research. This article’s purpose is to provide a conceptual guide through the steps involved in conceptualizing, conducting, and reporting meta-analyses. This article describes the steps involved in meta-analyses conceptually and relies minimally on mathematical formulas—instead, we provide references to the sources detailing the formulas for interested readers.

**STEP 1. CONCEPTUALIZING A META-ANALYTIC STUDY**

Traditionally, researchers applied meta-analyses to determine the effect sizes associated with different interventions or between control and intervention groups. The classical meta-analysis is Smith and Glass’s (1977) review of psychotherapy research in response to Eysenck’s (1952) doubts about psychotherapy’s efficacy. Meta-analyses have grown considerably in acceptance and application since Smith and Glass’s study. For example,
Lipsey and Wilson (1993) found 290 meta-analyses that had been conducted in the 15 years after Smith and Glass’s study. Lipsey and Wilson meta-analyzed those meta-analyses and found strong support for the effectiveness of interventions across three kinds of interventions (behavioral, educational, and psychological), which contrasted with the conclusions of small or no effects by narrative analyses reviewing the same bodies of research. Similarly, counseling psychologists have used meta-analyses to assess the effects of treatments and interventions including grief therapy (Allumbaugh & Hoyt, 1999), religion-accommodative counseling (McCullough, 1999), and career intervention (Whiston, Sexton, & Lasoff, 1998).

In addition, researchers often use meta-analyses to estimate the strength of relationships among variables, such as impact of ethnicity on counseling (Coleman, Wampold, & Casali, 1995; Shin et al., 2005), between conceptual-level and counseling-related tasks (Holloway & Wampold, 1986), among antecedents and effects of perceived therapist credibility (Hoyt, 1996), and self-efficacy beliefs on academic outcomes (Multon, Brown, & Lent, 1991).

For reasons elaborated throughout this article, meta-analyses are also well suited to resolve broad theoretical and empirical questions and have been critical in forming fairly conclusive decisions about issues that had been previously obfuscated by a plethora of studies yielding mixed or inconsistent results. For example, in a few meta-analyses, Wampold and his colleagues (Ahn & Wampold, 2001; Baskin, Tierney, Minami, & Wampold, 2003; Wampold, 2001; Wampold, Minami, Baskin, & Tierney, 2002; Wampold et al., 1997) have seriously challenged the applicability of the medical model to psychotherapy research by demonstrating that different schools of psychotherapy do not differ in efficacy and that specific ingredients (i.e., therapeutic techniques), in and of themselves, do not contribute to outcome. No single study could have had the impact on the field as this kind of well-designed meta-analysis. Furthermore, researchers could use meta-analyses creatively to assess theories and measures, as done by Behrens (1997) on the White Racial Identity Attitude Inventory (Helms & Carter, 1990) and by Rounds and Tracey (1996) on the cross-cultural structural equivalence of the Realistic, Investigative, Artistic, Social, Entrepreneurial, Conventional model (Holland, 1985).

Meta-analyses can address the same questions as primary studies. One advantage of meta-analyses over primary studies is a significant increase in power (Cohn & Becker, 2003). As mentioned before, because many primary studies in counseling and other areas of research have low levels of statistical power and may fail to reject the null hypothesis (i.e., committing Type II errors), it may be difficult to detect a prevailing trend when considering the studies individually. In contrast, meta-analyses aggregate findings
across studies and can compensate for the low statistical power within individual studies. For example, it is possible to detect significance in a meta-analysis of a set of primary studies with small to moderate sample sizes, although each of the primary studies may have failed to reveal statistical significance (Lipsey & Wilson, 1993). Hence, meta-analyses investigate with considerably more power the same questions as primary studies. Moreover, because of the increase in power, a null finding (i.e., a statistically nonsignificant finding, which would be a failure to reject the null hypothesis) in a meta-analysis may reveal important scientific information, whereas a null finding in primary studies is difficult to interpret (Cohn & Becker, 2003; Fagley, 1985). For example, a meta-analysis comparing efficacy of religion-accommodative counseling and standard approaches found no differences between the two approaches. This study has more implications than any of the five studies individually, particularly given that the individual studies averaged only 22 participants per study (McCullough, 1999). Interpreted individually, it is inconclusive if the lack of differences between religion-accommodative counseling and standard approaches is because of lack of power in the study or to lack of true differences in the population. However, in a sufficiently powered meta-analysis, lack of power is unlikely to be the reason for not finding a meaningful difference.

There are additional advantages of meta-analyses, relative to primary studies, in external validity because meta-analyses can generalize effects across the wide range of settings, participants, and interventions sampled in the original primary studies. Hall, Rosenthal, Tickle-Degnen, and Mosteller (1994) elaborate on the kinds of problems to which researchers can apply meta-analysis.

In addition to these considerable advantages over primary studies, meta-analyses also address questions that are either impossible or impractical to address within a primary study. For example, Eagly and Carli (1981) detected significant differences in findings about sex roles in research conducted by male compared with female authors. In another meta-analysis, Ahn and Wampold (2001) investigated the widely held assumption that there are therapeutic effects associated with the unique skills and techniques as specified in different kinds of therapy and counseling. In short, they found no support for the assumption of specific effects. Neither of these two theoretical issues would have been practical to investigate in a primary study. Moreover, the conclusions of these two meta-analyses were particularly compelling because they generalized across a range of studies. As another example, Newell (in press) found differences in effect sizes of school-based interventions for behavior disorders based on the proportion of minority youth in the study. That is, Newell found that the effect sizes of treatments were larger when the sample was predominately White as
opposed to predominately Black. A meta-analyst could also hypothesize that the intervention effects are increasing over time or that the strength of children’s prejudice is declining across the five decades since landmark civil rights legislation, neither of which would be a practical hypothesis for a primary study.

We illustrate the steps in a meta-analysis using a recently published study by Baskin et al. (2003). Baskin et al. meta-analyzed psychotherapy research to test the hypothesis that differences between treatment and placebo groups was due not to the superiority of the treatments over the placebo groups, as the primary studies hypothesized, but to structural differences (e.g., number of sessions, format of sessions) between the two conditions. Baskin et al. suspected and found that many clinical trials have failed to construct structurally equivalent placebo groups to the treatment groups and that these structural differences would be associated with whether there are significant effect-size differences between the two conditions: When the placebo groups were not equivalent, they would have lower effect sizes than the treatment groups; however, when the placebo and treatment groups were equivalent, there would be no difference in effect sizes. Clearly, the researchers could not effectively address this in a primary study. In sum, researchers can apply meta-analyses to most any question addressed by a number of primary studies, while providing significant increase in statistical power over primary studies. Furthermore, meta-analyses can address critical questions that are impractical to address with primary studies.

**Preliminary Power Analyses**

As with primary research, researchers should conduct meta-analyses only when there is sufficient statistical power to yield interpretable results. For example, a meta-analysis with low statistical power is likely to yield an effect-size estimate with a wide CI and, consequently, undermine the ability to draw firm conclusions from the results. Low statistical power will yield effect-size estimates with much error variance, creating a wide CI. A wide CI makes it difficult to reject null hypothesis. Because the power in meta-analysis often reflects the number of studies to meta-analyze, when meta-analysis does not have a sufficient number of studies (and thus a sufficient overall $N$) to aggregate and test its hypotheses, the advantages of conducting a meta-analysis are diminished. There are useful procedures and guidelines for conducting power analyses prior to conducting a complete meta-analysis (Donner, Piaggio, & Villar, 2003; Hedges & Pigott, 2001). As in power analyses for primary studies (Fagley, 1985), power in meta-analyses is based on the effect size relative to its variability. Because
both are rarely known a priori to a meta-analysis, they need to be estimated. For estimating the power of the meta-analysis using the methods Hedges and Pigott (2001) illustrated, several values need to be estimated: (a) the expected observed effect size between the conditions within the studies, (b) average sample sizes in each condition per study, and (c) number of studies to aggregate. In addition, power analyses can be valuable after completing a meta-analysis, particularly when results fail to reject the null hypotheses (for considerations for interpreting null findings, see Fagley, 1985). As mentioned earlier, failing to reject the null hypothesis may be scientifically important in a meta-analysis because of enhanced levels of power, whereas in primary studies, this is often difficult to interpret (Fagley, 1985). A power analysis conducted subsequent to failing to reject the null hypothesis in a meta-analysis could help rule out the possibility that the rejection of the null hypothesis was because of low power. Conversely, as in the case of primary studies, meta-analyses that fail to reject the null hypothesis but have low statistical power are difficult to interpret (Fagley, 1985).

Baskin et al. (2003) did not report a power calculation for their study, but we calculate it here for illustration. Using Hedges and Pigott (2001), first, the researchers must specify the magnitude of effect size that would be scientifically meaningful or would reflect practical significance. For Baskin et al., the effect size to be observed is the difference between the treatment and the placebo controls. Here, hypothetically, we specify our expected effect-size difference between the treatment and the placebo conditions as small (i.e., \( d = 0.2 \)), following Cohen’s (1988) suggestion. Second, for the purpose of power analysis, we hypothesize that the sample sizes in each condition within a study are equal to 25. Thus, each study will have 25 participants in both treatment and placebo control conditions, resulting in 50 total participants per study. Third, we assume that we will find approximately 20 studies that compare treatments to placebo control groups. Using Hedges and Pigott’s (2001) formulas, we found that the power to detect a small effect size \( (d = 0.2) \) with 20 studies that have 25 participants in each condition is .934 for a one-tailed hypothesis and .884 for a two-tailed hypothesis. Thus, under these parameters (i.e., \( n = 25 \) per study, \( d = 2 \), number of studies = 20), Baskin et al. had considerable power to warrant conducting the meta-analysis.

**STEP 2. DATA COLLECTION: IDENTIFYING AND SELECTING STUDIES FOR A META-ANALYSIS**

Data collection for meta-analyses often requires exhaustive efforts in using database searches (e.g., PsycINFO, Web of Science) and in reviewing
published articles, books, and other outlets. Meta-analytic researchers need to carefully consider the parameters (e.g., keywords, database selection) of their search to cast a wide net for studies and to minimize bias in selecting studies. Barber and Milrod (2004) recommended identifying studies by examining the reference sections of relevant articles and contacting experts in addition to using search engines. If search procedures identify more studies than can be feasibly examined with meta-analysis, it may be tempting to limit the search to a subset of journals. It is, however, generally preferable to select studies randomly across various sources than to arbitrarily limit the search to certain sources (Rosenthal, 1995). The idea of using randomly selected participants from a population in primary studies is applicable to meta-analyses in which primary studies are, ideally, randomly selected from the extant population of studies—in both situations, random selection increases external validity. It is particularly critical to search for unpublished studies (e.g., dissertations), which require considerably more effort to obtain than do published reports. White (1994) provided useful suggestions for maximizing the number of studies for a meta-analysis, including procedures to locate unpublished studies. The meaningfulness of meta-analyses is based on the ability to include unpublished studies because meta-analyses limited to published studies could inflate effect-size estimates (Conn, Valentine, Cooper, & Rantz, 2003; Lipsey & Wilson, 1993).

After searching for and identifying all possible studies, the meta-analytic researcher must formulate the criteria for including studies in the analysis. The inclusion criteria should involve methodological as well as content considerations. To minimize a concern raised about early large-scale meta-analyses (e.g., Eysenck, 1978), *Garbage In Garbage Out*, it may be important to include only the studies that meet some basic methodological standards—such as random assignment to treatment group, use of a control group, or appropriate instrumentation—to exclude studies that are unable to contribute meaningful information about the topic (see Landman & Dawes, 1982). What constitutes basic methodological requirements may vary across research areas. We recommend that the meta-analyst specify requirements a priori and base these decisions on a careful understanding of the research area. Another alternative to address the *Garbage In Garbage Out* criticism is to include methodologically weak studies and, thereby, investigate if method quality is associated with effect sizes (see section on moderator effects). Bowman, Scogin, Floyd, and Mckendree-Smith (2001) and Conn et al. (2003) provide good examples of including the quality of study methodology as a variable in meta-analyses.

There is research recommending excluding studies with little statistical power from the meta-analyses to minimize problems with the file-drawer problem mentioned earlier (Kraemer, Gardner, Brooks, & Yesavage, 1998;
Muncer, Craige, & Holmes, 2003). This recommendation appears counterintuitive as it suggests systematically excluding some studies (i.e., those with low power) to increase the results’ external validity. This recommendation appears to depart from accepted practices of increasing external validity by including as much data as possible and by not systematically excluding studies. Nonetheless, the logic underlying this recommendation is that of all the studies conducted with low statistical power, there is likely a strong publication bias associated with the ones lucky enough to demonstrate statistical significance (Kraemer et al., 1998). Moreover, adding primary studies with low statistical power to a meta-analysis can reduce the statistical power of the meta-analysis because effect sizes of these studies tend to be unreliable and can increase error associated with the meta-analysis, which reduces the power of the meta-analysis (Muncer et al., 2003). At present, we recommend that researchers not exclude studies based on statistical power. Instead, researchers can weight studies with small samples less than those with larger samples (Hedges & Olkin, 1985; Shadish & Haddock, 1994). In addition, journal editors and reviewers may evaluate negatively a meta-analysis that systematically excludes studies, although excluding studies has an ironic effect on the study’s external validity. In the future, excluding studies with low power may become a more common and accepted practice in meta-analyses.

Researchers should also limit the studies by basing their selection on conceptual or theoretical considerations related to the reason for conducting the study. Lipsey and Wilson (2001) describe considerations involved in forming inclusion and exclusion criteria for meta-analyses. In the example mentioned earlier, Baskin et al. (2003) conducted a thorough literature search to retrieve published clinical trials that met their inclusion and exclusion criteria. First, they manually searched the contents between 1994 and 2000 of the six journals that publish the majority of the psychotherapy outcome studies. Second, they searched using the PsycINFO database for the same period using terms such as placebo, placebo control, alternative treatment, alternative therapy, control facilitation, control group, supportive, supportive therapy, and nondirective. They included only those studies with placebo groups and information sufficient to classify them as equivalent or nonequivalent. In addition, they further assessed the retrieved studies against inclusion criteria such as participants being adults, participant randomization into conditions, experimental condition being bona fide psychotherapy or counseling, and participants receiving more than one session. These procedures identified 21 studies that met the inclusion criteria and reported sufficient information to conduct a meta-analysis.

To reiterate guidelines for collecting the data for meta-analyses, researchers need to give careful attention to how a sample of studies is identified to
ensure that it is representative of the body of research conducted on the topic. Search procedures should attempt to identify all relevant or applicable studies, giving particular attention to finding unpublished studies. Next, researchers need to develop specific inclusion and exclusion criteria related to methodological and theoretical considerations of the studies to select only those studies that contribute meaningfully to the research questions addressed in the meta-analysis.

**STEP 3. CODING STUDIES**

An important advantage of meta-analysis over narrative reviews is that important characteristics within primary studies tend to be coded using systematic and reliable criteria, and the coding is usually independent of the findings of primary studies in meta-analyses. One reason to code primary studies is to form homogenous subgroups. Interpreting homogenous sets of studies is more straightforward and meaningful than interpreting a heterogeneous set of studies. To illustrate, Whiston et al. (1998) coded studies based on what types of career intervention (e.g., individual, computer), rather than simply career intervention, so that the separately aggregated effect sizes could be referred to the specific intervention. A second reason to code studies is when some factors possibly moderate the effect sizes of the studies (see section on moderator effects). Thus, for example, the researcher could test whether multicultural competence is higher among students instructed under an infused curriculum. In addition, meta-analysts are often interested in the research design’s potential impact on the effect sizes. Researchers can code studies according to a variety of characteristics, ranging from the use of objective criteria for a single dimension (e.g., presence or absence of random assignment) to more subjective ratings of a generalized criterion (e.g., method quality based on design, adequacy of measurement strategy). However, as the strength of meta-analyses over narrative reviews lies in its increased objectivity, we strongly recommend that researchers define coding variables precisely (e.g., weeks in treatment vs. treatment length) and objectively (e.g., recruitment method vs. adequacy of recruitment). This decreases the likelihood of biased coding. For further discussions about coding studies in meta-analysis, we refer readers to Lipsey (1994), Orwin, (1994), Stock (1994), and Wortman (1994).

Nonetheless, which dimensions should be coded ultimately depends on the hypotheses of the meta-analysis and the nature of the research area (Rosenthal, 1995). For example, Baskin et al. (2003) coded studies based on whether the placebo groups were structurally equivalent to the active treatment. Specifically, the placebo group and the active treatment had to
be equivalent in six prespecified structural criteria, such as number and duration of sessions. If the placebo group was equivalent to the treatment in all six criteria, the authors considered the placebo equivalent; otherwise, they considered it nonequivalent. Among the 21 studies, the authors considered 13 to have equivalent structure between the treatment and the placebo and 8 to have nonequivalent structure. This is an example of coding based on a priori hypotheses and conceptualization of the study. As another example, Newell (in press) coded studies based on whether the sample was predominately African American in a meta-analysis investigating school-based interventions of behavior disorders. She hypothesized (and found) that treatment was less effective for African Americans, relative to samples with predominately White Americans. She also coded the studies according to level of cultural validity to test a hypothesis that studies with low levels of cultural validity would demonstrate smaller effect sizes relative to studies demonstrating greater levels of cultural validity. She based the coding of cultural validity on criteria described by Quintana, Troyano, and Taylor (2001).

The coding of primary studies needs to be based on important differences within the range of studies to control for differences across studies. These differences might not have been identified before the study began but surfaced as it became clear that the surveyed studies varied on important dimensions (e.g., random assignment vs. quasi-experimental design). For example, Newell (in press) discovered that nearly half the primary studies identified for her meta-analysis involved cognitive-behavioral interventions. Consequently, she classified the studies according to the interventions’ theoretical orientations. This kind of coding could be considered a priori because it occurred before there was an aggregation of effect sizes. It may become necessary, however, to code studies based on a dimension that was identified only after some preliminary data analyses. Specifically, statistical analyses might reveal that the collection of primary analyses is heterogeneous, which may signal the need to divide the collection of studies into smaller, homogeneous groups (see section on homogeneity).

Regardless of when the researcher codes primary studies, he or she should conduct the coding procedures blind to both the effect sizes of primary studies and the hypotheses of the meta-analysis. Researchers should use blind coding because failing to do so may unintentionally bias the coding toward the coders’ speculations (e.g., confirmatory inferential bias), which diminishes the advantage of performing a meta-analysis over a narrative review. Hence, researchers should calculate the primary studies’ effect sizes independently of the coding of all variables of interest in the study. Where possible, intercoder consistencies should be determined to indicate the reliability of the coding procedures, such as by calculating the interrater reliability (e.g., Cohen’s
kappa, intraclass correlation coefficient). Procedures for calculating interrater reliability in meta-analyses are identical to those in primary studies (see Helms, Henze, Sass, & Mifsud, 2006 [TCP, special issue, part 1]).

**STEP 4. STATISTICAL PROCEDURES IN META-ANALYSES**

**Converting Reported Statistical Information Into a Common Effect-Size Metric**

Researchers need to examine each study selected for a meta-analysis to extract its effect sizes. To aggregate across studies, researchers need to convert statistical information reported from the individual studies into a common effect-size metric. There are numerous metrics available, each with various rationales for its use (for further discussion, see Haase, Ellis, & Ladany, 1989; Strube, 1988; Vacha-Haase & Thompson, 2004). Metrics primarily used for continuous variables include $d$ (e.g., Hedges & Olkin, 1985), $r$ (e.g., Rosnow, Rosenthal, & Rubin, 2000), $\eta^2$ (e.g., Haase, Waechter, & Solomon, 1982; Olejnik & Algina, 2003), and $\omega^2$ (e.g., Olejnik & Algina, 2003; Paquin, 1983). A commonly used metric for dichotomous variables is the odds ratio (e.g., Fleiss, 1994; Sánchez-Meca, Marín-Martínez, & Chacón-Moscoso, 2003).

The two main families of effect sizes commonly used in meta-analyses are $r$ and $d$ (Henson, 2006). The $r$, which is the well-known Pearson product moment correlation, measures the relationship between two variables. Cohen’s $d$, on the other hand, measures the standardized difference between two means, and thus, the unit for the size of the difference is expressed in standard deviations (e.g., $d = 0.5$; a difference of one half of a standard deviation). However, the choice of which effect size to aggregate obviously depends on the nature of the question that the meta-analysis is investigating. As regards the $r$ and the $d$, the $r$ family is more appropriate when the meta-analysis is investigating relationships among two variables $X$ and $Y$ (e.g., relationship between conceptual level and counselor competence), and the $d$ family may be more appropriate for comparisons among different experimental groups (e.g., interventions vs. control groups) or conditions (e.g., pretreatment vs. posttreatment). In part, choosing among effect sizes may be a matter of preference, given that it is easy to convert across effect sizes (i.e., $r = d / \sqrt{(d^2 + 4) / 2}$; Cohen, 1988; and $d = 2 (r) / \sqrt{(1 - r^2)}$; Friedman, 1968). Some (e.g., Rosenthal, 1994) suggest that the $r$ statistic is a more general statistic investigating the relationships between variables; $r$ can describe relationships between variables when one is continuous and the second is either dichotomous (analogous to group differences) or
continuous. As is well known, the $r$ can be converted (i.e., $r^2$) into the percentage of variance the predictor variable explains in the criterion variable, with $r^2 = .01$, .09, and .25 (or $r = .1$, .3, and .5), representing small, medium, and large effects, respectively (Cohen, 1988). Conversely, there may be some intuitive appeal to using an effect size within the $d$ family when the meta-analysis is investigating group differences on some continuous index. Cohen’s $d$, for example, involves dividing the mean difference between two groups by a standard deviation. A $d$ of .5, for example, refers to differences between the groups equivalent to .5 of a standard deviation. A $d$ of .2, .5, and .8 are considered to be small, medium, and large effects, respectively (Cohen, 1988). As an historical note, Haase et al. (1982) found the average effect size for counseling research during the 1970s was slightly larger than a medium effect size (i.e., $\eta^2 = .083$, note: $\eta^2 = r^2$, where $r^2$ describes relationships between continuous variables and $\eta^2$ describes the prediction of a continuous variable from a categorical variable; Marascuilo & Serlin, 1988). Conversely, Wampold (2001) concluded that the effect size for psychotherapy was large ($d = .80$).

There is a specific effect-size metric for the situations in which both variables are dichotomous: odds-ratio test (Fleiss, 1994; Sánchez-Meca et al., 2003). Intuitively, the odds-ratio test identifies the odds of the presence of one variable dependent on the presence or absence of the second variable. For example, the odds ratio indexes the odds that a negative outcome, such as dropping out of school, is dependent on the presence of a risk factor, such as parental divorce. When the CI of the odds ratio includes 1.0, the null hypothesis that the criterion variable is not dependent on the predictor variable cannot be rejected.

Having decided on an effect-size metric, the researcher needs to cull the necessary statistical information to calculate effect sizes. The statistical information required to conduct meta-analyses is basic and can include the sample size and either a statistical value (e.g., $t$ statistic) or means and standard deviations associated with an effect. It is worth noting that researchers can also compute effect sizes using only the sample size and $p$ values (Rosenthal & Rubin, 2003), although the failure to report exact $p$ values (e.g., $p < .05$) may complicate this conversion. In these situations, although inexact $p$ values can be converted to inexact effect sizes, the resulting effect sizes will most likely result in conservative (i.e., underestimating) estimates.

Regardless of the specific effect-size metric, Hunter and Schmidt (1994) suggest that effect sizes be adjusted to correct for measurement error. The presence of measurement error will attenuate or underestimate the size of the relationship between imperfectly measured constructs (for more on construct validation, see Hoyt, Warbasse, & Chu, 2006 [this issue]). Researchers can correct for the attenuation in effect sizes associated with measurement error...
if the primary studies report information about the reliability of instruments used. Rosenthal (1994) recommends that these corrections be applied with caution as inappropriately applying them could yield misleading results.

Although the statistical information required for meta-analyses from primary studies is basic, one frustrating aspect of meta-analyses is securing the requisite information to calculate effect sizes. However, there are numerous sources available with formulas to convert basic statistical values into effect sizes (e.g., Rosenthal, 1994; Rosenthal & Rubin, 2003; Rosnow & Rosenthal, 2003). Moreover, journal policies encouraging authors to report effect sizes are making it easier to conduct meta-analyses. Nonetheless, the editorial policy encouraging reporting of effect sizes is not consistently followed (Vacha-Haase & Thompson, 2004). Moreover, the increasing reliance on multivariate statistics further complicates deriving simple effect sizes from published information.

It is important to note that effect sizes are most easily and consistently interpreted when they describe a relationship or association between two variables. As primary studies become increasingly complex such as by using multiple indices of the criterion variable (e.g., client outcome), many studies report multiple effects generated in a single primary study (structural equation modeling also enables more complex studies; see Martens & Haase, 2006 [this issue]; Weston & Gore, 2006). The inclusion of multiple effect sizes from a single study in a meta-analysis violates the assumption that effect sizes are independent when aggregating across studies, unless the dependencies are controlled. Early meta-analyses routinely violated this assumption. More recently, the most common, albeit conservative, solution is to allow each study to contribute only one effect size (Rosenthal, 1994).

There are ways involving matrix algebra to combine multiple effect sizes into one effect that accounts for the interdependencies among the variables (see Cheung & Chan, 2004; Gleser & Olkin, 1994). There are also procedures that correct for the dependencies involved in allowing multiple effect sizes from a single study to be aggregated with effects from other studies (Hedges & Olkin, 1985; Rosenthal, 1991). These complex procedures increase the power analysis but should be used cautiously to prevent generating misleading results.

Another difficulty involved in converting the statistical information reported in primary studies into a common effect-size metric occurs when the effect sizes are partial correlations, beta coefficients in regression analyses, analysis of covariance, or from factorial designs in which the relationship between two variables is calculated after controlling for other variables. It would be problematic to attempt to aggregate effect sizes without some adjustment, when some are based on univariate (e.g., correctional) analyses and others on analyses that control for the variance associated with third
variables. Describing the procedures to adjust these kinds of effect sizes is beyond the scope of this article, but the interested reader can find guidelines and formulas in a number of sources (Gillett, 2003; Keef & Roberts, 2004; Peterson & Brown, 2005).

We return to our illustrative study to describe the ways researchers can convert primary study data into a consistent metric for a meta-analysis. After selecting the studies for inclusion and coding them based on equivalence, Baskin et al. (2003) calculated the relative effect sizes between the bona fide treatment and the placebo control group within each study. Specifically, Baskin et al. compared posttest means between the bona fide treatment and the placebo control group, as did the individual studies. They did this by taking the difference between the means of the two conditions at posttest and dividing by the pooled standard deviation (Hedges & Olkin, 1985). Because effects sizes are associated with positive and negative indices (e.g., well-being vs. symptoms), the direction of the mean differences was made consistent before aggregating across studies so that a positive mean difference indicated that bona fide treatments had better outcomes than did the placebo control groups. Several primary studies in the Baskin et al. meta-analysis used multiple outcome measures and, therefore, combined effect-size estimates within them following Gleser and Olkin’s (1994) guidelines. This method allowed for aggregation of these within-study effect sizes considering the dependency among them. Thus, each study was represented with a single effect size.

**Aggregating Effect Sizes**

Aggregating effect sizes is complicated with different formulas associated with different underlying statistical models and assumptions (i.e., fixed- vs. random-effects models). We first discuss procedures to aggregate effect sizes based on the fixed-effects model to provide the reader with an intuitive sense of the process of aggregation and then describe the issues involved in choosing between random-effects and fixed-effects statistical models.

Early meta-analyses involved simply averaging across the effect sizes of obtained studies. There are several problems with such unweighted averages. First, it seems problematic to give equal weight to an effect size based on a sample of 10 and to an effect size based on a sample of 1,000, because the former is less reliable. Consequently, we recommend that researchers use the procedures described by Hedges and Olkin (1985) and Shadish and Haddock (1994) to aggregate across primary studies. These procedures involve weighting the contribution of a primary study’s effect size by the
inverse of its variance, which gives more weight to those studies with larger samples. Hedges and Olkin’s formula will yield an estimation of the population effect size from the sample of studies. An analogous formula, using similar weighting procedures, yields the variance of that effect size in which more weight is given to larger samples when determining the variability of effect sizes. These two procedures produce a mean effect size as well its variability, which are the two elements necessary to construct a CI. One simple test of statistical significance is to determine if 0 is included in a 95% CI (Hedges & Olkin). There are other formulas available to detect the statistical significance of the aggregated effect sizes. One such formula is the Stouffer method (see Hedges, 1994; Rosenthal, 1995), which involves converting the \( p \) values of individual effect sizes into \( Z \) statistic and aggregating the \( Z \) values. The aggregated effects size is significant if the aggregated \( Z \) statistic is greater than 1.96.

**Homogeneity of Effect Sizes**

The significance of the aggregated effect size is based on the variability of the estimate. In addition to error variance, other sources of variance may be influencing the estimate of effect size. Consequently, it would be important to test for the homogeneity of the effect sizes across the sampled studies to determine if there are subgroupings of studies to be analyzed separately. Specifically, the homogeneity test statistic \( Q \) determines if effect sizes from the studies are homogeneous or reasonably similar to one another (Hedges & Olkin, 1985). The \( Q \) statistic tests if the variability observed among the effect sizes is within expectations because of sampling error in the distribution of the effect sizes. The \( Q \) statistic has a \( \chi^2 \) distribution with \( k - 1 \) degrees of freedom (where \( k = \) number of studies) under the null hypothesis (Hedges & Olkin, 1985). A significant \( Q \) statistic suggests that there could be unexamined systematic differences among the studies such as differences in study population or outcome measure. For example, in Shin et al. (2005), homogeneity statistics for all their aggregated effect sizes were significant, indicating that the primary studies in their analyses differed from one another more so than what could be expected by chance. Therefore, although systematic differences among the studies (e.g., setting, treatment modality) could have masked any significant effects, Shin et al. could not follow up to investigate these possibilities, because they could not obtain enough information about these factors from the original studies.

A significant \( Q \) statistic would suggest the following next steps: (a) examining the distribution of effect sizes for possible outliers and, if so, excluding the outliers and (b) determining if there is a moderator variable (e.g.,
one of the coded variables) accounting for the variability. If the moderator is a continuous or linear variable, then the researcher can use regression procedures (see, e.g., Hunter & Schmidt, 1994). An alternative procedure would be to categorize the studies into subgroups and calculate the $Q$ statistics on the subgroups until the $Q$ values become nonsignificant. For example, in two of the Shin et al. (2005) analyses that examined African American and Caucasian functioning based on client-clinician racial and ethnic matching, there was one study that could have been suspected as an outlier based on the observed effect size (i.e., Chinman, Rosenheck, & Lam, 2000). Compared with the other studies in the analyses, it could be suspected that Chinman et al. (2000) has significantly influenced the aggregated effect size. Shin et al. could have removed this potential outlier and recomputed the $Q$ statistic to determine if it was no longer significant.

However, it is important to note that decisions to subdivide the studies into subgroups or eliminate studies as outliers cannot be based solely on the observed statistics (i.e., homogeneity statistic and effect sizes of the studies) or on examining studies for the effect they have on the distribution of effect sizes, as there are situations in which a significant $Q$ statistic should not prohibit the researchers from aggregating effect sizes across a heterogeneous grouping of studies. The first case is where a significant degree of heterogeneity (Shadish & Haddock, 1994) occurs because, as is true for all statistical tests, the high statistical power leads to statistical significance in the absence of practical significance. In other words, large within-study sample sizes (i.e., $N$) could sometimes lead to rejection of the null hypothesis of homogeneity because of differences among studies that are statistically different but do not have practical significance (Shadish & Haddock, 1994). A second and more important case is that although the effect sizes may differ, there may be no theoretical or methodological basis to exclude the studies contributing to the heterogeneity. For example, eliminating outliers without any justifiable reason not only reduces the data but could also lead to doubts by other researchers regarding the study’s objectivity. Therefore, when outliers are identified, researchers must consider whether removing them from the meta-analysis is justifiable on theoretical and methodological bases (see, e.g., Hedges, 1987; Johnson & Eagly, 1989). The third case is when researchers wish to calculate an aggregate effect size despite the known sources of heterogeneity to demonstrate the robustness of the association between the variables (Shadish, Cook, & Campbell, 2002). For example, if the researchers were interested in demonstrating that psychotherapy treatment produces substantial differences between pretest and posttest regardless of possible differences among schools of therapy (e.g., cognitive behavioral therapy, interpersonal psychotherapy), they may still choose to aggregate the effect sizes from different therapies even if the
observed effect sizes among the schools were known to differ at statistically significant yet clinically insignificant levels. Therefore, when aggregating the effect sizes is of particular interest, it may still be reasonable even with heterogeneous estimates. In any of the three cases, however, researchers should not interpret the aggregated effect size \( d \) as an estimate of a single parameter; rather, they should consider it as simply the mean of the observed effect sizes (Shadish & Haddock, 1994). Based on the earlier reasons, in the case of Shin et al. (2005), it would not have been appropriate for the researchers to eliminate a potential outlier based solely on the observed effect size of the study or the significance of the \( Q \) statistic unless they provided a solid explanation for doing so.

Baskin et al.’s (2003) hypotheses suggested that the effect sizes would systematically differ based on whether the placebo control groups in the individual studies structurally differed from the treatment groups. Thus, they expected that the \( Q \) statistic would be significant (i.e., homogeneity rejected) because of these structural differences. Although they did not report this statistic in their study, we calculate it for instructional purposes. As Baskin et al. expected, the \( Q \) statistic is significant \((Q = 59.26, df = 20, p < .001)\). It is important to note that the \( Q \) statistic does not need to be significant for the researchers to test for moderator variables if hypothesized a priori. We advise that researchers should categorize into groups based on possible moderators before aggregating effect sizes, as was the case for Baskin et al., to minimize potential bias. However, if no moderator was hypothesized and the \( Q \) statistic is nonsignificant, it would be inappropriate to explore for possible moderators as such an exploration would likely capitalize on chance variation and could be potentially misleading (Hunter & Schmidt, 1990). One last caveat is important to note when interpreting the \( Q \) statistic: As in all other statistical tests, a nonsignificant \( Q \) statistic may not provide sufficient evidence to establish homogeneity if statistical power is not relatively large. In short, the \( Q \) statistic can be helpful to (a) detect unanticipated levels of heterogeneity and, if so, implicate the search for the heterogeneity’s source and (b) provide statistical support for an a priori hypothesized degree of heterogeneity. Researchers should interpret the \( Q \) statistic in the context of conceptual factors about possible sources of heterogeneity.

Another crucial point is that significant effect-size differences among groups that were categorized a priori on a hypothesized moderator effect may not necessarily support the intended moderator. In other words, although the researchers intended to categorize the groups on a certain moderator, there may be other confounding variables in the groupings influencing the results. Allumbaugh and Hoyt’s (1999) meta-analysis of grief therapy is a good example. The researchers divided the primary studies on
whether all clients had a common relationship to the deceased (e.g., groups with clients who lost their spouses vs. groups mixed with clients who lost spouses, children, etc.). However, when the effect size was inconsistent with their hypothesis, they reexamined the studies for potential confounds and found that the groups were also roughly divided on whether the clients were self-referred or were recruited. Therefore, it is imperative that researchers consider potential confounds regardless of whether their hypothesis was supported.

**Fixed-Effects Model**

A complicated, yet critical component in aggregating effect sizes is deciding whether to adopt a fixed- or random-effects model. The aggregation procedures detailed earlier apply to the fixed-effects model; procedures for aggregating effect sizes using a random-effects model are described further on. Deciding which model to use depends on whether the researchers wish to generalize beyond their sample of studies. Typically, researchers conduct meta-analyses for one of two purposes with regards to generalization: to estimate an effect size for the specific sample of studies or to make inferences about the population from which the studies were supposedly sampled. For example, counselors may wish to evaluate their clinic’s current psychotherapy treatment effectiveness by aggregating all of their effect sizes associated with each counselor’s work with clients (i.e., fixed effects); however, the same counselors may wish to infer about other similar clinics based on what is observed in their clinic (i.e., random effects). When using the fixed-effects model, the researcher is interested only in the observed effect among the sampled studies, which Hedges and Vevea (1998) refer to as conditional inference:

> Conditional inferences apply to this collection of studies and say nothing about other studies that may be done later, could have been done earlier, or may have already been done but are not included among the observed studies. (p. 487)

When making a conditional inference, a fixed-effects meta-analysis model is appropriate. In contrast, if the researcher wishes to generalize to the population of studies (as opposed to the specific sample of studies), then the random-effects model is appropriate. For the remainder of the article, the notation $d_+$ indicates an aggregated effect-size estimate under the fixed-effects model.

Conceptually, the fixed-effects model assumes that all effect sizes under consideration derive from a common parameter of interest and are expected to converge around a single point. For example, under this model, researchers interested in estimating a clinic’s overall psychotherapy effectiveness assume
that the differences in the observed effect sizes associated with each counselor are because of random variability across a range of factors, such as therapeutic alliance or trivial factors contributing to sampling error. If the only source of difference is because of random variability, then there would be one estimate of effect size across all the counselors. Therefore, if all counselors in the agency were to be measured infinitely under ideal conditions, their results would converge; in other words, the fixed-effects model assumes that all counselors produce comparable outcomes and that their differences are random variation.

The earlier assumption regarding the fixed-effects model significantly limits the potential inferences regarding the aggregated effect size. Specifically, as the only source of variability assumed under this model is sampling error, the possible conclusion pertains only to the aggregated set of effect sizes. For example, the counselors interested in assessing their clinic’s overall effectiveness could interpret only the aggregated effect size as a current snapshot of their clinic’s current level of effectiveness as observed among the counselors providing the data.

In addition, heterogeneity of effect sizes (i.e., a significant $Q$ statistic) when using a fixed-effects model poses a threat to the validity of the meta-analysis because heterogeneity could potentially inflate Type I error (the likelihood of incorrectly rejecting the null hypothesis). Specifically, because the CI of the aggregated effect size narrows by the increase in aggregated studies under this model (Cohn & Becker, 2003), including irrelevant effect sizes could erroneously increase the power of the meta-analysis. For example, if the counselors also erroneously included effect sizes of psychiatrists in the clinic who solely prescribed medication, this may lead to an inflated Type I error rate even if the psychiatrists did not produce effect sizes as large as counselors. Therefore, when the $Q$ statistic is significant under the fixed-effects model, it is crucial to reexamine the aggregated effect sizes to identify potential outliers and moderating variables (such as counselors vs. psychiatrists; please see section on moderator effects further on).

**Random-Effects Model**

In contrast to the fixed-effects model, the random-effects model does not assume that all effect sizes are similar, and consequently, the model simply incorporates this variability into the aggregated estimate. Thus, in a random-effects model, researchers make unconditional inferences:

The analyst wishes to make inferences about the parameters of a population of studies that is larger than the set of observed studies and that may not be strictly identical to them. (Hedges & Vevea, 1998, p. 487)
The effect size aggregated using a random-effects model is denoted as $d_{+*}$ for the remainder of the article.

If the purpose of the researchers is to infer the clinic’s effectiveness as a whole without limiting the inferences to those counselors who provided data, we would advise them to use the random-effects model. Unlike the fixed-effects model, the random-effects model incorporates in its effect-size estimate any variability in the data other than what is expected from sampling error. Thus, if the aggregated pre-effect, post-effect size under the random-effects model was significant (i.e., CI did not include 0), then the researchers could claim that their center as a whole was effective, rather than just the counselors who provided the data.

Comparing Fixed-Effects and Random-Effects Models

In more technical terms, compared to the fixed-effects model, the random-effects model assumes that there is an additional source of variance that needs to be accounted for in the model (Raudenbush, 1994). Specifically, whereas the only variance that the fixed-effects model considers is associated with sampling of participants within the studies (and hence its aggregate for $d_{+}$), the random-effects model also accounts for the variance in sampling of studies. Thus, for the random-effects model, the researcher needs to estimate the variance associated with study sampling. Therefore, in a random-effects model, the $Q$ statistic is used for significance testing of the study sampling variance and for estimating its magnitude (Hedges & Vevea, 1998; Shadish & Haddock, 1994), whereas in the fixed-effects model, the test statistic is interpreted differently, indicating whether the effect sizes among the studies, which, in this case, are not samples from a population but the actual population of interest, are homogeneous. Therefore, in a random-effects model, if the $Q$ statistic is significant, it indicates that the study sampling variance is significantly different from 0 and that this variance needs to be incorporated throughout the analysis. In addition, as the random-effects model wishes to incorporate the variability among studies, the $p < .05$ for the $Q$ statistic is not overly restrictive. Although it appears that the random-effects model yields more meaningful and powerful conclusions related to interpreting and generalizing the statistical results, Cohn and Becker (2003) note important limitations with random-effects models. They note that the studies included in a meta-analysis rarely represent a random selection of studies, which violates a critical assumption of the random-effects model.

In many cases, the comparison between the two effect sizes $d_{+}$ and $d_{+*}$ estimated under the fixed-effects and random-effects models appears to differ little, if at all, in magnitude. Indeed, for technical reasons, when the
homogeneity statistic $Q$ is smaller than the degrees of freedom (i.e., $k - 1$), the two estimates $d_+$ and $d_+^*$ are identical. Nonetheless, the potential inferences based on the fixed-effects versus random-effects model differ in important ways. The aggregated effect size $d_+$ is the mean of the specific set of effect sizes in the given sample, whereas $d_+^*$ is the estimate of the mean of a sample population. Hence, researchers cannot use fixed-effect estimates to infer estimates of effect sizes in the population.

On the other hand, when the $Q$ statistic is larger than the degrees of freedom, the variance and hence the CI of the aggregated effect size $d_+^*$ becomes larger than $d_+$, obviously because of the significant amount of variance associated with study sampling (Cohn & Becker, 2003; Hedges & Vevea, 1998). In other words, when there is considerable variability among the studies’ effect sizes, the random-effects model has a reduced possibility of rejecting the null hypothesis and thus becomes more conservative. It is important to note that selecting a fixed-effects model based simply on either increased statistical power or a nonsignificant $Q$ statistic is not justifiable, and researchers should determine this only by thoroughly considering the inferences that they wish to make (Hedges & Vevea, 1998; Shadish & Haddock, 1994). For example, comparison of two or more treatment groups in clinical trials specifies treatments as fixed but therapists as random (Serlin, Wampold, & Levin, 2003; Wampold & Serlin, 2000). With regard to treatment, researchers aim to infer about the specific treatments that were investigated and not the mean of the population of different treatments. However, researchers do not want to limit their inference on the specific counselors in the trial but rather intend to infer about the whole population of possible counselors.

Baskin et al.’s (2003) use of the fixed-effects model therefore limits their inference to only the primary studies they included. However, for illustration, we compute both the fixed-effects and the random-effects models using their data. As mentioned earlier, because the $Q$ statistic is significant, the overall $Q$ statistic in Baskin et al. (2003) was $Q = 59.26$ ($df = 20, p < .001$). Thus, aggregation based on the fixed-effects and random-effects models will differ as the effect sizes among the studies varied considerably from one another.

Using the fixed-effects model, the aggregate effect size of the eight studies that compared treatments with placebo controls with nonequivalent structures resulted in $d_+ = 0.465$, with a 95% CI of $d = 0.309 \sim 0.621$. The aggregated effect size for the 13 studies comparing treatments with equivalent placebos resulted in $d_+ = 0.149$, with a 95% CI of $d = 0.055 \sim 0.292$. Under the random-effects model, effect sizes were $d_+^* = 0.472$ (95% CI: $d = 0.189 \sim 0.755$) for the eight nonequivalent comparisons and $d_+^* = 0.139$ (95% CI: $d = -0.090 \sim 0.369$) for the 13 equivalent comparisons. As illustrated,
although the estimates of the aggregate effect sizes between the fixed-effects and random-effects models do not differ greatly in magnitude, the CI under the random-effects model is considerably larger than under the fixed-effects model. In particular, the nonequivalent comparisons in the 13 aggregated studies fail to reject the null hypothesis under the random-effects model (i.e., CI includes 0), thus leading to a different conclusion than the fixed-effects model. Hence, under the random-effects model, statistical support for concluding a difference was provided only for the population of studies comparing treatments and structurally nonequivalent control groups but not for the population of studies comparing treatments and structurally equivalent control groups.

**Moderator Effects**

Detecting a significant moderator effect is an important component of meta-analyses. Because effect sizes index the strength of an association between the independent and the dependent variables, detecting a moderator variable in meta-analysis requires only a test for a linear association between the effect size and the moderator variable (Shadish & Sweeney, 1991). The appropriate statistical methods to test moderator effects depend on two factors. The first factor is whether the moderator variable is categorical or continuous. When the moderator variable is categorical, researchers use a statistical procedure that is analogous to a weighted ANOVA; on the other hand, researchers test effects of continuous moderators by applying the weighted multiple regression (Hedges, 1994; Hedges & Pigott, 2004). Because ANOVA is a subset of regression analyses, researchers could use weighted regression when the moderator variable is categorical or continuous. The second factor associated with a test for moderator effects is whether aggregated effects sizes were tested under the fixed-effects or the random-effects model. Following Hedges and Pigott (2004), when the aggregated studies within the groups are considered random, we will refer to this moderator analysis as a mixed-effects model because although the moderator variable itself is fixed, the effect sizes of the groups are random. Although we advise readers to refer to other sources for technical details (e.g., Hedges, 1994; Hedges & Pigott, 2004; Raudenbush, 1994), we provide a conceptual overview of moderator analysis.

For the fixed-effects model, the omnibus test uses the $Q$ statistic in a straightforward way. Recall that for the fixed-effects model, the homogeneity statistic $Q$ tested whether the effect sizes significantly differed from one another. As studies are categorized into groups based on a priori criteria, the sources of variance are also divided into within-groups and between-groups variance. Under the fixed-effects model, all variance is divided
between these two sources as the model assumes that all variance is attributable to the included studies. Consequently, under this model, the overall $Q$ statistic is completely divided into two statistics, $Q_W$ and $Q_B$, respectively, for within-groups and between-group test statistic for homogeneity (i.e., $Q = Q_W + Q_B$; Hedges & Olkin, 1985). This is analogous to the sum of squares within and between, $SS_W$ and $SS_B$, respectively, in an ANOVA. As $Q_B$ is the between-groups variance, the test of its significance is compared to the $\chi^2$ critical value at $df = p - 1$ (where $p = \text{number of groups}$) and this becomes the omnibus test for between-groups differences.

After detecting a significant $Q_B$ statistic, researchers may be interested in comparing the mean effect sizes among the different groups. As in primary studies, researchers test differences among group means by using contrasts (on using discriminant analysis to compare groups, see Sherry, 2006 [TCP, special issue, part 1]). For example, let us assume a meta-analysis comparing the effectiveness of therapy based on client age. If the clients were categorized in, say, three age groups—notably, younger than 18 (Group A), between 18 and 59 (Group B), and 60 and older (Group C)—the researchers may be interested in investigating whether there are differences in effectiveness between the two age groups within adults (i.e., comparing Groups B and C). Or the researchers may be interested in comparing possible differences in effectiveness between children and adolescents versus adults (i.e., comparing Group A with average of Groups B and C). In cases where researchers are interested in a categorical moderator, they use procedures identical to tests of contrasts that follow an ANOVA (for detail, see Hedges, 1994). In cases where age is not categorized but is kept continuous, researchers use a multiple regression analysis (Hedges, 1994).

In a mixed-effects model (i.e., involving a random-effects model for the aggregated effect size and a fixed-effects model for the moderator analysis), researchers conduct a similar omnibus test using $Q_B^*$, which is the between-groups variance for the mixed-effects model. The difference between $Q_B$ and $Q_B^*$ is that when $Q$ is larger than its degree of freedom (i.e., $k - 1$), $Q_B^*$ becomes smaller than $Q_B$ because more variance is accounted for as within-groups variance resulting from two unknown factors (i.e., $Q_W^*$, or within-group homogeneity statistic, becomes larger; however, $Q \neq Q_W^* + Q_B^*$ if $Q$ is significant). Therefore, it is harder for $Q_B^*$ to reach the $\chi^2$ critical value at $df = p - 1$, meaning that the null hypothesis that the groups do not differ is harder to reject. Hedges and Pigott (2004) provide procedures for determining the power associated with tests for moderators in meta-analyses. These power analyses are useful to conduct prior to conducting the study to determine if there is sufficient power to yield meaningful results. Moreover, power analyses may be important to conduct subsequent to failing to reject the null hypothesis for a test of moderation to facilitate interpretation of the
null results to rule out that the moderator was nonsignificant because of limited statistical power (Hedges & Pigott, 2004).

In Baskin et al. (2003), the researchers were interested in investigating whether the structural equivalence of the placebo controls affected the relative efficacy between the treatment and the placebo control conditions. In other words, the researchers suspected that the equivalent or nonequivalent variable moderated differences in effect sizes associated with the treatment and placebo conditions. Baskin et al. used a fixed-effects model in their study, which implied that any demonstrated effect by the moderator would generalize only to the sample of studies included in the meta-analysis. On the other hand, researchers could also use a mixed-effects model for this analysis, which would allow researchers to extend the interpretation to a population of studies comparing treatments with placebo control groups. For illustration, we describe both fixed-effects and mixed-effects models. In Baskin et al., the moderator variable was dichotomous—the structure of the placebo control group was either equivalent or nonequivalent. In this case, the omnibus test evaluates the difference between these two groups, and no tests of other contrasts are needed because there are only two groups.

Whether the $Q_B$ statistic was significant indicated whether structural equivalence between the treatment and the placebo group moderated the relative effectiveness between the two groups of comparisons. In other words, the between-groups (i.e., $Q_B$) comparison of equivalent versus nonequivalent conditions tested whether structural equivalence affected the effect-size differences between placebo and treatment. As Baskin et al. reported, $Q_B = 8.57$ ($df = 1$, $p < .005$) was significant, indicating that in the aggregated studies, the relative effect sizes between treatments and nonequivalent placebos were significantly different than effect sizes between treatments and equivalent placebos. This analysis indicated (for these studies) that the magnitude of the effect-size differences between the placebo and the treatment was related to nonequivalence in structure. However, a reanalysis of the same test using the mixed-effects model implicated a different statistical conclusion, as $Q_B^*$ resulted in nonsignificance ($Q_B^* = 3.21$, $df = 1$, $p = .073$). Specifically, it indicated that (for the population of comparisons between placebo and treatments) there was no evidence to suggest that the magnitude of the effect-size differences between the placebo and the treatment was related to nonequivalence in structure. The reason for this difference is that in the mixed-effects model, more within-group variance is taken into consideration than in the fixed-effects model. Given that research often intends to infer about the population, it is unfortunate that their results were not generalizable.

Based on these results, the potential inferences differ. Results from the fixed-effects model would allow the researchers to conclude that the studies
(included in this meta-analysis) comparing the treatments against structurally nonequivalent placebo groups produced larger relative effect sizes as compared with the studies (again, included in this meta-analysis) comparing the treatments against structurally equivalent placebo groups. However, based on the random-effects model, the researchers should conclude that there was insufficient evidence that the population of studies comparing treatments against structurally nonequivalent placebo control groups produces larger relative effect sizes as compared with the population of studies comparing the treatments against structurally equivalent placebo control groups. Hence, Baskin et al. (2003) had sufficient statistical power to conclude that there were significant differences based on the structure of the control groups for the sample of studies investigated but not to conclude that these differences could be generalized to the population of studies conducted.

**Counternull Value**

If a meta-analysis yields a nonsignificant aggregated effect size, it may be tempting to conclude that the effect is 0, but this conclusion is simply accepting null results, which can be problematic. To avoid the problem, Rosnow and Rosenthal (2003) have suggested that when the CI of the aggregated effect size includes 0, it would be important to determine a counternull value. The counternull value is double the effect size (for technical justification, see Rosnow & Rosenthal, 2003), which is the same distance from the estimate of the effect size as 0 is. For example, an aggregated effect size (d) of .20 with a CI ranging from –0.10 and 0.50 would be considered nonsignificant because 0 is included in the CI. The counternull value is 0.40. Rosnow and Rosenthal point out that in this example, there is as much evidence supporting the conclusion that the best estimator of the effect size is 0 (null finding) as there is supporting the conclusion that the effect size is 0.40. Shin et al. (2005) recently conducted a meta-analysis testing the effects of racial matching between counselor and client. They concluded that there were no overall effects of client-clinician racial and ethnic matching for African American and Caucasian American clients, but they did not consider a counternull value in their interpretation. Computing a counternull value for one of their comparisons suggested that there was as much statistical support for concluding that there was no effect as there was for concluding that there was an effect size of –.584, which would be a moderately large effect size suggesting that African Americans fared worse in their functioning after counseling when seen by a Caucasian therapist than when seen by an African American therapist. It would seem, therefore, premature to conclude that there were no effects associated with the influence of client-therapist racial matching on client outcome. This
example demonstrates the importance of considering counternull values to minimize the logic problems associated with accepting a null hypothesis by demonstrating that another point estimate, the counternull value, is as reasonable as concluding the effect size is 0.

File-Drawer Analysis

Unlike narrative reviews, there are procedures developed in meta-analyses to address the bias associated with disproportionately sampling among studies with publishable results and underrepresenting those studies languishing in researchers’ file drawers and on computer hard drives. To address the file-drawer problem, Rosenthal (1979) reasoned that the meta-analyst could calculate the number of unpublished and unavailable studies that would need to exist to change the conclusions of a significant meta-analysis. The larger the number of studies included in a meta-analyses (other factors being equal) the smaller the CI and, consequently, the more studies that would be needed to increase the size of confidence such that 0 would be included and, thereby, overturn the results of the meta-analysis. Specifically, Rosenthal (1979) derived the Fail Safe $N$ that suggests the number of unpublished and unavailable studies with an average effect size of 0 that would have to exist before the meta-analysis results would be overturned. In addition, he proposed a criterion that the number of studies that the Fail Safe $N$ should be just more than 5 times the number of studies (actually $5K + 10$, where $K$ equals the number of studies in the meta-analysis) that were included in the meta-analysis before the file-drawer problem could be reasonably ignored. However, Iyengar and Greenhouse (1988; see also Hsu, 2002) noted that Rosenthal’s Fail Safe $N$ was based on the assumption that the only studies languishing in file drawers would be non-significant ones but noted that studies that were significant but in the direction opposite to what was hypothesized may also be lost in file drawers. Consequently, Iyengar and Greenhouse reports on an adjusted formula (adjusted Fail Safe $N$) that corrects for this potential bias, which researchers could use in situations where unpublished studies demonstrating significance but in the wrong direction may be a concern to the validity of the meta-analysis (for instructions on solving Iyengar and Greenhouse’s formula, see Hsu, 2002).

Hedges and Pigott (2001) suggest another way to address the file-drawer problem, which is that the publication bias involved in meta-analyses is likely greatest for studies with small sample sizes. The logic behind this suggestion is that large studies are more likely to be published in part because of the increase in statistical power over studies with small samples and in part because they are likely to be reviewed more positively irrespective
of statistical significance relative to studies with small samples. Conn et al. (2003) found that the biggest difference between published and unpublished studies was statistical significance; surprisingly, there were minimal differences in methodological sophistication. Hedges and Pigott recommend that whenever possible, the meta-analysis include only studies with large samples to further minimize the file-drawer problem. However, we recommend that researchers use the aggregation procedures described earlier, which give greater weight to studies with small variances (i.e., large samples) to reduce the file-drawer problem rather than excluding the studies with small samples overall.

A third strategy for a file-drawer problem is to use a funnel plot, in which researchers plot the effect sizes against the sample sizes of the effect sizes. A publication bias would be suggested when the shape of the plot indicates that only those studies with large sample sizes correspond to large effect sizes and in which the center of the plot (corresponding to small effect sizes and sample sizes) is hollow (for more details as well as an alternative plotting strategy to detect publication bias, see Wang and Bushman, 1998). Each of these three strategies provides ways to address the file-drawer problem but does not eliminate it.

Here, for illustration, we calculate the Fail Safe N based on formulas of Rosenthal (1979) and the adjusted Fail Safe N by Iyengar and Greenhouse (1988; Hsu, 2002). In Baskin et al. (2003), the comparison that was significant under the random-effects models was between the treatment group and the placebo control group that was not structurally equivalent with the treatment. As reported, the effect size of the difference between the groups was \( d^* = 0.472 \) (random-effects model, 8 studies). This effect size had a standard deviation of 0.144. Using these values, Rosenthal’s formula for the Fail Safe N results in 246 (see Rosenthal for details on calculation). In other words, it indicates that there needs to be at least 246 studies with null results to refute the conclusion that treatments exceed nonequivalent placebos in efficacy. This exceeds Rosenthal’s criterion introduced earlier (i.e., \( 5K + 10 \)), which, in this case, is 50. Therefore, we could claim that the result is unlikely to be refuted. As a comparison, Iyengar and Greenhouse’s adjusted Fail Safe N formula is more conservative, which, for this example, yields 91. Here, it is important to note that even this conservative estimate exceeds Rosenthal’s criterion.

STEP 5. WRITING UP A META-ANALYSIS

We summarize and add to Rosenthal’s (1995) useful recommendations for what to include when writing up the results of a meta-analysis, which
will also serve to summarize the guidelines discussed earlier. Table 1 illustrates the flow of conducting a meta-analysis.

Introduction

The introduction of a meta-analysis is similar to one for a primary study because both involve summarizing the work that has been done in the area, reviewing literature to support the study hypotheses, and making a case for the value of the study and how it differs from previous research. Specifically, the introduction of a meta-analysis should include, if applicable, other meta-analyses on the topic and why the present one is different and, in some ways, better. The meta-analyst should describe the kind of primary studies that will be included in the meta-analysis and may also present one or more

---

### TABLE 1: Summary of Steps for Meta-Analyses

<table>
<thead>
<tr>
<th>Step 1: Conceptualization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identify need for meta-analysis</td>
</tr>
<tr>
<td>Identify area of research require synthesis</td>
</tr>
<tr>
<td>Investigate areas that are impractical for individual studies</td>
</tr>
<tr>
<td>Formulate a priori hypotheses and determine critical variables for meta-analysis</td>
</tr>
<tr>
<td>Preliminary power analyses</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step 2: Data collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identification of studies</td>
</tr>
<tr>
<td>Search parameters (keywords, search engines)</td>
</tr>
<tr>
<td>Sources (journals, dissertations, reference sections, etc.) to search</td>
</tr>
<tr>
<td>Selection (inclusion or exclusion) criteria (e.g., methodological quality)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step 3: Coding studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding of variables (predictor, criterion, moderator, and control variables)</td>
</tr>
<tr>
<td>Abstract information necessary for calculating effect sizes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step 4: Statistical procedures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conversion to common metric (d, r, \text{ etc.})</td>
</tr>
<tr>
<td>Adjust effect size for measurement error</td>
</tr>
<tr>
<td>Impute missing information when necessary to calculate effect size</td>
</tr>
<tr>
<td>Control for dependencies among effect sizes</td>
</tr>
<tr>
<td>Determine if adopting fixed-effects or random-effects model</td>
</tr>
<tr>
<td>If fixed effects: Test for homogeneity (if heterogeneous, consider outliers, additional moderators, subdividing into homogenous groups)</td>
</tr>
<tr>
<td>Investigate moderator effects</td>
</tr>
<tr>
<td>Aggregate effect sizes</td>
</tr>
<tr>
<td>Weight studies according to individual characteristics (e.g., sample size)</td>
</tr>
<tr>
<td>Derive confidence intervals or significance levels</td>
</tr>
<tr>
<td>If nonsignificant, then compute counternull</td>
</tr>
<tr>
<td>If significant, then consider extent of file-drawer problem</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step 5: Write up meta-analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOTE: Please see text for details about each of these steps.</td>
</tr>
</tbody>
</table>
illustrative primary studies in some detail (Rosenthal, 1995). When appropriate, the meta-analyst would need to discuss methodological innovations associated with the current meta-analysis in the introduction. As with primary studies, the introduction should end with the specification of a priori research hypotheses.

Methods

The meta-analyst should inform readers about what search techniques he or she used (e.g., search engines) as well as the parameters of the search, including the keywords. The meta-analyst should describe other procedures used to identify possible studies, which may include examining the reference sections and contacting experts in the field about published and unpublished data. It is critical to illustrate what strategies were used and how effective they were in identifying unpublished studies. The meta-analyst will also need to identify carefully and specifically the criteria for inclusion and exclusion and how he or she applied those criteria, such as by using raters to determine eligibility for inclusion. The criteria should be conceptually and, when possible, empirically justified. For example, if some studies are excluded because of problems with methodology, the threshold for study quality (e.g., random assignment) should be carefully justified. Moreover, the meta-analyst should describe how she or he chose to handle studies that report more than one effect size and situations when multiple studies were published from a single dataset. Rosenthal (1995) recommends that a preliminary power analysis be conducted and reported to determine if the number of studies obtained are sufficient to make sound conclusions.

The meta-analyst will need to describe procedures to impute statistical information that was missing from the obtained studies, which may happen, for example, when an author fails to give an exact p value for a nonsignificant effect or when some important statistical information is missing (e.g., reliability estimates used to correct for attenuation because of measurement error). Finally, the meta-analyst should report which study characteristics were coded to be examined as possible moderator variables, including which were conceptualized a priori and which were identified post hoc.

Results

As in primary studies, reports of meta-analyses usually provide descriptive information about the sample, which would describe the range and types of primary studies collected for the meta-analysis. A main focus, however, for the results section of a meta-analysis will be describing the
characteristics of the effect sizes. Rosenthal (1995) recommends that authors present the effect sizes visually in a stem-and-leaf display to describe the distribution of effect sizes. If outliers were detected, the meta-analyst should describe the factors associated with deciding which studies were outliers. Importantly, the central tendency of effect sizes will need to be characterized by, for example, means and medians of weighted and unweighted effect sizes, along with the total number of studies involved in the aggregation. The meta-analyst should describe the variability of the effect sizes, which may include standard deviation, the range, as well as the 75th and 25th percentiles (Rosenthal, 1995). Obviously, the meta-analyst should list the CIs around each of the mean effect sizes and, usually, inspect them to determine if 0 is included in the CI. Any other procedures used to determine statistical significance should be reported as well, including if random-effects or fixed-effects models were assumed. If the analyst used a fixed-effects model, then he or she should report the results of heterogeneity tests—significant levels of heterogeneity may compromise the interpretation of results from statistical analyses based on the fixed-effects model.

A method that illustrates the practical importance of an effect size is the binomial effect-size display (BESD; Rosenthal & Rubin, 1982). The BESD converts the Pearson’s $r$ into success rates between two conditions, under the assumption that the overall success rate is 50%. It is displayed in a twocolumn (e.g., treatment or control) by two-level-of-outcome (e.g., success or failure) table, with each row and column summing 100%. The success rate of the treatment condition and the failure rate of the control condition are calculated as $.50 + r/2$, and the failure rate of the treatment condition and the success rate of the control condition are calculated as $.50 - r/2$ (Rosenthal & Rubin, 1982). As any effect-size metric can be converted into $r$, the BESD is not limited to $r$. To illustrate, a meta-analyst could convert the $d$ that Baskin et al. (2003) used to $r$ and present this using BESD. With these data, when treatments are compared against placebos with nonequivalent structure, relative success rates between treatment and placebo conditions differ greatly (Comparison 1 in Table 2). However, when treatments are compared against placebos with equivalent structure, the differences in relative success rates become small (Comparison 2 in Table 2). This method of displaying an effect size has intuitive appeal over simply reporting the actual effect sizes. Nonetheless, it requires an arbitrary assumption that the overall success rate is .50 (Wampold, 2001) and does not provide any unique statistical information earlier reporting the $r$ or other effect-size metrics.

Meta-analysts should conduct and report tests of hypotheses, including analyses for moderator variables and, when investigating group differences, results for omnibus tests and tests of specific hypotheses (e.g., contrasts). If investigating relationships among variables, then the analyst will
TABLE 2:  Binomial Effect-Size Display of Relative Success Rates Between Treatment and Placebo With and Without Equivalent Structure

<table>
<thead>
<tr>
<th>Group</th>
<th>Outcome</th>
<th>% Success</th>
<th>% Failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td></td>
<td>61</td>
<td>39</td>
</tr>
<tr>
<td>Placebo (nonequivalent)</td>
<td></td>
<td>39</td>
<td>61</td>
</tr>
<tr>
<td>Comparison 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td></td>
<td>53</td>
<td>47</td>
</tr>
<tr>
<td>Placebo (equivalent)</td>
<td></td>
<td>47</td>
<td>53</td>
</tr>
</tbody>
</table>

conducted and report correlational or regression analyses. If some results fail to reject the null hypotheses, analysts could determine and report the counterfactual value to guard against the temptation to conclude that the effect sizes were 0. They may also calculate and report power for those analyses in which the null hypothesis could not be rejected to determine if the failure could be attributed to low power. Finally, they should report one or more of the three procedures to address the file-drawer problem, such as funnel plots, calculation of a Fail Safe $N$ or Adjusted Fail Safe $N$, and elimination of those studies with small sample sizes that are likely to reflect greatest levels of publication bias.

**Discussion Section**

As with primary studies, the discussion section for a meta-analysis usually begins with a reiteration of the major findings. The meta-analyst should discuss the implications of the findings as well as any caveats associated with those findings because of limitations in the meta-analysis procedures or data that were available and collected. Meta-analysts usually provide a complete list of the studies included in the meta-analysis, sometimes in the reference section, in which case the studies analyzed would be distinguished from those cited in the body of the meta-analysis report. This list may also appear in an appendix, and if so, meta-analysts should list the effect size(s), sample size, and other descriptive information about each study.

**CONCLUSION**

By now, it should be apparent that meta-analysis represents a powerful and important tool in the development of the science of counseling psychology.
Many of the criticisms associated with meta-analyses are not unique to quantitative syntheses and may be equally or more problematic for traditional narrative reviews. Researchers are just recognizing the use and flexibility of meta-analyses. For example, Newell (in press) combined quantitative analyses with a qualitative discourse analyses in her investigation of race differences in the effectiveness of school-based interventions for behavior disorders. As mentioned earlier, she found that those studies with a majority of African American participants had smaller effect sizes than did those with a majority of White participants. Her analysis of the discourse (i.e., introduction and discussion sections) associated with this research indicated that the justification given for those studies with predominately White participants was focused on the implications of behavior disorder for impeding academic achievement. In contrast, those studies with a majority of African Americans justified the importance of the research in terms of the potential for these participants to commit violent crimes and to become involved in the criminal justice system. The discourse analysis suggested that one reason for the diminished effect sizes for African Americans samples was the way they are engaged in this kind of intervention. Behavioral interventions may be less effective, for example, if the main purpose is to prevent criminality rather than improve the academic achievement of their participants. Newell is just one example of how quantitative meta-analyses may be combined with qualitative as well as traditional narrative syntheses of research.

In addition, we assume that the current review of meta-analysis illustrated the effort required throughout the process to conduct a rigorous analysis. In fact, the actual calculations could perhaps be the easiest part, especially with the available statistical software. There are several commercial software programs specifically for meta-analysis, such as Comprehensive Meta-Analysis (Borenstein, 2001) and MetaWin (Rosenberg, Adams, & Gurevitch, 1999), in addition to commonly used statistical packages (such as SAS and SPSS) and freely downloadable software on the Web. However, meta-analysts spend most of their time, effort, and perhaps patience reviewing the literature, selecting relevant studies, and coding. Therefore, a well-conducted meta-analysis, from conception to write-up, could easily take months and up to several years. Finally, we want to add a caveat that what we have described earlier are only guidelines and that the results of meta-analyses should not be rigidly interpreted (e.g., rigid application of small, medium, and large effect sizes). Rather, researchers need to meaningfully interpret meta-analytic results in the context of the field and the nature of the study (Rosenthal, 1990).

In closing, we envision a time when meta-analyses become so commonplace that most reviews of literatures—including introductions of journal articles, thesis, and dissertations—will involve some meta-analyses.
REFERENCES


Hsu, L. M. (2002). Fail-Safe Ns for one versus two tailed tests lead to different conclusions about publication bias. *Understanding Statistics, 1*, 85-100.


