2014

Meta-analysis of social-personality psychological research

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

Alice H. Eagly
Northwestern University

Johnson, B. T., & Eagly, A. H. (2014). Meta-analysis of social-personality psychological research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd Ed., pp. 675-707). London: Cambridge University Press.

CHAPTER TWENTY-SIX

Meta-Analysis of Research in Social and Personality Psychology

BLAIR T. JOHNSON AND ALICE H. EAGLY*

As in other scientific fields, the progress of social and personality psychology hinges on the orderly and accurate accumulation of empirical evidence about phenomena. This evidence, consisting of multiple studies recording systematic observations of a phenomenon, exists as a literature on the topic. Although new studies rarely replicate earlier studies exactly, many studies are conceptual replications that use different stimulus materials and dependent measures to test the same hypothesis, and still others contain exact replications embedded within designs that add new experimental conditions.

To reach conclusions about empirical support for a particular phenomenon, it is necessary to compare and contrast the findings of relevant studies. Therefore, comparisons of study outcomes – reviews of research – are essential to the scientific enterprise. Until recent decades these comparisons nearly always used methods now known as *narrative reviewing*, which informally draw conclusions about the general trend of the studies' findings, sometimes guided by counts of studies that had either produced or failed to produce statistically significant findings in the hypothesized direction. Such narrative reviews still serve a useful purpose when conducting a comprehensive literature review is not desired or feasible. For example, textbooks typically contain narrative reviews of many hypotheses, and introductions to journal articles reporting primary research usually include brief narrative reviews. These qualitative reviews may suggest useful hypotheses for further scientific investigations.

Despite the usefulness of narrative reviewing, the method does not yield definitive conclusions about the degree of empirical support for a phenomenon or a theory of the phenomenon. One result of this inadequacy is that independent narrative reviews of the same literature often reach differing conclusions. For example, two separate reviews (Brubaker & Powers, 1976; Green, 1981) concluded that those surveyed like younger adults better than older adults, but another review (Lutsky, 1981) concluded that there was little difference. In such cases, it is difficult to determine which conclusion is more accurate.

Critics have pointed to four general faults in narrative reviewing (e.g., Cooper, 2010; Eagly, 1987; Rosenthal, 1991). Although these faults are not necessarily inherent in narrative reviewing, they typify narrative reviewing in practice: (1) Narrative reviewing generally involves the use of a convenience sample of studies, perhaps consisting of only those studies that a reviewer happens to know. Any criteria by which a reviewer selected these studies typically go unstated and may never have been formalized by the reviewer. (2) Narrative reviewers generally do not state their procedures for cataloging studies' characteristics or evaluating the quality of the studies' methods. Any rules or procedures are often not applied uniformly to

* The preparation of this chapter was facilitated by U.S. Public Health Service grants K18 AI094581–01 and R01 MH58563–13 to Blair T. Johnson.

We thank Tania B. Huedo-Medina and Cleo Protogerou for their helpful comments on previous drafts of this chapter.

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

all of the studies in the sample. (3) When study findings differ, narrative reviewers have difficulty reaching clear conclusions about whether differences in study methods explain differences in results. Because such reviewers usually do not systematically code studies' methods, their procedures are poorly suited to account for inconsistencies in findings. (4) Narrative reviewers typically rely on statistical significance to judge studies' findings and not on the magnitude of the findings. Statistical significance is a poor basis for comparing studies that have different sample sizes because effects of identical magnitude can differ in statistical significance. Because of this problem, narrative reviewers often reach erroneous conclusions about the confirmation of a hypothesis in a series of studies, even in literatures as small as 10 studies (Cooper & Rosenthal, 1980). All four of these problems can render narrative reviews inadequate in most contexts in which research is aggregated and integrated.

These potential flaws in the review process become increasingly aggravated as the number of studies available mounts. In contemporary psychology, large research literatures are not uncommon. For example, even as early as 1978, there were at least 345 studies examining interpersonal expectancy effects (Rosenthal & Rubin, 1978). Similarly, by 1983, there were more than 1,000 studies evaluating whether birth order relates to personality (Ernst & Angst, 1983). As the number of studies increases, the conclusions reached by narrative reviewers typically become more unreliable because of the informality of their methods (Johnson & Boynton, 2008).

Because of the importance of comparing study findings accurately, scholars have dedicated considerable effort to making the review process as reliable and valid as possible and thereby avoiding the criticisms that narrative reviews often engender. The result has been the emergence of review techniques that summarize scientific literatures by methods that are themselves consistent with scientific norms. *Quantitative research synthesis* or *meta-analysis* statistically cumulates the results of independent empirical tests of a particular relation between variables. More recently, integrative data analysis of individual-level data has also emerged (e.g., Cooper & Pattall, 2009). Although scientists have cumulated empirical data from independent studies since the early 1800s (Stigler, 1986), relatively sophisticated techniques emerged only after the advent of standardized indexes such as *r*-, *d*-, and *p*-values. In the first published monograph related to these strategies, Glass, McGaw, and Smith (1981) emphasized that reviewing scientific literature is a scientific practice that should follow disciplined and transparent steps. Reflecting the maturation of meta-analysis, Hedges and Olkin (1985) presented a sophisticated version of its statistical bases. Standards for meta-analysis have grown increasingly rigorous, as apparent in the two editions of *The Handbook of Research Synthesis and Meta-Analysis* (Cooper & Hedges, 1994; Cooper, Hedges, & Valentine, 2009).

Social psychologists' first rudimentary applications of quantitative review techniques occurred in the 1960s (e.g., Rosenthal, 1968; Wicker, 1969), but it was not until the late 1970s and early 1980s that scholars applied these techniques to a wide range of social psychological phenomena (e.g., Bond & Titus, 1983; Cooper, 1979; Hall, 1978). In many instances, meta-analyses have overturned or enhanced prior narrative reviewers' conclusions. As one example, Sidanius, Pratto, and Bobo (1994) proposed the gender invariance hypothesis – that, across cultures, males score higher in social dominance orientation than do females. Lee, Pratto, and Johnson's (2011) meta-analysis revealed gender differences that varied considerably in magnitude but did not disappear across the cultures investigated. Within social and personality psychology, as in many other sciences, quantitative research synthesis is now well accepted because scholars realize that careful application of these techniques yields the clearest conclusions about a research literature (Card, 2012; Cooper et al., 2009).

To provide a general introduction to meta-analysis, in the remainder of this chapter we (1) present the steps involved in synthesizing research, (2) consider some options that reviewers should consider as they proceed through these steps, (3) discuss standards for conducting and evaluating quantitative reviews, and (4) evaluate meta-analysis relative to primary research and other methods of testing hypotheses. In treating this subject, consistent with convention, we use the term "meta-analysis" to refer broadly to the entirety of the process, including both quantitative and qualitative aspects.[1]

[1] Strictly speaking, meta-analysis concerns only a statistical integration, the "analyses of analyses" that the term literally connotes. Nonetheless, in practice, reviews that include analyses of analyses are usually labeled meta-analyses, meaning more broadly the entire research synthesis process. For clarity, a *systematic review* is generally one that attempts to grade evidence relevant to a question; it may or may not include meta-analysis per se.

PROCEDURES FOR META-ANALYSIS

An Overview of the Process of Quantitative Synthesis

The research process underlying quantitative synthesis can be broken into discrete steps (Cooper, 2010). Each stage contributes to the next stage; careful work in the early stages makes the later stages easier to accomplish and improves the quality of the overall review. As a preview to a more detailed exposition, we list the stages and some of the questions that often accompany them:

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

Conceptual Analysis of the Literature

The initial conceptual exploration of a research literature is critical because these ideas affect the methods that follow, such as the criteria for including and excluding studies. The first conceptual step is to specify, with great clarity, the phenomenon under review by defining the variables whose relation is the focus of the review. Ordinarily, a synthesis evaluates evidence relevant to a single hypothesis that is defined as a relation between two variables, often stated as the influence of an independent variable on a dependent variable (e.g., the effects of ego depletion on self-control, synthesized by Hagger, Wood, Stiff, and Chatzisarantis, 2010). Moreover, a synthesis must take study quality into account at an early point to determine the kinds of operations that constitute acceptable operationalizations of these conceptual variables. Typically, studies testing a particular hypothesis differ in the operations used to establish the independent and the dependent variables. If the differences in studies' operations can be appropriately judged or categorized, analysts can probably explain some of this variability using these differences as moderator variables.

The research problem's history and its typical studies are essential to this conceptual analysis. Theoretical articles, earlier reviews, and empirical articles should be examined for their interpretations of the phenomenon under investigation. Authors' theories or even their more informal insights may suggest moderators of the effect that could potentially be coded in the studies and examined for their explanatory power. If scholars have debated different theories, the synthesis should be designed to address them, if possible.

The most common way to test competing explanations is to examine how the findings pattern across studies. Specifically, a theory might imply that a third variable should influence the relation between the independent and dependent variables: The relation should be larger or smaller with a higher level of this third variable. Treating this third variable as a potential moderator, the analyst would code the studies for

their status on the moderator. This meta-analytic strategy, known as the *moderator variable* or *effect modifier approach*, is analogous to the examination of interactions with primary-level data (see the section on Estimating Effect Sizes). However, instead of testing the interaction within one study's data, the meta-analysis tests whether the moderator affects the examined relation across the studies included in the sample. Such an analysis determines *when* the magnitude or sign of the relationship varies. Using this strategy, Malle (2006) found that the tendency to explain one's own behavior with situational causes and others' behavior with personal causes holds only for negative events; the opposite asymmetry holds for positive events.

In addition to this moderator variable approach, other strategies have proven to be useful. In particular, a theory might suggest that a third variable serves as a mediator of the critical relation because it conveys the causal impact of the independent variable on the dependent variable (see Judd, Yzerbyt, & Muller, Chapter 25 in this volume; Shadish, 1996). If at least some of the primary studies have evaluated this mediating process, mediator relations can be tested within a meta-analytic framework by performing correlational analyses that are an extension of path analysis with primary-level data. Using such techniques, Albarracín, Johnson, Fishbein, and Muellerleile's (2001) examination of 96 independent studies showed that, consistent with reasoned action approaches, intentions generally mediated the influence of attitudes, subjective norms, and perceived behavioral control on action.

Setting Boundaries for the Sample of Studies

In beginning a meta-analysis, the reviewer should consider whether all possible tests of a relationship should be included. This decision is important because the inferential power of any meta-analysis is limited by the methods of the studies that it integrates. To the extent that all (or most) of the reviewed studies share a particular methodological limitation, any synthesis of these studies would be limited in this respect. For example, a synthesis of correlational studies will produce only correlational evidence about the association in question. Yet if the critical hypothesis were tested with true experiments, defined by one or more manipulated independent variables and the random assignment of participants to conditions, the meta-analysis would gauge the causal effect of the independent variables on the dependent variable across the studies reviewed. Nevertheless, in all meta-analyses, most relations between moderator variables and the effect of interest are correlational and therefore causally ambiguous. For example, Koenig, Eagly, Mitchell, and Ristikari (2011) found that, across three research paradigms, the cultural masculinity of the leader stereotype has decreased over time. Effects of year of publication, like many other study characteristics, can be difficult to interpret because of potential confounds with other variables (e.g., cultural change or change in methods).

Moderator tests can yield stronger causal claims if the moderator reflects within-studies manipulations. In such cases, random assignment of participants to levels of the moderator in the primary studies makes it less likely that confounds were associated with the moderator. In this strategy, the results of each study are divided to produce separate effect sizes within levels of the moderator. For example, Baas, De Dreu, and Nijstad (2008) showed that creativity was enhanced more by positive moods than by neutral ones; moreover, mood valence was experimentally manipulated in most of the studies. If an analysis were limited to the studies that contained this manipulation, any moderation could be more confidently attributed to the manipulated variable, barring confounds with other variables.

In deciding whether some studies may be insufficiently rigorous to include in the meta-analysis, a reviewer should take into account methodological standards within the research area. Although a large number of potential threats to methodological rigor have been identified (see Brewer & Crano, Chapter 2 in this volume; Campbell & Stanley, 1963; Shadish, Cook, & Campbell, 2002; Valentine, 2009), there are few absolute standards of study quality. For example, there are hundreds of scales purporting to gauge methodological quality (Deeks, Dinnes, D'Amico, Sowden, Sakarovitch, Song, Petticrew, & Altman, 2003). Moreover, in practice, the characteristics considered essential to ensure high study quality vary widely across research areas. In some literatures, it is known that a certain method (e.g., a measure or a manipulation) yields seriously flawed results; if so, an analyst might eliminate studies that used this method. Indeed, one possible strategy is to omit obviously flawed studies to restrict the synthesis to studies of high quality, a practice known as *best-evidence synthesis* (Greenwald & Russell, 1991).

Another option is to attempt to correct the effect sizes for certain methodological biases (see the section on Correcting Effect Sizes for Bias). Retaining potentially flawed studies and representing their quality-relevant features in the coding scheme is another defensible strategy, given that methods always contain some degree of error. For example, if a given variable was not manipulated or assessed uniformly across the studies, a coding of the variable's quality (e.g., its reliability) may predict effect size magnitude. More generally, meta-analyses should examine whether variant methods yield differing findings (for an example, see Heinsman & Shadish, 1996; see also Moyer & Finney, 2002).

In addition to study quality, many other considerations enter into setting the boundaries of a research literature. Boundary-setting forces reviewers to weigh conceptual and practical issues, which are particularly acute in literatures featuring a variety of methods. Sometimes boundaries include only studies that are relatively homogeneous methodologically (e.g., only experimental studies), and sometimes boundaries encompass different methods (e.g., both experimental and correlational studies). In general, boundaries should be wide enough to allow the testing of interesting hypotheses about moderator variables. Yet if very diverse methods are included, some moderator variables may exist only within particular methods (e.g., participants' organizational status exists only within studies conducted in organizations). In general, including a wide variety of methods might make a meta-analysis unwieldy. In such instances, meta-analysts may divide a literature into two or more reviews, each addressing a different aspect of a broad research question.

If the boundaries of a meta-analysis are too wide, researchers may be the targets of what is known as the "apples and oranges" critique (Glass et al., 1981) – that is, combining studies that used markedly different methods. Methodologists have been generally unsympathetic to this criticism because they regard it as the task of the meta-analyst to examine whether differences in methods produce consequential differences in study outcomes. This demonstration is achieved by dividing studies into various categories or ranges, as we discuss in the section on Analyzing the Meta-Analytic Database. Of course, meta-analyses that fail to consider moderators can warrant the criticism of ignoring the possible effects that diverse methods have on study outcomes.

Analysts often set the boundaries of the synthesis so that the methods of included studies differ dramatically only on critical moderator dimensions. If other extraneous dimensions are thereby held relatively constant across the reviewed studies, moderator variable analyses can be more clearly interpreted. Meta-analysts proceed by dividing studies based on the moderator variable, where possible, and analyzing the effect of interest within the levels of the moderator (or treating such moderators as continuous variables). Such designs appear frequently in social and personality psychology. For example, because argument quality moderates the effects of involvement on message-based persuasion, Johnson and Eagly (1989) calculated involvement effect sizes within the levels of quality.

Meta-analysts should include all studies or portions of studies that satisfy the selection criteria. If some studies meeting preliminary criteria established conditions that are judged to be extremely atypical (e.g., mentally disabled or ill participant populations), the selection criteria may be modified to exclude them. Developing selection criteria often continues as meta-analysts examine more studies and thereby discover the full range of research designs that have investigated a particular hypothesis.

One issue that generally arises when setting boundaries is whether to include unpublished studies (Rothstein & Hopewell, 2009). Although these studies are usually more difficult to access, their omission typically biases the review's findings in favor of larger effects (e.g., Dickersin, 1997; Johnson, Scott-Sheldon, & Carey, 2010; Lipsey & Wilson, 1993). The frequent omission of nonsignificant findings from the research record is most likely responsible for the so-called *decline effect* (Schooler, 2011), whereby the strength of findings supporting a particular hypothesis decreases after initially appearing robust. Moreover, the withholding of nonsignificant findings from publication appears to be a widespread practice that can compromise the validity of many published effects (Francis, 2012; Ioannidis, 2005). In a discussion of unpublished studies, Rosenthal (1979) referred to them as producing "a file-drawer problem" because they may be buried in researchers' file drawers and therefore inaccessible to reviewers. In fact, surveys of researchers suggest that as much as two-thirds of the studies that are conducted are never published (Cooper, DeNeve, & Charlton, 1997; Rotton, Foos, Van Meek, & Levitt, 1995). Of course, many additional factors affect studies' publication status (e.g., author productivity; Sommer, 1987). A partial solution to the problem of published literatures that are biased in favor of hypotheses is to seek studies that are reported in dissertations and master's theses and as poster sessions and talks at conferences. These studies are less likely to be screened for statistical significance than studies published in journals. Meta-analysts can also ask

researchers in an area if they have additional, unpublished data sets that they can share.

Given these considerations, every effort should be made to obtain unpublished studies. The goal of meta-analysis is to describe the *universe* of studies on a topic, or at least an unbiased sample of that universe (White, 2009). Disregarding this goal compromises the validity of the meta-analysis as a representation of the research literature. Ironically, a meta-analyst would not even learn that this unpublished literature exists without searching for it. Another benefit of including unpublished studies is that they enlarge the number of studies in the meta-analysis, thereby increasing statistical power to estimate mean effect sizes and to detect moderators of effect sizes.

Regardless of studies' publication status, analysts should judge them against a set of inclusion and exclusion criteria and code their quality-relevant features. Uniform implementation of these procedures helps circumvent the potential criticism that unpublished studies are generally of unacceptable quality because of the absence of peer review. Rather than merely assume (perhaps incorrectly) that unpublished studies are of inadequate quality, a meta-analyst should remove all studies, published or unpublished, that do not meet the review's quality criteria and code the remaining studies on quality-relevant study characteristics (e.g., reliability of measures).

A further decision that often arises is whether the sample of studies should be restricted to one country or culture. The reasoning that encourages sampling unpublished studies also encourages sampling studies from all countries and cultures. Moreover, including such studies increases the inclusiveness of the meta-analysis by permitting an analyst to answer questions about the generality of the studied effect across diverse cultures. Indeed, it seems meritorious for meta-analyses with large enough samples of studies to conduct such tests routinely. For example, Bond and Smith (1996) found that conformity in Asch-style line-judgment experiments was more marked in collectivistic than in individualistic cultures (although the conformity effect was significant within both types of cultures). Yet, in many research literatures, it may not be possible to address this issue meta-analytically because only a very small number of studies are available from countries other than the one in which the research paradigm first appeared (e.g., Eagly, Makhijani, & Klonsky, 1992; Lee et al., 2011). Therefore, as a general rule, studies from multiple cultures should be included in the sample if they are available in at least modest numbers. Although computer applications (e.g., Google Translate) can help overcome foreign language barriers, knowledge of a culture's practices can be crucial to coding such studies accurately. Therefore, meta-analysts should typically seek the assistance of native or other highly skilled speakers of the foreign languages represented in the included studies (e.g., Pettigrew & Tropp, 2006).

A final issue is the completeness with which very large research literatures are reviewed. Some literatures are so enormous that including *all* studies would be impractical. In these instances, meta-analysts might take a random sample of the entire research literature (Card, 2012), with sample size guided by statistical power considerations (Cafri, Kromrey, & Brannick, 2009; Valentine, Pigott, & Rothstein, 2010).[2] Specifically, a meta-analyst would list all the studies in the pertinent literature, decide how many would make a sufficient sample, and randomly select this number of studies. An example of such sampling is Rosenthal and Rubin's (1978) meta-analysis of the interpersonal expectancy effect literature.

Locating Relevant Studies

Because including a large number of studies generally increases the value of a quantitative synthesis, it is important to locate as many studies as possible that might be suitable for inclusion. When a literature consists of findings whose presence in reports cannot necessarily be discerned from reading titles and abstracts, a reviewer may have to retrieve all studies in the general research area to identify the finding of interest. For example, Kotov, Gamez, Schmidt, and Watson (2010) screened 7,156 abstracts of studies on traits and anxiety, depression, and substance use; 175 studies fit their inclusion criteria.

Reviewers are well advised to err in the direction of being overly inclusive in their searching procedures. As described elsewhere (e.g., Cooper, 2010; Johnson & Boynton, 2008; Lipsey & Wilson, 2001; White, 2009), there are many ways to find relevant studies; ordinarily, analysts should use all of these techniques. Unfortunately, computer searches of databases such as PsycINFO and Google Scholar seldom locate all of the available studies, although such searches are extremely useful. There are many other databases aside from the most familiar aforementioned ones. Some of these databases cover literature primarily in English (e.g., ProQuest Dissertations and Theses, Web of Science, Sociological Abstracts, MEDLINE, ABI/Inform Global, ERIC). Other databases contain primarily studies published in foreign languages (Psicodoc for Spanish and Portuguese; PSYNDEX for German). Also, other nations maintain databases of dissertations (e.g., Index to Theses and Electronic Theses Online Service, United Kingdom and Ireland; Deutsche Nationalbibliothek and Dissonline, Germany; DART-Europe, pan-European portal for dissertations and theses; China Doctor Dissertations Database). Finally, conference papers and other types of unpublished papers appear in PsycEXTRA (from the American Psychological Association) and ERIC. These databases thus provide partial access to the fugitive literature of unpublished studies (Rothstein & Hopewell, 2009). Databases also increasingly afford full-text searches, which can be very important for literatures in which the focal comparison is less likely to appear in abstracts (e.g., comparison of cooperative behavior of women and men as reviewed by Balliet, Li, Macfarlan, & Van Vugt, 2011). Librarians can provide helpful advice to novice searchers, and many databases offer excellent tutorials (Reed & Baxter, 2009).

Finally, to enable evaluation of search procedures as well as their replication, the review should describe in detail its methods of locating studies, including the names of the databases that were searched, and for each database the time period covered and the keywords used. Reviewers should also describe their inclusion and exclusion criteria and provide a rationale for these criteria, consistent with meta-analysis reporting standards (MARS; American Psychological Association, 2008). More comprehensive standards (e.g., PRISMA; Moher, Liberati, Tetzlaff, & Altman, 2009) include other features, such as a chart describing the flow of study reports into the meta-analysis and a listing of excluded as well as included studies.

STUDY CHARACTERISTICS. In conceptualizing the meta-analysis, reviewers have usually developed ideas about the study characteristics that should be coded. The most important of these characteristics are potential moderator variables that may account for variation among the studies' effect sizes. It is also important to consider whether studies that differ along a critical moderator dimension also differ on other dimensions. Because such confounds could produce interpretational difficulties, coding these additional characteristics potentially permits a meta-analysis to determine which variables explain unique variation in predicting effect size magnitude and which do not. Finally, it is also important to code the studies for numerous other characteristics such as their date of publication and participant population, even if these characteristics are not expected to account for variation in studies' outcomes (Lipsey, 2009), because such features help set an interpretative context for the review.

Study characteristics may be either continuous or categorical. Variables on a *categorical* metric consist of a discrete number of values that reflect qualitative differences between those values. For example, among the categorical study characteristics that Freund and Kasten (2012) coded in a meta-analysis of the validity of self-estimates of cognitive ability were ability type, order of self-estimate and ability test, and gender of participants. Variables on a *continuous* metric consist of values that exist along ratio, interval, or ordinal scales (see Wilson, 2009 for examples).

Some important features of studies are difficult to code accurately by reading study reports. For example, in a meta-analysis on sex-related differences in aggression, Eagly and Steffen (1986) wished to determine whether women and men differed in how unfavorably they perceived aggressive acts. Therefore, they asked female and male students to rate the extent to which each such act would produce harm to the target of aggression, guilt and anxiety in oneself as the aggressor, and danger to oneself. From these ratings Eagly and Steffen estimated sex differences in these students' perceptions of the aggressive acts and related these scores to the effect sizes that represented sex differences in aggressive behavior. In other instances, experts' ratings could be obtained based on their reading of the method sections of the reports or of the actual stimulus materials used in the studies (e.g., Johnson & Eagly, 1989; Marcus-Newhall, Pedersen, Carlson, & Miller, 2000). Similarly, in a review of the involvement and persuasion literature, Johnson and Eagly (1989) provided undergraduate judges samples of the arguments these studies had used and asked them to rate them in terms of their strength in supporting the message position. Such operations help assess dimensions that can prove important in moderator analyses.

Convergent evidence of the reliability and validity of the judges' ratings used by these methods is desirable, because these judges function only as observers of studies' methods. Interjudge reliability estimates

[2] Meta-analysts are wise to consider the potential coverage of the moderators planned for analyses (Card, 2012). Merely randomly sampling studies from the frame of available studies may leave some moderators relatively sparse at values of theoretical interest. In oversampling among extreme values on the moderator, stratified random samples maximize available moderator variance and thus make statistical tests more sensitive.

can be calculated (e.g., Marcus-Newhall et al., 2000). In addition, the validity of judges' ratings of manipulation effectiveness can be estimated by comparing them with effect sizes representing the manipulation checks present in the studies (e.g., Bettencourt & Miller, 1996; Miller, Lee, & Carlson, 1991).

RELIABILITY OF CODING. Given the importance to meta-analyses of accurate coding of the included studies, two or more individuals should perform the coding independently, followed by the calculation of an appropriate index of interrater reliability (such as the intraclass correlation or Cohen's, 1960, *kappa*; Orwin & Vevea, 2009). In most cases, disagreements can be resolved by discussion, or perhaps by averaging. Given that coding can be extremely time consuming, an alternative is to conduct dual coding on only a subset of studies, and if reliability is high, do only single coding on the remaining studies (Card, 2012). However, random sampling should determine which studies enter the initial sample of studies to be double-coded. Then, once reliability is established, studies should be chosen at random for double-coding (and included in the final reliability calculations). The better procedure, if feasible, is to double-code all studies.

CULTURAL AND SOCIAL STRUCTURAL CHARACTERISTICS. Although meta-analysts rely mainly on information in the source reports, they often incorporate information available elsewhere. Such information ranges from physical dimensions of social milieus to descriptions of social collectives such as organizations, communities, and nations. For example, Mullen and Felleman (1989) learned what specific dormitories had been studied in studies of crowding and then obtained from college administrators blueprints that allowed them to gauge physical features that were relevant to crowding effects. Similarly, Eagly, Johannesen-Schmidt, and van Engen's (2001) synthesis of sex differences and similarities of leadership styles obtained data from the U.S. Bureau of Labor Statistics and other sources to estimate the distribution of the sexes in studies' leadership roles when that information was missing from the reports.

Many additional databases relevant to social and personality phenomena track trends over decades or even centuries. For example, Gapminder (2012) tracks nation-level indicators on hundreds of dimensions (e.g., economic and health statistics). Among the databases that make U. S. survey data available are the American National Election Studies (2012) and the General Social Survey (2012). Many other nations and collectives (e.g., International Social Survey Programme, 2012) conduct similar opinion surveys. Hofstede's (2001) and others' surveys on cultural dimensions such as individualism, uncertainty avoidance, and masculinity are available for many nations (Taras, Kirkman, & Steel, 2010). The Cingranelli-Richards Human Rights Project (2012) gauges government respect for human rights across most nations. The World Values Survey (2012) compiles political and sociocultural indicators for many nations. The United Nations Statistics Division (2012) offers economic and sociopolitical data, as do the International Labor Organization (2012) and the World Bank (2013).

Estimating Effect Sizes in Individual Studies

To be included in the meta-analysis, a study must report a quantitative test of the hypothesis under scrutiny. In theory, each study j provides an observed estimate, T_j, of the underlying population phenomenon, θ. Hence, an observed study result is not the "truth" but an estimate of it. In general, past meta-analyses in personality and social psychology have emphasized two-variable quantitative tests, such as how maternal employment relates to children's achievement (Goldberg, Prause, Lucas-Thompson, & Himsel, 2008). Other meta-analyses have used the arithmetic means of one or more variables as effect sizes – for example, how much well-being, burnout, and anxiety are present in particular nations (Fischer & Boer, 2011). This section considers two-variable effect size indexes, otherwise known as indexes of association, and the following section addresses arithmetic means.

EFFECT SIZE INDEXES OF ASSOCIATION. There are many effect size indexes that gauge associations between two variables, as Table 26.1 shows. The table indicates that the measurement features of the variables in question guide the choice of effect size and the particular effect size index. As a general principle, if two or more studies report any one of Table 26.1's effect size metrics, they can be meta-analyzed, although all results must be converted to a single metric.[3] In addition to an effect size index T_j for each

[3] Similarly, use of an unstandardized outcome as the effect size (e.g., unstandardized regression slope or unstandardized mean difference) requires that each study assessed the phenomenon using the same operations. For example, Kirsch et al. (2008) used the unstandardized difference in improvement in depression scores as T because every study in their meta-analysis used exactly the same measure of depression.

TABLE 26.1. Potential Two-Variable Effect Sizes Dependent on the Measurement Features of the Two Variables (adapted from Johnson & Boynton, 2008).

Nature of Second Variable	Nature of First Variable: Continuous	Ordinal	Categorical
Continuous	• Pearson correlation (r) • Standardized regression slopes (β) • Unstandardized regression slopes	• Biserial correlation (r_b)	• Standardized mean difference • Unstandardized mean difference • Point-biserial correlation (r_{pb})
Ordinal		• Spearman correlation (ρ or rho) • Tetrachoric correlation (r_{tet})	• Rank-biserial correlation
Categorical			• Phi coefficient (φ) • Odds ratio (OR) • Risk ratio (RR) • Risk difference (RD)

Note: (a) Whether a variable is "first" or "second" is arbitrary. (b) "Categorical" assumes two discrete categories (e.g., male vs. female or experimental vs. control group), but it is of course possible to have more than two categories. (c) Any continuous or ordinal variable(s) could artificially be placed in a coarser category. (d) Some forms of effect size have subtypes not listed here (e.g., standardized mean difference can gauge either the means of two independent groups or of two time points for a single group).

study j, the sampling error associated with each study's effects must be estimated or recorded because it is used in all analyses. In social and personality psychology, because a diversity of measures appears to be the rule, analysts have nearly always used standardized effect size indexes, especially the standardized mean difference and the correlation coefficient. These effect sizes yield a common metric for comparing studies' findings.

Table 26.2 provides equations for the most commonly used forms of the standardized mean difference, the product-moment correlation coefficient, and the logged odds ratio. The table also highlights the systematic biases of estimates of effect sizes that are typically corrected in analyses. In addition, this table notes changes in the naming conventions for standardized mean differences. For example, Hedges's d (line 2) also has been labeled g^* and Hedges's g. Hedges (1981) developed this particular index of T specifically to apply to between-groups comparisons at a single point in time, providing proofs and documentation pertaining to this type of comparison. (Other sources consider complexities such as adjusting baseline differences between groups or gauging their change over time; e.g., Becker, 1988; Table 26.2, line 3). Consequently, the term "Hedges's d" should be restricted to the comparison specified in Table 26.2's line 2. The same principle holds regarding the other indexes of T.

THE DIRECTION OF EFFECT SIZES GAUGING ASSOCIATIONS. No matter the type of T used in a meta-analysis, its direction must be maintained consistently across the included studies by making T positive or negative so that studies with opposite outcomes have opposing signs. Ordinarily, a positive sign is given to outcomes in the expected, hypothesized, or typical, direction for the meta-analysis as a whole, whereas the negative sign is given to outcomes that reverse this direction. Only a relation that is exactly null would have no sign, because a standardized mean difference effect size (or r) would be 0.00 (and the Odds Ratio would be 1.00).[4] Illustrating this practice is Kite and Whitley's (1996) meta-analysis of sex-related differences in attitudes toward homosexuals, in which the expected direction of the findings was that women would evaluate homosexuals more positively than do men. Therefore, the positive sign for effect sizes indicated that women's evaluations were more positive

[4] In parallel, if the odds ratio is T, one might define values greater than one as positive and those smaller than one as negative.

TABLE 26.2. Common Two-Variable Effect Size Equations, Inverse Variance, and Usage Notes

Number	Effect Size	Equation(s)	Inverse Variance	Terms	Classic Citation and Notes
Standardized Mean Difference					
	Two-group comparison				
1	Cohen's *d*	$d = \frac{M_A - M_B}{SD_P}$	Not formally defined	M_A = mean for group A; M_B = mean for group B; SD_p = pooled standard deviation	Cohen's (1969) *d* is often called "uncorrected effect size" or *g* to distinguish it from Hedge's (sample- size corrected) *d*.
2	Hedges's *d*	$d = J(m) \times$ Cohen's *d*, where $J(m) \approx 1 - \frac{3}{4m-1}$	$\frac{2(n_a+n_b)n_a \times n_b}{2(n_a+n_b)^2 + n_a n_b d^2}$	$m = n_A + n_B - 2$ n_a = sample size for group a; n_b = sample size for group b	Hedges's (1981) *d* is often termed "Hedges's *g*" and sometimes g^*, where the asterisk implies the sample-size correction, $J(m)$. In the inverse variance equation, *d* is Hedges's *d*.
	One-group temporal comparison				
3	Becker's *d*	$d = \frac{M_{Pre} - M_{Post}}{SD_{Pre}}$	$\frac{2N}{4(1-r)+d^2}$	M_{Pre} = pretest mean; M_{Post} = posttest mean; SD_{Pre} = pretest standard deviation; r = correlation between pretest and posttest; N = sample size	Becker (1988)
Correlation between two variables					
4	Pearson's product-moment *r*	$r = \frac{\sum_{i=1}^{N} z_{X_i} z_{Y_i}}{N}$	Not formally defined	z_{X_i} and z_{Y_i} = standardized forms of *X* and *Y* being related for each case *i*	Pearson (1895)
5	Correction to Pearson's *r*	$\tilde{G}_{(r)} \cong r + \frac{r(1-r^2)}{2(N-3)}$	Not formally defined		Rarely used because bias is small when $n > 20$.
6	Fisher's *r*-to-*z* transform	$z_r = \frac{1}{2}\log_e \frac{1+r}{1-r}$	$N - 3$	$\log_e$ = natural logarithm	Fisher (1921)
7	Fisher's *z*-to-*r* transform	$r = \frac{e^{(2Z_r)} - 1}{e^{(2Z_r)} + 1}$	Not formally defined	e = base of the natural logarithm	Fisher (1921)
Odds ratio					
8	Logged odds ratio (OR)	$LOR = \log_e\left(\frac{ab}{bc}\right)$	$\frac{abcd}{ab(c+d)+cd(a+b)}$	$\log_e$ = natural logarithm Observed cases in a 2 × 2 contingency table:	A. W. F. Edwards (1963), J. H. Edwards (1957)
9	Transform of logged odds ratio to OR	$OR = e^{LOR}$	Not formally defined	e = inverse natural logarithm function	Is used to convert the LOR back into its original units for purposes of display and interpretation.

	+	−
+	a	b
−	c	d

Note: The inverse variance is provided only for fixed-effects assumptions. For random-effects assumptions, see the text.

than men's, and the negative sign that men's evaluations were more positive than women's. Alternatively, when experimental groups are compared with control groups, differences in favor of the experimental group might be given a positive sign, and differences in favor of the control group given a negative sign. Finally, meta-analyses may examine omnibus *T*s, such as multiple *R*, which gauges the amount of variance explained in a dependent variable attributable to more than one predictor variable; such *T*s take only positive signs.

MULTIPLE REPORTS FROM INDIVIDUAL STUDIES. When a given study provides multiple reports of the relation of interest, the analyst must decide whether to average the effect sizes to represent the study with a single effect size estimate or to treat them as separate estimates. To preserve the independence of the effect sizes in a meta-analysis, each must come from a different study. That is, the participants whose data contribute to a given effect size must not contribute to any other effect sizes in the analysis.[5] Therefore, the analyst would ordinarily average multiple effect sizes calculated from a single study. Instead of or in addition to averaging, an analyst may wish to investigate whether the results of the studies varied depending on the different operations by which their dependent variables were defined. For this purpose, the preservation of the separate effect size estimates made within individual studies may enable subsequent analyses examining whether the operations produced differences in the effect sizes. For example, in a meta-analysis of sex differences in leaders' effectiveness, Eagly, Karau, and Makhijani (1995) analyzed effect sizes according to the identity of the raters who provided the effectiveness measure and the basic type of measure (e.g., objective vs. subjective). Although many individual studies contributed several effect sizes to these analyses, each study's effect sizes were subsequently aggregated into a single study-level effect size that was used in additional analyses that did satisfy the assumption that effect sizes are independent. Analyses using multiple effect sizes from single studies can be informative even though they violate the assumption of independence of the effect sizes and thus can make statistical tests more liberal than they ought to be.

When a study examined the focal relation within levels of another variable, effect sizes may be calculated within these levels as well as for the study as a whole. How seriously the use of such within-level effect sizes violates the independence assumption depends on whether these levels were created on a within-subjects or a between-subjects basis. If the same participants took part at all levels of the variable (i.e., a within-subjects variable), the effect sizes would be highly dependent. The effect sizes would also be dependent if one control group served as a comparison for more than one treatment group. Even if the participants at the different levels were not the same individuals, the effect sizes would be dependent because they came from the same study, which was carried out under conditions existing in a particular place at a particular point in time (Hedges, 1990). For example, effect sizes might be calculated separately for the male and female participants of studies to enable examination of sex-related differences in the relation (e.g., Koenig et al., 2011), even though these effect sizes would not be independent.

PRECISION OF REPORTED STATISTICAL INFORMATION. Reports may contain more than one form of statistical information that could be used to calculate a given effect size. Some of these should converge within rounding error. For example, *F*-tests or *t*-tests should produce the same *T* as do the means and standard deviations that underlie them. The analyst should compute the effect size from both such sources to make sure that the results agree. As long as the effect sizes are similar, they should be averaged. If the effect size estimates are dissimilar, there may be errors in the information reported or the analyst's calculations. Sometimes inspection of the report's quantitative information for its internal consistency suggests that one form of the information is more accurate.

Similarly, for many reasons, some source reports contain less than desirable amounts of information for estimating *T*s, especially when *T* is gauged as a standardized mean difference. Some routes to estimating effect sizes merely require a great deal of effort on the part of the analyst (e.g., reanalyzing raw data found in an appendix of a dissertation). In other instances, deriving an effect size may require the application of several nonroutine techniques in sequence. (We provide some of these strategies in the Appendix.) Each

[5] A more subtle form of nonindependence occurs when samples within particular studies are related, such as husbands in one sample and wives in another, or when single investigators contribute more than one study. Current convention offers no satisfactory solution to this problem except to conduct sensitivity analyses to determine whether including dependent cases affects statistical inferences (Greenhouse & Iyengar, 2009) or to conduct individual participant meta-analyses that can directly accommodate the dependencies (Stewart, Tierney, & Burdett, 2005).

meta-analysis poses statistical challenges that may call for novel solutions.

Meta-analysts should contact studies' authors, if possible, to acquire essential information that is not included in a report. In our experience, cordial invitations to authors have produced moderate success rates (e.g., 40%). Obtaining such information allows the report to be adequately represented; failing to obtain the needed information renders the meta-analysis less comprehensive and potentially less representative. Finally, a lack of statistical detail in reports does not necessarily reflect their authors' oversights, errors, or poor methods. Rather, omissions generally occur because the authors' goals differed from those pursued in a subsequent meta-analysis. For example, a small sex-of-employee effect on job performance might have warranted only a brief acknowledgement of its nonsignificance, but for a meta-analysis on this subject (e.g., Roth, Purvis, & Bobko, 2012), such findings are crucial.

DEALING WITH NONREPORTED RESULTS. Reports that describe the effect of interest merely as "nonsignificant" are highly problematic in meta-analysis (Bushman & Wang, 1996). It is common to represent such effects as though they are exactly null (e.g., $d = 0.00$), but such estimates are obviously crude. If the N in the study was small, its actual effect size could be quite large, yet not significant. Introducing such effect sizes into a meta-analysis as though they were null biases a mean effect size toward the null (Schmidt, 1996); when these studies actually have results in the opposite direction, then assuming a null value is also unsatisfactory. Especially if many such reports exist in a literature, it may be advisable to conduct analyses with and without these 0.00 values.

At the synthesis stage of a meta-analysis, one way to incorporate imprecisely reported results, including those described as nonsignificant, is to use so-called "vote-counting procedures" to summarize findings (Bushman & Wang, 1996; Darlington & Hayes, 2000). In these procedures, rather than using effect size estimates to represent the studies' outcomes, an analyst examines how many studies obtained a result in the hypothesized direction or how many obtained a significant result in this direction. Because the strategy relies only on findings' directions or significance levels, it allows an analyst to include even the imprecisely reported nonsignificant results. More formally, calculating what is sometimes called the "sign test" determines the exact p of the observed distribution of positive and negative outcomes (or one more extreme), given that the probability of obtaining a positive result is .5, according to the null hypothesis, which specifies that half of the results should be positive and half negative following the binomial distribution. This probability can be calculated by standard statistics packages or spreadsheet software. An analyst can also use the binomial distribution to calculate a p-value for obtaining the observed distribution of significant positive findings versus other findings (nonsignificant and reversed), given that the probability of obtaining a significant result in one tail of the distribution is .025, according to the null hypothesis and assuming .05 for two-tailed significance testing. The p-values associated with the proportion of the studies that have a positive direction or that produced a significant positive result can be used to estimate a mean effect size for a sample of studies. These estimated effect sizes can then be compared to the exact mean effect size based on the studies that permitted this calculation (Bushman & Wang, 1996). For example, Wood (1987) used these techniques to estimate the mean effect size for sex-related differences in group performance because many of the studies did not permit an effect size to be estimated. Of course, it is much better to calculate the mean effect size by averaging effect sizes from individual studies when the majority of studies permit this strategy.

RELIABILITY OF EFFECT SIZE CALCULATIONS. At least two analysts should compute effect sizes independently for each of the studies and then compare solutions and resolve discrepancies. Given the complexity of many research designs and the ambiguity of some research reports, errors of effect size estimation are not uncommon. Moreover, sometimes one analyst may discover an indirect route to computing an effect size that is missed by a second analyst. Calculations by two or more analysts minimize such errors and omissions (see the section on Reliability of Coding).

CORRECTING EFFECT SIZES FOR BIASED METHODS. In addition to correcting the raw g and r for their inherent bias as estimators of the population effect size (see prior subsection on Effect Size Indexes), analysts may correct for many other biases that accrue from the methods used in each study. For example, as the reliability of a measure increases (and its measurement error therefore decreases), its relations with other variables will also increase (Cronbach, 1990). Increased measurement error decreases a measure's ability to predict another variable. Corrections for measurement unreliability and other forms of error or bias allow estimation of the strength of a relation absent such artifacts. In their presentations of such corrections for independent and dependent variables, Hunter and Schmidt (2004) and their colleagues (e.g., Schmidt, Le, & Oh, 2009) explained how to implement corrections for measurement error, artificial dichotomization of a continuous variable, imperfect construct validity, and range restriction. In theory, correcting for such errors permits a more accurate estimation of the true population effect size.

These corrections are quite popular in industrial and organizational psychology (e.g., Chiaburu, Oh, Berry, Li, & Gardner, 2011). They have seldom been used in social psychological meta-analyses because in most research areas relatively few studies include the information that would be required to perform the corrections (e.g., reliability or validity statistics). Nevertheless, meta-analysts may perform such corrections in research literatures in which reliabilities and other relevant information are routinely provided.

When meta-analysts do implement these corrections, the resultant corrected mean effect size yields an idealized estimate of the magnitude of the population effect rather than an estimate of the relation that is reported in a typical study if the corrections were not implemented. Nonetheless, because the correction procedures assume that the different biases are uncorrelated, the bias-adjusted corrections can yield irrational effect sizes (e.g., correlations larger than 1.00; Rosenthal, 1991). Therefore, analysts should consider their goals when deciding whether to use such corrections. If the goal is to estimate the effect size that would exist if there were no contamination by artifacts of measurement, the corrections would be desirable. In contrast, if the goal is to show how large a relation is in practice, then the corrections would be less useful.[6]

Regardless of whether these corrections are implemented, various biases may enter into studies' effect sizes. Consider that effect size estimates are a ratio of signal to noise, like all inferential statistics. For example, in a between-groups design, the signal is the difference in means, and the noise is the pooled standard deviation. Methodological factors can influence the effect size through their impact on signal, noise, or both factors. If two identical studies are conducted and one controls for noise that the other study does not (e.g., by statistically controlling for an individual difference characteristic), the first study will have a smaller error term (standard deviation), and the effect size will be larger for the first than the second study. To minimize this type of variation in effect sizes, meta-analysts should equate as much as possible the comparisons that the studies yield, so that the effect sizes are not influenced by differing statistical operations. For example, one such recommendation is that in meta-analyses of experimentally manipulated effects, analysts return irrelevant individual difference factors to the error term if they were included in the analysis in only some of the included studies. Reconstituting the error term in this way would not be necessary if the variable in question were controlled in all of the studies in the review. Similarly, many contemporary statistics already invoke corrections. For example, causal models with a latent variable structure effectively correct for unreliability and invalidity. Consequently, including results from such studies along with studies without latent variable structures introduces methodological noise across a literature. One method to reduce this influence is introducing the Hunter-and-Schmidt bias corrections to studies that lack the corrections (Card, 2012).

Additional problems can arise from the inclusion of studies that used within-subjects designs. For example, a researcher might have implemented a within-subjects design that required each participant to judge two objects along the same dimension. Such multiple assessments can produce many complications, including carryover, priming, and contrast effects (Smith, Chapter 3 in this volume). In analyzing such data, researchers nearly always use a repeated-measures inferential statistic that removes within-subjects variation from the error term. Consequently, these tests are more statistically powerful than those produced by a comparable between-subjects design (Dunlap, Cortina, Vaslow, & Burke, 1996; Morris & DeShon, 2002). If the meta-analyst uses these within-subjects error terms to calculate effect sizes, it is likely that these effect sizes will be larger than those based on standard deviations pooled from the cells of the design (e.g., Kite & Johnson, 1988; for an exception, see Symons & Johnson, 1997). Some sources recommend not mixing effect sizes from these two types of designs in the same analysis (e.g., Lipsey & Wilson, 2001), but others suggest using type of design as a moderator variable (e.g., Card, 2012). A growing convention

[6] Because the corrections information may sometimes be correlated with moderator dimensions, it seems that the most defensible strategy is to use the corrections as moderators themselves so that model testing can incorporate both types of information simultaneously and thus determine which aspects uniquely explain variation in the effect sizes.

is to estimate within-subjects cases using a between-subjects approximation (Becker, 1988).

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

Using Arithmetic Means to Gauge a Quantity's Magnitude

In the last 15 years, some meta-analysts in personality and social psychology have conducted meta-analyses by analyzing arithmetic means from studies as their estimate of *T*. With such strategies, analysts examine how low or high a sample scored on a certain criterion and model these outcomes using information about the samples (e.g., gender, recruitment strategies) and their milieus (e.g., economic success of women). For example, Twenge and her colleagues have examined temporal trends in U.S. samples in terms of levels of such variables as anxiety (Twenge, 2000), depression (Twenge & Nolen-Hoeksema, 2002), psychopathology (Twenge, Gentile, DeWall, Ma, Lacefield, & Schurtz, 2010), and narcissism (Twenge, Konrath, Foster, Campbell, & Bushman, 2008). Noguchi, Albarracín, Durantini, & Glasman (2007) examined interventions' recruitment and retention rates as factors that might relate to risk for acquiring or transmitting human immunodeficiency virus (HIV). Fischer, Hanke, and Sibling (2012) examined how social dominance orientation varies across 27 nations.

STANDARDIZING ARITHMETIC MEANS ACROSS STUDIES. If every study in a research literature operationalized the criterion of interest in exactly the same fashion, then meta-analyses can proceed without converting it to any other dimension (Bond, Wiitala, & Richard, 2003; Johnson & Boynton, 2008; Lipsey & Wilson, 2001). Doing so might be particularly advantageous when the measure is well known – measures of intelligence are good examples – as readers' familiarity with the measure helps make results easier to understand. Another alternative is mathematically converting results obtained on one scale to be equivalent with another scale. A mean value obtained on a 1-to-5 scale can be converted to the equivalent on a 1-to-7 scale or whatever target scale an analyst wishes to use across the literature of studies. Indeed, an argument can be made to move all such arithmetic means to their equivalents on a 0-to-100 scale, where 0 implies the lowest possible score and 100 is the maximum possible score. Targeting primary-level research, Cohen, Cohen, Aiken, & West (1999) advocated just such a procedure to convert means into percent of maximum possible (POMP) scores:

$$M_{\text{POMP}} = \frac{M - \text{minimum possible score}}{\text{maximum possible score} - \text{minimum possible score}} \times 100, \tag{26.1}$$

where M is the observed mean. The advantage of the POMP procedure is that the transformed values now take a more immediately interpretable meaning – those close to 0 are low and those close to 100 are high, and 50 is the mid-point. Putting all observed *M*s in a literature on the POMP metric also serves the statistical purpose of putting the study results on a common metric. If effect sizes of association are the focus of the meta-analysis, now the POMP scores could serve as moderators of those *T*s. Lennon, Huedo-Medina, Gerwien, and Johnson (2012) provided an example of this moderator strategy, showing that HIV prevention interventions for women succeeded to a greater extent in samples for which depression (represented by POMP scores) was more marked.

Putting arithmetic means on the same metric also implies that they can plausibly be used as *T*s themselves. To date, this strategy has been relatively rare (for an example, see Fischer et al., 2012). In order to invoke this strategy, not only the arithmetic means must be put into POMP metric but also their accompanying standard deviations:

$$SD_{\text{POMP}} = \frac{SD}{\text{Maximum possible score} - \text{Minimum possible score}} \times 100. \tag{26.2}$$

As we explain in the next subsection, SD_{POMP} is needed to estimate the inverse variance that is used as a weight in analyses of *T*s.

Some cautions about POMP scores are in order. Converting study results to a common metric assumes that they can be scaled in this fashion. That is, values may not have the same meaning on every scale converted into a common metric (e.g., Rosenthal & Rosnow, 1991). Therefore, the same sample of individuals may exhibit varying levels on differing scales intended to measure the same feature. If enough studies have multiple measures, meta-analyses can quantitatively test this assumption by examining whether different scales yield different M_{POMP} values.

ARITHMETIC MEANS VERSUS STANDARDIZED MEAN DIFFERENCE EFFECT SIZES. The fact that meta-analytic procedures allow use of the arithmetic mean as *T* might present a difficult decision for analysts who examine literatures in which two or more groups are compared on a continuous outcome (see Table 26.1). Historically, meta-analyses have defaulted to the standardized mean difference as *T*, but they could instead analyze the arithmetic means for each group. As Johnson and Boynton (2008) described, results from arithmetic means can provide even more detailed information about a literature than do results from the standardized mean difference. As we have noted, the latter form of *T* describes a difference between two means, where the sign of the *T* denotes whether one group is higher or lower than the other. Moderation patterns related to the standardized mean difference can leave unclear which of the two groups is changing most over the values of the moderator or moderators. As an example, Johnson and Boynton (2008) showed how mean sample age related positively to gender differences in social dominance orientation: As sample ages increased, standardized mean differences grew smaller. Yet, men may have decreased their support of social dominance, or women may have increased it. Johnson and Boynton used the arithmetic means separately for samples of females and males to show that the trend across the studies on the standardized mean difference index was primarily attributable to changes in the female samples. This example illustrates the use of *both* methods to gauge studies' effects.

There are some important caveats to using arithmetic means as *T* in a meta-analysis. First, many factors can affect the levels that arithmetic means take. For example, how positive participants are toward the position advocated in a persuasion experiment might be related to such factors as positive or negative mood, gender, personality traits, related attitudes, and of course the experimental condition itself. A meta-analysis could treat the mean for each condition as though it is an independent study, and if gender is the focus, subdivide each condition's data. Although some factors could be coded and used as moderators, many factors would not be possible to control. In contrast, meta-analyses that treat study information as two-variable effect sizes (Table 26.1) effectively control for the "noise" of variables that are not the focus of the meta-analysis. A comparison between, say, males and females from the same study controls for every factor *except* gender (and its correlates). Second, no matter the scale used for standardization (including POMP), the inverse variance for the arithmetic mean, which is used for weighting in analyses, relies on each study's observed standard deviation (Lipsey & Wilson, 2001):

$$\textit{Inverse variance} = \frac{n}{SD^2}. \tag{26.3}$$

One problem with POMP scores is related to the zero or near-zero standard deviations that may appear under some circumstances. For example, when observed arithmetic means take the maximum or the minimum possible value, their standard deviations will be zero, which implies that a weight cannot be calculated. Such studies might need to be omitted from analyses or examined with alternative assumptions.

Analyzing the Meta-Analytic Database

PRELIMINARY CONSIDERATIONS. The general steps involved in the analysis of any effect size, *T*, usually are the following: (1) aggregate effect sizes across the studies to determine the overall magnitude of the weighted mean *T*; (2) analyze the consistency of the effect sizes across the studies; (3) diagnose statistical outliers among the effect sizes; (4) examine the distribution of effect sizes to determine whether any irregularities exist; and (5) perform tests of whether study attributes moderate the magnitude of the effect sizes.

MEAN EFFECT SIZE AND HOMOGENEITY OF EFFECT SIZES. The model-testing procedures that we present are analogous to techniques used in data analysis in primary research and take advantage of weighted general linear models, where the weights are defined as the inverse variance, as we will explain. Models that divide results for categorical features are known as *subgroup analyses* or *categorical models*, and those that use continuous features are known as *meta-regressions* (which may also include categorical variables). Statistical analyses in meta-analysis differ from those in primary research in two main respects. The first difference pertains to the heterogeneity of the variances ordinarily associated with the individual effect sizes, which would likely violate the homoscedasticity

assumption of conventional regressions and ANOVAs (Hedges & Olkin, 1985), which is that standard deviations of the error terms do not vary and do not depend on predictors' values. Because this nonsystematic variance of an effect size is in general inversely proportional to the sample size of the study and sample sizes vary widely across the studies, the error variances of the effect sizes are ordinarily quite heterogeneous. Meta-analytic statistics aim to overcome this limitation (see the next subsection). The second difference between the statistical procedures of meta-analysis and primary research is that meta-analytic statistics permit an analysis of the consistency (or homogeneity) of the effect sizes across the studies – a highly informative analysis.

As a first step in a quantitative synthesis, the study outcomes are combined by averaging the *T*-values with each T_j for each study j is weighted by the reciprocal of its variance. The weighted mean effect size T_+ is a weighted average of the individual studies' effect sizes,

$$T_+ = \frac{\sum_{j=1}^{k} w_j T_j}{\sum_{j=1}^{k} w_j}, \tag{26.4}$$

where k is the number of effect sizes and w_j is the weight for each study j. The weights may be defined as a simple function of the sampling error associated with each effect size j, which follows *fixed-effects assumptions*. In this case, the inverse variance for each T serves as the weight (see examples in Table 26.2). Alternatively, analysts can define the weights to incorporate an estimate of the variance in the population of effect sizes, τ^2 (Hedges & Vevea, 1998), which follows *random-effects assumptions*. In either version of weighting, Equation 26.4 gives greater weight to the more reliably estimated study outcomes.

Cochran's (1954) Q evaluates the hypothesis that the effect sizes are homogeneous. Specifically, Q is a model specification statistic that evaluates how closely individual T_j correspond with T_+,

$$Q = \sum_{j=1}^{k} W_j (T_j - T_+)^2, \tag{26.5}$$

where k is the number of effect sizes in the class and W_j is based on fixed-effects assumptions (see examples in Table 26.2).[7] Q has an approximate χ^2 distribution with $k - 1$ degrees of freedom. If Q is significant, the hypothesis of the homogeneity (or consistency) of the effect sizes is rejected, and heterogeneity is inferred. In other words, there is more variability in the observed *T*s than would be expected on the basis of the sampling error alone. In this event, the weighted mean effect size may not adequately describe the outcomes of the set of studies because it is likely that quite different mean effects exist in different groups of studies, and these differences may include differences in the direction (or sign) of the relation. In some subgroups of studies, *X* might have had a large positive effect on *Y*, and in other studies it might have had a smaller positive effect or even a negative effect on *Y*.

Values of Q are highly correlated with the numbers of *T*s entering into this statistic, making it difficult to compare levels of heterogeneity between meta-analyses and within portions of meta-analysis. To address this issue, Higgins and Thompson (2002) introduced a homogeneity index, I^2, based on Q and its degrees of freedom. Values of I^2 range from 0 to 100%, where high values indicate more variability among the effect sizes and 0 implies homogeneity. Yet, I^2 is subject to the same conditions and qualifications as is Q (Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006) such that both statistics are underpowered in small samples of studies. Moreover, values of I^2 at 25%, 50%, and 75% are often taken to describe small, moderate, and large amounts of heterogeneity, respectively. Yet, these cut points are best taken only as suggestions: Even a "small" I^2 can hide statistically significant variability in *T*s.

Even if a homogeneity test is nonsignificant, significant moderators could be present, especially when Q or I^2 are relatively large (Johnson & Turco, 1992). Also, Q and I^2 can be significant even though the effect sizes are very close in value, especially if the sample sizes are very large. Therefore, heterogeneity deserves careful interpretation, in conjunction with inspecting the values of the effect sizes. Nonetheless, in a meta-analysis that attempts to determine *X*'s impact on *Y*, rejecting the hypothesis of homogeneity could be troublesome because it implies that the association between these two variables likely is complicated by the presence of interacting conditions. Because analysts usually anticipate the presence of one or more moderators of effect-size magnitude, establishing that, overall, effect sizes lack homogeneity is ordinarily of no concern, unless analysts cannot determine the sources of the heterogeneity.

The fact that *T*s may differ widely in magnitude should give analysts pause about the meaning of a weighted mean effect size, T_+. In the face of heterogeneity, T_+ may lack a clear meaning, even if it is evaluated with random-effects assumptions, which are relatively conservative compared to fixed-effects assumptions. That is, incorporating random-effects assumptions will yield wider confidence intervals around T_+ than will those based on fixed-effects assumptions. Thus, a random-effects mean may disguise meaningful subpopulations of *T*s.

In practice, the fixed- and random-effects variance components are summed to form new weights:

$$W_j = \frac{1}{Variance_{FE} + \tau^2}$$

where $Variance_{FE}$ is the fixed-effects variance for each study and τ^2 is a constant for each study. The standard deviation of the population of effect sizes, τ, takes the same metric as *T*, and τ^2 is in the same metric as T^2 (for calculations, see Borenstein, Hedges, Higgins, & Rothstein, 2009). Using these weights in Equation 26.5 produces a mean based on random-effects assumptions. In the unlikely event that $\tau^2 = 0$, random-effects assumptions reduce to fixed-effects assumptions.

The variance, v_+, of the weighted mean effect size T_+ is

$$v_+ = \frac{1}{\sum_{j=1}^{k} W_j}. \tag{26.6}$$

As a test for significance of this weighted mean effect size, one can calculate a confidence interval around this mean, based on its standard deviation, $T_+ \pm 1.96\sqrt{v+}$ where 1.96 is the unit-normal value for a 95% CI (assuming a nondirectional hypothesis). If the confidence interval (*CI*) includes zero (0.00), the value indicating exactly no difference, it may be concluded that, aggregated across all studies, there is no significant association between the independent and dependent variable (*X* and *Y*). The fixed-effects mean is known to be overpowered in the face of heterogeneity (Hedges & Vevea, 1998; Huedo-Medina, Sánchez Meca, & Marín Martínez, 2004). In other words, when study results are inconsistent, a fixed-effects mean is more likely to reach statistical significance than is a random-effects mean, other factors being equal. Thus, assuming fixed-effects assumptions should be considered a relatively risky strategy of statistical inference.

Finally, analysts often present other measures of central tendency in addition to the weighted mean effect size (Borenstein et al., 2009). For example, the unweighted mean effect size shows the typical effect without weighting studies with larger sample sizes more heavily. A substantial difference in the values of the unweighted and weighted mean effect sizes suggests that one or more studies with large sample sizes may deviate from the rest of the sample. It is possible that larger studies used different methods than smaller studies did. Also, the median effect size describes a typical effect size but would be less affected than a mean effect size by outliers and other anomalies of the distribution of effect sizes.

EVALUATING THE POTENTIAL FOR PUBLICATION BIAS. Asymmetries in the distribution of effect sizes often are taken as evidence of publication bias, that is, the possibility that published results differ systematically from those that are not published (Sutton, 2009). *Funnel plots* (Light & Pillemer, 1984) are scatter plots of inverse variances versus effect sizes. When there is no publication bias, the scatterplot should take the shape of a funnel sitting on end in the sense that the effect sizes from smaller studies, which are less reliable, would show more scatter than the effect sizes from the larger studies, which would center on the best estimate of the population effect. Yet, if there is a publication bias in the literature, a funnel plot should reveal few entries in the smaller effect size portion of the graph for smaller sample sizes. There are many variations on such displays that are often quite sophisticated (Borman & Grigg, 2009). The most popular quantitative alternatives to examine for asymmetries include Egger, Smith, Schneider, and Minder's (1997) and Begg's (1985) tests, which provide estimates of the extent to which asymmetry is present in a distribution of effect sizes. Another popular tool is the trim-and-fill technique (Duval & Tweedie, 2000), which quantitatively assesses whether such asymmetries would change inferences about the significance of T_+. An important caveat to all of these strategies is that each assumes a single population of effect sizes. Under heterogeneity, the tests may not be diagnostic of publication bias (e.g., Lau, Ioannidis, Terrin, Schmid, & Olkin, 2006; Sutton, 2009).

Analysts sometimes calculate the number of studies averaging a null effect that would be necessary to bring an overall meta-analytic mean to the point of nonsignificance (Rosenthal, 1979). If this "fail-safe *N*" (N_{fs}) is small, then the result seems less trustworthy. Specifically, one would calculate

$$N_{fs} = \frac{\left(\sum_{j=1}^{k} Z_j\right)^2}{z_\alpha^2}, \tag{26.7}$$

where k is the number of studies, Z_j is the unit normal value corresponding to a one-tailed test of significance, and z_α is the critical value (i.e., 1.645 for a one-tailed hypothesis). Orwin (1983) offered a variant of

[7] Q and I^2 may also be defined using random-effects assumptions.

this equation that estimates N_{fs} directly from the mean weighted effect size. Although N_{fs} may have heuristic value in some instances, the equation for N_{fs} assumes that unretrieved studies would average null when in fact they may have the same pattern as the retrieved studies or even a reversed pattern. Also, it is difficult to evaluate the magnitude of N_{fs} because it has no statistical distribution theory (Becker, 2005).

TESTING MODELS OF META-ANALYTIC MODERATORS. To determine the relation between study characteristics and the magnitude of the effect sizes, analysts fit models using a form of weighted ordinary least squares regressions (for statistical methods, see Borenstein et al., 2009; Harbord & Higgins, 2008; Hedges & Olkin, 1985; Higgins & Thompson, 2004; Huedo-Medina & Johnson, 2010). Moderators, which are also called *effect modifiers*, can take the form of either categorical or continuous dimensions; they can be entered either solely (bivariate) or in a combined form. For example, in a continuous model, Hart, Albarracín, Eagly, Brechan, Lindberg, & Merrill (2009) found that, to the extent that information was more congenial, greater selective exposure resulted. Similarly, in a categorical model (also called subgroup analysis) they found that individuals preferred congenial over uncongenial information, especially when the issue was of high versus low value-relevance.

As noted, categorical and continuous features may be evaluated in meta-regression procedures, dummy-coding categorical variables as necessary. The unstandardized regression *(b)* coefficient(s) provide tests for the significance of the predictor's association with the effect sizes. Under fixed-effects assumptions, the models use the inverse *variance* for each effect size as the weights. Such models are known to be overpowered in the face of heterogeneity (Hedges & Vevea, 1998). Under fixed-effects assumptions, the fit of meta-regression models is estimated by the error sum of squares statistic, Q_E, which has an approximate chi-square distribution with $k - p - 1$ degrees of freedom, where k is the number of effect sizes and p is the number of predictors (not including the intercept). Q_E can be converted to I^2 for evaluation.

Contemporary software permits easy incorporation of random-effects assumptions in such models. Such models are ordinarily *mixed-effects models* because differences between groups of *T*s (i.e., the slopes) are fixed and the constant (or intercept) follows random-effects assumptions (e.g., Harbord & Higgins, 2008). By convention, most analysts label these models *random-effects meta-regressions*, and this set of assumptions has become the most conventional for most meta-analytic situations. These models estimate the population variance, τ^2, after removing the variance attributable to the moderators included in the model. Thus τ^2 can and does change from model to model. Commonly available output in these models includes I^2 residual, which is an assessment of the between-studies variability that is not explained by the model.

OUTLIER DIAGNOSES. Because meta-analyses weight studies for their inverse variance, outliers with larger weights can dramatically alter meta-regression results (for a more general discussion of the topic of data outliers, see McClelland, Chapter 23 in this volume). Under such circumstances, these outliers can be removed from subsequent phases of the data analysis. Alternatively, *T*s that are far distant from other *T*s can be winsorized so that they are not so extreme. The same can be done for inverse variance estimates that are relatively extreme. Outliers might be detected in many ways, but one that is highly recommended is to examine the residuals in meta-regression models.

DEPICTIONS OF EFFECT SIZE MAGNITUDE. In some instances, visual presentations can assist greatly in the interpretation of meta-analytic results (Borman & Grigg, 2009; Johnson & Huedo-Medina, 2011). For example, visually examining study outcomes enhances the analyst's potential for finding anomalies in the meta-analytic data. By examining how effect sizes vary over the range of a moderator, an analyst may determine that effect sizes are related to a continuous predictor in a nonmonotonic fashion – an outcome that would not be detected by the linear regressions that have been described to this point in the chapter. Meta-regression models may include tests of nonlinear associations, yet unless nonmonotonic associations are expected on an a priori basis, they are unlikely to be discovered except by the use of visual displays.

Depictions of model results in either graphical or tabled form can help describe results in presentations and written reports. Johnson and Huedo-Medina (2011) described *the moving constant technique*, with which analysts can use meta-regression to create graphs of effect sizes plotted against moderator values, including confidence bands around the meta-regression line. This technique can also be used to estimate mean effect size values and confidence intervals at moderator values of interest. Specifically, analysts may move the intercept to reflect interesting points

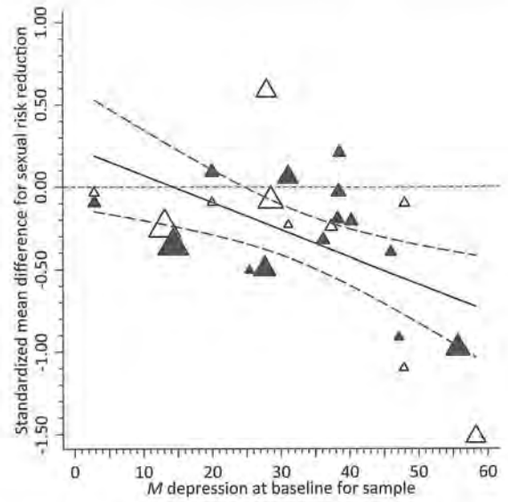

Figure 26.1. Sexual risk reduction following a behavioral intervention as a function of each sample's baseline depression. Sexual risk behavior declined following the intervention at the last available follow-up to the extent that samples had higher levels of baseline depression (treatment [control] group effects appear as darker [white] triangles and the size of each plotted value reflects its weight in the analysis). The solid regression line indicates trends across initial levels of depression; dashed lines provide 95% confidence bands for these trends. Reproduced from Lennon et al. (2012).

along or beyond a range of independent variable values. For example, Lennon et al. (2012) found that HIV prevention efforts for women succeeded better for samples with higher baseline depression. Using the moving constant technique, they estimated the amount of risk reduction for samples with the highest mean levels of depression to be large and significant, whereas for samples with lower levels of depression, on average, interventions failed to impact risk (see Figure 26.1). Results presented in this form help show for what levels of a moderator an effect exists. Such estimates, in turn, can be highly informative when interpreting the nature of the phenomenon being studied in the meta-analysis, especially when a comparison to an absolute or a practical criterion is important. The moving constant technique also permits analysts to estimate confidence intervals for an effect size at particular values of one or more independent variables (and thus to avoid artificially dichotomizing continuous predictor variables).

DEALING WITH NONINDEPENDENT EFFECT SIZES. We have indicated that, as a general rule, it is wise to represent studies' participants only once in effect size calculations. Thus, analysts should ordinarily combine effect sizes representing conceptually similar measures from any given study. If such effect sizes were not combined, the nonindependence that would result could have several effects on the findings of a meta-analysis, depending on the source of the nonindependence (Gleser & Olkin, 2009). If the nonindependence results from producing more than one effect size from the same participants on correlated measures, the meta-analysis will be likely to reach a liberal estimate of the significance of the weighted mean effect size: Its *CI* will grow tighter. Including more effect sizes from the same groups of participants may also affect inferences from model-fit statistics (Q or I^2).

Despite these concerns, representing studies multiple times may be defensible to address certain meta-analytic questions. One such question is whether an effect generalizes across various types of measures of a dependent variable. In such a case, the analyst could examine a model to determine if the effect sizes differed according to the type of measure used. If the synthesis forgoes this analysis to uphold the assumption that effect sizes are independent, potentially valuable information about a moderator would be lost. Therefore, one defensible strategy is to conduct a two-stage meta-analysis that shifts its units of analysis (Cooper, 2010). In the first stage, the meta-analysis would address the study-level effect sizes, which represent the information from each study only once. A second stage would divide study outcomes into the various groupings specified by moderators and would permit information for a group of study participants to appear more than once, in order to examine the differences across the moderator (for examples of this strategy, see Gerrard, Gibbons, & Bushman, 1996; Kolodziej & Johnson, 1996). This ordering of the stages enables analysts to learn the overall, more general pattern in the literature prior to answering specific questions about moderators. This combination of approaches should help allay concerns about nonindependence while still yielding the desired information. Other alternatives include (a) using multivariate procedures for the analysis of multiple effect sizes from each study (Gleser & Olkin, 2009); (b) representing effect sizes nested within studies in terms of multilevel models (Hedges, 2009); or (c) pursuing individual-level meta-analyses of studies whose raw data are available, in a practice also known as integrated data analysis (Cooper & Patall, 2009; Stewart,

TABLE 26.3. Cohen's (1969) Guidelines for Magnitude of *d* and *r*

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

Tierney, & Burdett, 2005). This latter option is often considered the gold standard of meta-analysis when the individual-level studies reviewed are highly representative of the often much larger literatures for which only study-level effects are available.

INTERPRETATIONS OF EFFECT SIZE INDEXES OF ASSOCIATION. Cohen (1969, 1988) tentatively proposed some guidelines for judging effect magnitude, based on his informal analysis of the magnitude of effects commonly yielded by psychological research. Cohen intended "that medium represents an effect of a size likely to be visible to the naked eye of a careful observer" (Cohen, 1992, p. 156). He intended that small effect sizes be "noticeably smaller yet not trivial," and that large effect sizes "be the same distance above medium as small is below it" (p. 156). As Table 26.3 shows, a "medium" effect turned out to be about $d = 0.50$ and $r = .30$, equivalent to the difference in intelligence scores between clerical and semiskilled workers. A "small" effect size was about $d = 0.20$ and $r = .10$, equivalent to the difference in height between 15- and 16-year-old girls. Finally, a large effect was about $d = 0.80$ and $r = .50$, equivalent to the difference in intelligence scores between college professors and college freshmen. Although these impressionistic guidelines for magnitude of effects are frequently cited, there are caveats about particular effect size indexes' magnitude (McGrath & Meyer, 2006). Many alternatives exist for interpreting the magnitude of effects.

One popular way to interpret mean effect sizes is to derive the equivalent r and square it. This procedure shows how much variability would be explained by an effect of the magnitude of the mean effect size (see Table 26.3). Thus, a mean d of 0.50 produces an R^2 of .09. However, this value must be interpreted carefully because R^2, or variance explained, is a directionless effect size. Therefore, if the individual effect sizes that produced the mean effect size varied in their signs (i.e., the effect sizes were not all negative or all positive), the variance in Y explained by the predictor X, calculated for each study and averaged, would be larger than this simple transformation of the mean effect size.

A number of methodologists have argued that even quantitatively small effects can be quite consequential (e.g., Abelson, 1985; Prentice & Miller, 1992; Rosenthal, 1990; Ross & Nisbett, 1991), and some have provided tools to help show how meaningful an implied effect size is in application. These tools include Rosenthal and Rubin's (1982) binomial effect size display (for caveats, see Thompson & Schumacker, 1997), McGraw and Wong's (1992) common language effect size statistic index, and Rosenthal and Rubin's (1994) counternull statistic. In using such tools, the meta-analyst attempts to reach some conclusion about how much the effect matters in terms of some tangible outcome.

Another method of interpreting the magnitude of effect sizes is to compare them with effect sizes in similar domains in which magnitude is already known. For example, Eagly (1995) argued that claims that sex-related differences in behavior are necessarily small should be evaluated in relation to the magnitude of other known effects in psychology. Following this strategy, Bettencourt and Miller (1996) compared the magnitude of sex-related differences in aggression to the magnitude of the effect of provocation on aggression, which was derived from the same sample of studies. More generally, meta-analysts ought to compare the magnitude of a newly derived meta-analytic effect size to the magnitude of known effects in the same or related research areas. It is also important to consider the implications of effect sizes in metrics that are sensible in natural settings (e.g., number of lives saved by treatments, proportions of girls and boys admitted to selective educational programs, given a particular ability sex difference).

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

CONDUCTING AND EVALUATING META-ANALYSES

Our treatment of meta-analytic methods has stressed the importance of high standards in conducting and evaluating these reviews. From the preceding sections of this chapter, a picture of a high-quality meta-analysis emerges:

1. Define the research problem clearly and, if possible, define hypotheses prior to commencing with the meta-analysis.
2. Use highly inclusive search strategies that locate unpublished as well as published studies.
3. Be explicit in the criteria for selecting studies and, if possible, define these a priori.
4. Thoroughly and accurately code moderator variables and other study-relevant information.
5. Represent study outcomes with high accuracy.
6. Conduct meta-analytic models, maintaining fidelity to the statistics' assumptions.
7. Interpret findings carefully in relation to the assumptions that underlie both individual studies and the meta-analysis itself.

Each of these dimensions appears in Shea et al.'s (2007) recent quality-coding protocol for meta-analysis. Nonetheless, even a quantitative review that meets high standards does not necessarily constitute an important scientific contribution.

One factor affecting the scientific contribution of a synthesis is that its conclusions are limited by the quality of the data that are synthesized. Serious methodological faults that are endemic in a research literature may well handicap a synthesis, unless it is designed to shed light on the influence of these faults. Also, to be regarded as important, the review must address an interesting question. Similarly, unless the paper reporting a meta-analysis "tells a good story," its full value may go unappreciated by readers. Although there are many paths to a good story, Sternberg's (1991) recommendations to authors of reviews are instructive: pick interesting questions, challenge conventional understandings if at all possible, take a unified perspective on the phenomenon, offer a clear take-home message, and write well.

Some reports of research syntheses may fail to tell a good story because they are overly complex. This complexity may arise from the fact that quantitative synthesis forces the reviewer to study the minute details of the studies' methods and findings. Although this close scrutiny can yield valuable insights, it may also foster a review that reflects too many complexities and thereby obscures its major findings. In short, even if a synthesis happens to solve a time-honored problem, it will have a poor reception if its message is mired in a forest of distracting minutiae. Excellent organization and skillful writing can overcome this challenge.

Although many critiques of meta-analyses have taken a narrative form by discussing their methods and findings, the most informative critiques take a quantitative approach by empirically evaluating the findings and conclusions. A critique that may seem reasonable based on sheer logic may become overwhelming when supported by appropriate data. In this manner, scientific disputes can be arbitrated by empirical tests. In primary research, the most influential critiques take the form of replications with variations, often showing how an effect disappears once a confound is controlled. Similarly, criticism of quantitative syntheses proceeds most effectively in an empirical fashion. In our view, replications of meta-analytic reviews should become more frequent, so that faults that may be present in one review are evaluated or eliminated in later reviews.

With meta-analyses having become commonplace, investigators should anticipate the recycling of their findings in meta-analyses. They should therefore redouble their efforts to report the method and results of their studies as accurately and completely as possible, aided by supplements and archives. Researchers can find excellent guidance in the Journal Article Reporting Standards (JARS) presented in the *Publication Manual* of the American Psychological Association (2010). In particular, for experimental studies, a table of means and standard deviations for each primary dependent variable, reported for all cells of the design, should be conventional. It is very helpful if exact statistics are provided even for auxiliary effects that may be nonsignificant (e.g., the comparison of female and male participants). For correlational studies, a complete matrix of the variables' intercorrelations should be conventional.

ADDITIONAL RESOURCES ON RESEARCH SYNTHESIS

Hunt (1997) provides a compelling and highly readable history on research synthesis. Essential reference works for conducting meta-analyses are *The Handbook of Research Synthesis and Meta-Analysis*, edited by Cooper, Hedges, and Valentine (2009), as well as

texts by Borenstein et al. (2009), Card (2012), Cooper (2010), Hedges and Olkin (1985), and Lipsey and Wilson (2001). Two of these offer either commercial software (Borenstein et al., 2009) or open-access macros for popular statistical platforms (Lipsey & Wilson, 2001). Viechtbauer (2010) authored a flexible and powerful set of tools for the open-source statistics software package, R. Other works may be particularly valuable for other aspects of meta-analysis: Hunter and Schmidt (2004) extensively addressed corrections to effect sizes; Glass et al.'s (1981) book remains a good source on derivations of effect sizes.

THE FUTURE OF META-ANALYSIS IN SOCIAL AND PERSONALITY PSYCHOLOGY

The growing numbers of studies on personality and social psychology's central phenomena dictate that, in the future, greater importance will be accorded to high-quality meta-analyses of these knowledge bases. In our opinion, the quality of meta-analyses has improved over the past decades. Meta-analysis should foster a healthy interaction between primary research and research synthesis, at once summarizing old research and suggesting promising directions for new research. One misperception that psychologists sometimes express is that a meta-analysis represents a point beyond which nothing more needs to be known. On the contrary, carefully conducted meta-analyses can often be the best medicine for a literature, by documenting the robustness with which certain associations are attained, resulting in a sturdier foundation on which future theories may rest. In addition, meta-analyses can show where knowledge is at its thinnest, to help plan additional, primary-level research (Wood & Eagly, 2009). As a consequence of a carefully conducted meta-analysis, new studies can be designed with the complete existing literature in mind and therefore have a better chance of contributing new knowledge. In this fashion, scientific resources can be directed more efficiently toward gains in knowledge.

The advent of computerized and readily accessible databases of psychological research literatures (e.g., PsycINFO) has meant that less time and financial resources are necessary to conduct meta-analyses than in the past. Despite these gains, psychologists face severe limitations in obtaining access to the data underlying completed research. In contrast to some other scientific fields (e.g., sociology, political science), few raw data from primary research are archived in psychology, and this omission greatly limits the opportunity for reviewers to perform the secondary analyses that can produce effect sizes for phenomena that have not been adequately reported. Primary researchers are often unable or unwilling to provide needed statistical information when they are contacted directly. Routine data archiving in a central location would remedy this unfortunate situation (Cooper et al., 1997).

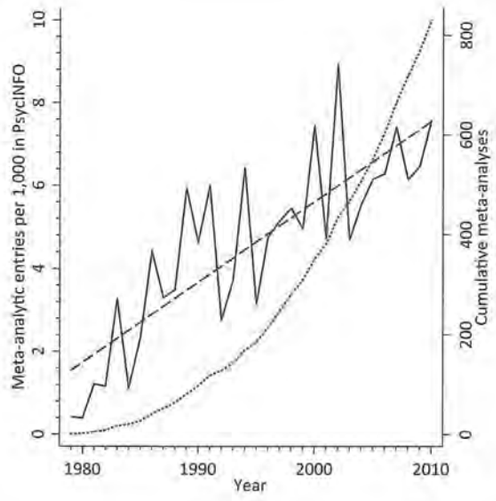

Figure 26.2. Publication trends in meta-analyses in social and personality psychology, where the solid line plots the number of reports per year per 1,000 recorded in PsycINFO; the dashed line is the best-fitting linear trend (both on the left axis), and the dotted line represents cumulative meta-analytic reports (right axis).

Psychologists and other scientists rely more and more on meta-analyses to inform them about the knowledge that has accumulated in their research. Although meta-analysis might become the purview of an elite class of researchers who specialize in research integration, as Schmidt (1992) argued, we believe that, on the contrary, meta-analysis will become a routine part of graduate training in many fields. With computer programs to aid calculations, most researchers should be able to integrate findings across studies as a normal and routine part of their research activities. Indeed, the publication trends[8] within social and personality psychology that we portray in Figure 26.2 suggest that this phenomenon is occurring. Meta-analysis has become central to these areas of research and to many others.

[8] This PsycINFO search was performed on April 2, 2012, with "meta-analysis" in title, abstract, or keywords; AND Content Classification Code = social psychology, personality psychology, personality scales and inventories, political processes and political issues, or sex roles and women's issues; AND Document type = journal article, chapter, or dissertation.

REFERENCES

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

Albarracín, D., Johnson, B. T., Fishbein, M., & Muellerleile, P. A. (2001). Theories of reasoned action and planned behavior as models of condom use: A meta-analysis. *Psychological Bulletin, 127*(1), 142–161.

American National Election Studies. (2012). Databases related to political trends. Retrieved from http://www.electionstudies.org/

American Psychological Association. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63*(9), 839–851.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Baas, M., De Dreu, C. K. W., & Nijstad, B. A. (2008). A meta-analysis of 25 years of mood-creativity research: Hedonic tone, activation, or regulatory focus? *Psychological Bulletin, 134*(6), 779–806.

Balliet, D., Li, N. P., Macfarlan, S. J., & Van Vugt, M. (2011). Sex differences in cooperation: A meta-analytic review of social dilemmas. *Psychological Bulletin, 137*(6), 881–909.

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

Becker, B. J. (2005). Failsafe N or file-drawer number. In H. Rothstein, A. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 111–126). Chichester, UK: Wiley.

Begg, C. B. (1985). A measure to aid in the interpretation of published clinical trials. *Statistics in Medicine, 4*, 1–9.

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

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

Bond, C. F., Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods, 8*, 406–418.

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

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

Borman, G. D., & Grigg, J. A. (2009). Visual and narrative interpretation. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 497–519). New York: Russell Sage Foundation.

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

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

Cafri, G., Kromrey, J. D., & Brannick, M. T. (2009). A SAS macro for statistical power calculations in meta-analysis. *Behavior Research Methods, 41*(1), 35–46.

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

Card, N. A. (2012). *Applied meta-analysis for social science research*. New York: Guilford Press.

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

Chiaburu, D. S., Oh, I. S., Berry, C. M., Li, N., & Gardner, R. G. (2011). The five-factor model of personality traits and organizational citizenship behaviors: A meta-analysis. *Journal of Applied Psychology, 96*(6), 1140–1166.

Cingranelli-Richards Human Rights Project. (2012). Human rights data across nations. Retrieved April 6, 2012, from http://www.humanrightsdata.org/

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10*(1), 101–129.

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

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

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

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

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstances for POMP. *Multivariate Behavioral Research, 34*, 315–346.

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

Cooper, H. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Los Angeles: Sage.

Cooper, H., DeNeve, K., & Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods, 2*(4), 447–452.

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

Cooper, H., Hedges, L., & Valentine, J. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage Foundation.

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods, 14,* 165–176.

Cooper, H., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin, 87,* 442–149.

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

Darlington, R. B., & Hayes, A. F. (2000). Combining independent *p* values: Extensions of the Stouffer and binomial methods. *Psychological Methods, 5*(4), 496–515.

Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitch, C., Song, F., Petticrew, M. & Altman, D. J. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment, 7*(27), 1–179.

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

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

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56,* 455–463.

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

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

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

Eagly, A. H., Johannesen-Schmidt, M. C., & van Engen, M. L. (2003). Transformational, transactional, and laissez-faire leadership styles: A meta-analysis comparing women and men. *Psychological Bulletin, 129*(4), 569–591.

Eagly, A. H., Karau, S. J., & Makhijani, M. G. (1995). Gender and the effectiveness of leaders: A meta-analysis. *Psychological Bulletin, 117*(1), 125–145.

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

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

Edwards, A. W. F. (1963). The measure of association in a 2 × 2 table. *Journal of the Royal Statistical Society: Series A, 126,* 109–114.

Edwards, J. H. (1957). A note on the practical interpretation of 2 × 2 tables. *British Journal of Preventive & Social Medicine, 11,* 73–78.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple graphical test. *British Medical Journal, 315,* 629–634.

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

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

Fischer, R., & Boer, D. (2011). What is more important for national well-being: Money or autonomy? A meta-analysis of well-being, burnout, and anxiety across 63 societies. *Journal of Personality and Social Psychology, 101*(1), 164–184.

Fischer, R., Hanke, K., & Sibley, C. G. (2012). Cultural and institutional determinants of Social Dominance Orientation: A cross-cultural meta-analysis of 27 societies. *Political Psychology, 33,* 437–467.

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

Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review, 19,* 151–156.

Freund, P. A., & Kasten, N. (2012). How smart do you think you are? A meta-analysis on the validity of self-estimates of cognitive ability. *Psychological Bulletin, 138*(2), 296–321.

Gapminder (2012). Social, political, and health databases. Retrieved April 6, 2012, from http://www.gapminder.org/

General Social Survey (2012). Survey data regarding the U.S. population. Retrieved April 6, 2012, from http://www3.norc.org/gss/

Gerrard, M., Gibbons, F. X., & Bushman, B. J. (1996). Relation between perceived vulnerability to HIV and precautionary sexual behavior. *Psychological Bulletin, 119*(3), 390–409.

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

Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 357–376). New York: Russell Sage Foundation.

Goldberg, W. A., Prause, J., Lucas-Thompson, R., & Himsel, A. (2008). Maternal employment and children's achievement in context: A meta-analysis of four decades of research. *Psychological Bulletin, 134,* 77–108.

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

Greenhouse, J. B., & Iyengar, S. (2009). Sensitivity analysis and diagnostics. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.) , *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 417–433). New York: Russell Sage Foundation.

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

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. D. (2010). Ego depletion and the strength model of self-control: A meta-analysis. *Psychological Bulletin, 136,* 495–525.

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

Harbord, R. M., & Higgins, J. P. T. (2008). Meta-regression in Stata. *Stata Journal, 8,* 493–519.

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin, 135*(4), 555–588.

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

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

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

Hedges, L.V. (2009). Statistical considerations. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 37–46). New York: Russell Sage Foundation.

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

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

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

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

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

Higgins, J. P. T., & Thompson, S. G. (2004). Controlling the risk of spurious findings from meta-regression. *Statistics & Medicine, 23,* 1663–1682.

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations* (2nd ed.). Thousand Oaks, CA: Sage.

Huedo-Medina, T. B., & Johnson, B. T. (2010). *Modelos estadísticos en meta-análisis* [*Statistical models in meta-analysis*]. Series in Methodology and Data Analysis in Social Sciences. La Coruña, Spain: Netbiblio.

Huedo-Medina, T. B., Sánchez-Meca J., & Marín-Martínez F. (2004). Estimación del tamaño del efecto medio en un meta-análisis: Una comparación entre los modelos de efectos fijos y aleatorios. [Estimating the average of the effect size in a meta-analysis: A comparison between fixed- and random-effects models.] *Metodología de las Ciencias del Comportamiento,* Volumen Especial, 307–315.

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

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

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

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

International Labor Organization. (2012). Databases. Retrieved April 6, 2012, from http://www.ilo.org/global/lang-en/index.htm.

International Social Survey Programme. (2012). Survey data from many nations. Retrieved April 6, 2012, from http://www.issp.org/

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. doi:10.1371/journal.pmed.0020124.

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

Johnson, B. T., & Boynton, M. H. (2008). Cumulating evidence about the social animal: Meta-analysis in social-personality psychology. *Social and Personality Psychology Compass, 2*(2), 817–841.

Johnson, B. T., & Eagly, A. H. (1989). Effects of involvement on persuasion: A meta-analysis. *Psychological Bulletin, 106*(2), 290–314.

Johnson, B. T., & Huedo-Medina, T. B. (2011). Depicting estimates using the intercept in meta-regression models: The moving constant technique. *Research Synthesis Methods, 2*(3), 204–220.

Johnson, B. T., Scott-Sheldon, L. A., & Carey, M. P. (2010). Meta-synthesis of health behavior change meta-analyses. *American Journal of Public Health, 100*(11), 2193–2198.

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

Kirsch, I., Deacon, B. J., Huedo-Medina, T. B., Scoboria, A., Moore, T. J., & Johnson, B. T. (2008). Initial severity and antidepressant benefits: A meta-analysis of data submitted to the FDA. *PLoS Medicine, 5,* 260–268.

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

Kite, M. E., & Whitley, B. R. (1996). Sex differences in attitudes toward homosexual persons, behaviors, and civil rights: A meta-analysis. *Personality and Social Psychology Bulletin, 22*(4), 336–353.

Koenig, A. M., Eagly, A. H., Mitchell, A. A., & Ristikari, T. (2011). Are leader stereotypes masculine? A meta-analysis of three research paradigms. *Psychological Bulletin, 137,* 616–642.

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

Kotov, R., Gamez, W., Schmidt, F., & Watson, D. (2010). Linking "big" personality traits to anxiety, depressive, and substance use disorders: A meta-analysis. *Psychological Bulletin, 136*(5), 768–821.

Lau, J., Ioannidis, J. P. A., Terrin, N., Schmid, C. H., Olkin, I. (2006). The case of the misleading funnel plot. *British Medical Journal, 333,* 597–600.

Lee, I., Pratto, F., & Johnson, B. T. (2011). Intergroup consensus/disagreement in support of group-based hierarchy: An examination of socio-structural and psychocultural factors. *Psychological Bulletin, 137*(6), 1029–1064.

Lennon, C. A., Huedo-Medina, T. B., Gerwien, D. P., & Johnson, B. T. (2012). A role for depression in sexual risk reduction for women? A meta-analysis of HIV prevention trials with depression outcomes. *Social Science & Medicine, 75*(4), 688–698.

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

Lipsey, M. W. (2009). Identifying potentially interesting variables and analysis opportunities. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 148–158). New York: Russell Sage Foundation.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*(12), 1181–1209.

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

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

Malle, B. F. (2006). The actor-observer asymmetry in attribution: A (surprising) meta-analysis. *Psychological Bulletin, 132*(6), 895–919.

Marcus-Newhall, A., Pedersen, W. C., Carlson, M., & Miller, N. (2000). Displaced aggression is alive and well: A meta-analytic review. *Journal of Personality and Social Psychology, 78*(4), 670–689.

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of r and d. *Psychological Methods, 11*(4), 386–401.

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

Miller, N., Lee, J., & Carlson, M. (1991). The validity of inferential judgments when used in theory-testing meta-analysis. *Personality and Social Psychology Bulletin, 17*(3), 335–343.

Moher, D., Liberati, A., Tetzlaf, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Journal of Clinical Epidemiology, 62*(10), 1006–1012.

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

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

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

Moyer, A., & Finney, J. W. (2002). Randomized versus nonrandomized studies of alcohol treatment: Participants, methodological features and posttreatment functioning. *Journal of Studies on Alcohol, 63*(5), 542–550.

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

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

Noguchi, K., Albarracín, D., Durantini, M. R., & Glasman, L. R. (2007). Who participates in which health promotion programs? A meta-analysis of motivations underlying enrollment and retention in HIV-prevention interventions. *Psychological Bulletin, 133,* 955–975.

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

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

Orwin, R. G., & Vevea, J. L. (2009). Evaluating coding decisions. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 177–203). New York: Russell Sage Foundation.

Pearson, K. (1895). Contributions to the mathematical theory of evolution II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London, 186,* 343–414.

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology, 90*(5), 751–783.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112*(1), 160–164.

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

Reed, J. G., & Baxter, P. M. (2009). Using reference databases. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 73–101). New York: Russell Sage Foundation.

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

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

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

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

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Rev. ed.). Beverly Hills, CA: Sage.

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

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

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

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

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74,* 166–169.

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

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

Roth, P. L., Purvis, K. L., & Bobko, P. (2012). A meta-analysis of gender group differences for measures of job performance in field studies. *Journal of Management, 38,* 719–739.

Rothstein, H. R., & Hopewell, S. (2009). Grey literature. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 103–125). New York: Russell Sage Foundation.

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

Sanchez-Meca, J., Marin-Martinez, F., & Chacon-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods, 8*(4), 448–467.

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

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

Schmidt, F. L., Le, H., & Oh, I. (2009). Correcting for the distorting effects of study artifacts in meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 317–333). New York: Russell Sage Foundation.

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature, 470,* 327.

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

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

Shea, B. J., Grimshaw, J. M., Wells, G. A. et al. (2007). Development of AMSTAR: A measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology, 7,* 10–17.

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

Sidanius, J., Pratto, F., & Bobo, L. (1994). Social dominance orientation and the political psychology of gender: A case of invariance? *Journal of Personality and Social Psychology, 67,* 998–1011.

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

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

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

Stewart, L., Tierney, J., & Burdett, S. (2005). Do systematic reviews based on individual patient data offer a means of circumventing biases associated with trial publications? In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 261–286). New York: Wiley.

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

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 435–452). New York: Russell Sage Foundation.

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

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

Taras, V., Kirkman, B. L., & Steel, P. (2010). Examining the impact of culture's consequences: A three-decade, multilevel, meta-analytic review of Hofstede's cultural value dimensions. *Journal of Applied Psychology, 95*(3), 405–439.

Thompson, K. N., & Schumacker, R. E. (1997). An evaluation of Rosenthal and Rubin's binomial effect size display. *Journal of Education and Behavioral Statistics, 22*(1), 109–117.

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

Twenge, J. M. (2000). The age of anxiety? The birth cohort change in anxiety and neuroticism, 1952–1993. *Journal of Personality and Social Psychology, 79*(6), 1007–1021.

Twenge, J. M., Gentile, B., DeWall, C. N., Ma, D., Lacefield, K., & Schurtz, D. R. (2010). Birth cohort increases in psychopathology among young Americans, 1938–2007: A cross-temporal meta-analysis of the MMPI. *Clinical Psychology Review, 30*, 145–154.

Twenge, J. M., Konrath, S., Foster, J. D., Campbell, W. K., & Bushman, B. J. (2008). Egos inflating over time: A cross-temporal meta-analysis of the narcissistic personality inventory. *Journal of Personality, 76*, 875–901.

Twenge, J. M., & Nolen-Hoeksema, S. (2002). Age, gender, race, socioeconomic status, and birth cohort difference on the children's depression inventory: A meta-analysis. *Journal of Abnormal Psychology, 111*(4), 578–588.

United Nations Statistics Division (2012). International databases. Retrieved April 6 from http://unstats.un.org/unsd/default.htm.

Valentine, J. C. (2009). Judging the quality of primary research. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 129–146). New York: Russell Sage Foundation.

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics, 35*, 215–247.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48.

White, H. D. (2009). Scientific communication and literature retrieval. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 51–71). New York: Russell Sage Foundation.

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

Wilson, D. B. (2009). Systematic coding. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 159–176). New York: Russell Sage Foundation.

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

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

Wood, W., & Eagly, A. H. (2009). Advantages of certainty and uncertainty. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 455–472). New York: Russell Sage Foundation.

World Bank (2013). *Data*. Retrieved from http://data.worldbank.org/

World Values Survey (2012). Retrieved April 6, 2012, from http://www.worldvaluessurvey.org/

APPENDIX A: ESTIMATING EFFECT SIZES IN INDIVIDUAL STUDIES

A comprehensive treatment of the formulas to convert primary-level statistics to effect sizes is beyond the scope of this chapter (see Card, 2012; Glass et al., 1981; Johnson, 1993; Lipsey & Wilson, 2001; Rosenthal, 1991). Here we offer only the most common transforms for deriving g, the standardized mean difference effect size. For producing r from various statistical reports, Glass et al. (1981) provided several useful formulas; alternatively, the standardized mean difference, g (see Table 26.2), may be calculated and transformed to r by this equation:

$$r = \frac{g}{\sqrt{g^2 + 4}}. \tag{26.1A}$$

Effect Sizes from Means and Standard Deviations

Table 26.2, line 1, shows the equation to transform two means and a standard deviation into an effect size, $(M_A - M_B)/SD_{pooled}$. Yet, there are many possible forms of the standard deviation that can appear in the dominator of the formula. To derive g from means and standard deviations in a between-subjects design, it is conventional to use the pooled standard deviation, SD,

$$SD_{pooled} = \sqrt{\frac{(n_{A-1})(SD_A)^2 + (n_{B-1})(SD_B)^2}{n_A + n_B - 2}}, \tag{26.2A}$$

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

For within-subjects designs, Becker (1988) recommended using the pretest SD as the denominator when pretest and posttest scores are compared. Other within-subjects comparisons may be calculated as between-subjects when cell standard deviations are available. Alternatively, SD_{pooled} can be replaced with SD_d, the standard deviation of the differences between paired observations,

$$SD_d = \sqrt{SD_A^2 + SD_B^2 - 2r_{EC}SD_ASD_B}, \tag{26.3A}$$

where r_{EC} is the correlation between the paired observations (e.g., Dunlap, Cortina, Vaslow, & Burke, 1996). This form of the SD is equivalent to the $\sqrt{MS_{Error}}$ term in a repeated measures analysis of variance or in a t-test, which will generally provide relatively liberal estimates of effect size. Most often all of the components of this formula are not provided, and a paired-observation t-test or a within-subjects F is given instead. As we indicate in the next subsection, these statistics may be directly converted into the effect size that has the standard deviation of the differences in its dominator.

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

Effect Sizes from t- and F-values

Calculations of g can also be based on summary statistics. In the case of the t-test for independent groups,

$$t = \frac{M_A - M_B}{\sqrt{\frac{SD_A^2}{n_A} + \frac{SD_B^2}{n_B}}}. \tag{26.4A}$$

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

$$g = t\sqrt{\frac{n_A + n_B}{n_An_B}} \tag{26.5A}$$

Or, if $n_A = n_B$,

$$g = t\sqrt{\frac{2}{n}} = \frac{2t}{\sqrt{2n}}. \tag{26.6A}$$

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

$$g = \sqrt{F\frac{n_A + n_B}{n_An_B}}. \tag{26.7A}$$

Or, if $n_A = n_B$,

$$g = \sqrt{F\frac{2F}{n}}, \tag{26.8A}$$

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

$$g = \frac{t}{\sqrt{n}}. \tag{26.9A}$$

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

$$g = \sqrt{\frac{F}{n}}. \tag{26.10A}$$

Note that because equations 26.9A and 26.10A assume a repeated measures error variance (see equation 26.2A), they generally will provide relatively large estimates of effect size.

F-values that derive from designs with three or more conditions require some special consideration. F-values that have more than one degree of freedom in the numerator cannot be directly converted into effect sizes because they do not directly gauge differences between individual means. Rather, a significant omnibus F-value implies that somewhere among the relevant means, one or more significant differences exist (see Judd, Yzerbyt, & Muller, Chapter 25 in this volume). Thus, for example, a significant F-value from a design that uses low, medium, and high levels of the independent variable must be decomposed in order to permit effect size derivations. If a linear contrast is reported, it will be equivalent to a comparison between the high and low levels. One could compare the means only for the high and low levels or also compare the medium level with the low and the high levels (e.g., Rhodes & Wood, 1992). Or, if the relation between the independent and dependent variables is expected to be linear, one could compute an F for the linear trend in the means and transform it into g (see Glass et al., 1981; Rosenthal & Rosnow, 1985). Of course, analysts should use the means in a particular study that would produce the most similar comparison to that used to represent the other studies in the sample. Treating studies' results in substantially

TABLE 26.1A. Hypothetical analysis of variance summary tables (a) before reconstitution and (b) after returning factor B's sums of squares to the error term degrees

Source	Sum of squares	Degrees of freedom	Mean squared error	*F*
(a) Before reconstituting				
A	430.33	1	430.33	15.22
B	200.12	1	200.12	7.08
A × B	43.55	1	43.55	1.54
Error	1,244.29	44	28.28	
(b) After reconstituting				
A	430.33	1	430.33	13.30
Error	1,487.96	46	32.35	

different ways would introduce noise into the effect sizes in the database.

Similar issues arise in designs with two or more factors. In such instances, to make effect size comparisons more similar across the studies in a meta-analytic sample, some methodologists have recommended producing one-way designs by returning the effects of irrelevant factors to the error term of the ANOVA (Glass et al., 1981; Hedges & Becker, 1986; Morris & DeShon, 1997). This procedure should be seriously considered for individual-difference variables that were crossed with the crucial independent variable in only some of the studies, because this source of variability would not have been removed from the error term in studies that did not assess these individual differences. When these irrelevant variables were instead manipulated, the decision is less straightforward, to the extent that researchers have created extreme conditions atypical of natural settings by means of powerful experimental manipulations. Variability stemming from extreme or atypical conditions would not be in the error term of typical studies. Therefore, adding sums of squares for such manipulated variables to the sum of squares error could greatly inflate these error terms in at least some instances and thus decrease the absolute magnitude of effect sizes based on these error terms. As Morris and DeShon (2002) concluded, in deciding whether to return irrelevant factors to the error term, analysts should keep as their goal the production of error terms that are based on the same sources of variability across the studies in the sample.

To illustrate how to return irrelevant factors to the error term, Table 26.1A contains a hypothetical ANOVA for a two-factor design. The top panel contains the ANOVA summary for the two factors. Suppose that Factor *A* is the focal independent variable, and that Factor *B* is a meta-analytically irrelevant variable. To represent the impact of Factor *A* on the dependent variable, the variation due to Factor *B* can be returned to the error term. This operation is performed by (a) adding the sum-of-squares due to Factor *B* and its interaction with Factor *A* to the error sum-of-squares and (b) adding the degrees of freedom due to Factor *B* and its interaction to the degrees of freedom for error. Once the sum-of-squares for error has been divided by its new degrees of freedom, the square root of the resulting mean-square for error would be interpretable as the standard deviation pooled within the two levels of *A*, or $SD = \sqrt{MS_e}$. The result of this reconstitution of the error term appears in Panel b. In this example, *g* may be derived by converting the *F*-value that resulted from the reconstitution procedure, or it may be derived by dividing the difference between the means of Factor *A* by *SD*. Morris and DeShon (1997) presented other equations and examples of this strategy; Nouri and Greenberg (1995) presented techniques for use with more complex ANOVA designs (e.g., those that mix between- and within-subjects factors).

If the effects of the focal independent variable on the dependent variable are expected to change within the levels of another independent variable, separate effect sizes can be calculated within levels of the second independent variable, as we already mentioned above (see subsection "Multiple Reports from Individual Studies"). Specifically, as an alternative to representing the effect of the focal independent variable aggregated over this other variable (i.e., as a main effect), the analyst can partition each study on this other variable and represent the effect of interest

Table 26.2A. A Hypothetical factorial design in which a focal independent variable is crossed with a moderator-independent variable

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

within levels of this variable (i.e., as a simple main effect). When interactions are expected, simple main effects are the desired comparison, and the other, interacting variable can function as a moderator of the relation between the focal variables. As an example, Table 26.2A displays a 2 × 3 factorial design in which the focal independent variable (IV_{focal}) and a moderator variable ($IV_{moderator}$) serve as the factors. Suppose that we expect the effect of IV_{focal} on the dependent variable to change depending on the level of $IV_{moderator}$. To represent these contrasting expectations, a separate effect size must be derived for each level of $IV_{moderator}$. Thus, the first *g* would result from a comparison of the means from cells *a* and *b*, the second from cells *c* and *d*, and the third from cells *e* and *f*. To perform this calculation, it is necessary to obtain all cell means and either (a) the within-cell standard deviations, (b) the standard deviations for each relevant level of $IV_{moderator}$ (and transformed to SD_{pooled}), or (c) MS_e for the ANOVA. The MS_e can be recovered when all cells means are reported and at least one *F*-value is known for the dependent variable, even when the available *F* is not the most relevant to the analysts' focal comparison (Johnson, 1993; Morris & DeShon, 1997). These calculations are facilitated if the source report contains a complete ANOVA table, but the components of the table can be estimated if the means, cell sizes, and one or more *F*-values are known (Johnson, 1993). Then, $SD = \sqrt{MS_e}$. Once this value or the standard deviations are known, effect-size derivations continue as though each condition were a separate experiment.

Finally, *F*-values derived from multivariate analysis of variance (MANOVA), in which one or more independent variables were examined for their simultaneous influence on two or more dependent measures, should not be transformed into effect sizes if the dependent variable of interest was combined with other, irrelevant dependent variables (see Morrison, 1976; Timm, 1975). If several measures of the same conceptual dependent variable were combined in a multivariate analysis, however, the analyst might derive an effect size by taking the square root of the proportion of variance that the independent variable accounts for in the best linear combination of the dependent variables and treating this value as an *r* (see Tabachnick & Fidell, 1996, pp. 388–391, discussion of Wilk's Lambda), even if univariate *F*-values from ANOVAs are not available. However, because such effect sizes would be dependent on the exact set of dependent variables included in the multivariate analysis, some meta-analysts recommend against such procedures (Hunter & Schmidt, 1990).

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

Effect Sizes from *r*-values

Although *r* can be readily transformed to *g*,

$$g = \frac{2r}{\sqrt{1 - r^2}}, \tag{26.11A}$$

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

$$r_b = \frac{r_{pb}\sqrt{n_A n_B}}{\mu N}, \tag{26.12A}$$

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

$$r_b = \frac{r_{pb}\sqrt{n_A n_B}}{\mu N}. \quad (26.13A)$$

Whereas standardized regression weights (β) deriving from simple linear regressions are r-values and can be so interpreted, βs deriving from regressions with more than one predictor *cannot* be directly interpreted as r-values. The β-value for a given predictor in a multiple regression equation is *adjusted* for the other independent variables present in the equation. In the case of suppressor variables (Cohen & Cohen, 1983), these adjustments can affect not only the value of β but also its sign, which could be reversed from the sign of the correlation between the two variables. Yet another problem with converting β-values to effect sizes is that under some circumstances β-values from multiple regression equations exceed |1|, whereas r-values never exceed ||1|. For example, if Equation 26.11A is used with a β of 1.1, the denominator of the equation will be the square root of a negative number, –0.21, which is an irrational mathematical operation. Therefore, as a general rule, in meta-analyses for which multiple regression results are the exception and other studies in the sample report statistics unadjusted for the other variables in the equation, multiple regression results should not be converted to effect sizes (see Hunter & Schmidt, 1990). Of course, before discarding a study because its findings were reported in a multiple regression, one should see whether a correlation matrix or comparable statistics appear in the report or could be obtained from its authors.

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

Effect Sizes from Chi-square Values

Chi-square (χ^2) values are sometimes used to test for the frequency with which groups meet some criterion or to test for the association between two variables (Hays, 1988). If the χ^2 results from a 2 × 2 classification table linking a predictor (X) to the outcome (Y), then r can be calculated:

$$r_\varphi = \sqrt{\frac{\chi^2}{n}}, \quad (26.14A)$$

where r_φ is a phi coefficient and approximates the product-moment r and can be converted to g:

$$g = \frac{2r}{\sqrt{1-r^2}}. \quad (26.15A)$$

Note: that if there is more than 1 degree of freedom in the χ^2 value, it cannot be directly converted into an effect size because the χ^2 may describe a nonlinear pattern. It may be possible to compute χ^2 for an appropriate 2 × 2 table based on the proportions of the relevant groups that meet a criterion (see the next subsection). If the data for these recomputations are not available, the study result cannot be used to derive an effect size.

Effect Sizes from Proportions Meeting a Criterion

In some designs, the proportion of individuals in one group (p_E) who meet a given criterion is compared with the proportion of individuals in another group (p_c) who meet it. For example, the proportion of people who help another person in one experimental condition can be compared to the proportion of people who help in another condition (see Eagly & Crowley, 1986). Although these proportions can be transformed into an effect size by using a probit transformation (Glass et al., 1981) or by treating the proportions as means (Snedecor & Cochran, 1980), the most efficient solution is to use the Cox transformation of the odds ratio gauging the effect size (see Table 26.2A, line 8),

$$g_{Cox} = \frac{LOR}{1.65}, \quad (26.16A)$$

where LOR is the logged odds ratio (Sánchez-Meca, Marín-Martínez, & Chacón-Moscoso, 2003). Note that this equation assumes that the proportions are in relation to the study's unit of analysis, which usually is the numbers of persons. The equations do not apply to proportions that represent values of dependent variables assessed for each unit of analysis. For example, if each participant's helping were assessed by a self-report of the proportion of occasions on which he or she helped, these data would produce an effect size by equations that use the variability of these proportions (e.g., Table 26.2, line 1) rather than Equation 26.16A.

Effect Sizes from Probabilities Associated with Inferential Statistics

Source reports sometimes contain only a p-value associated with the critical effect (e.g., $p = .0439$), which can be used to calculate an effect size if the direction of the finding and the sample size (n) are known. To do so, the analyst would use a statistical package's (e.g., SAS, IMSL, SPSS, Stata) or a spreadsheet's inverse probability distribution functions, which provide an exact solution of a test statistic from p. For example, SAS provides BETAINV, which yields F from p and df, after which the F can be converted to g using Equations 26.7A through 26.10A (assuming that the F compares the means of only two groups). Obviously, an exact p allows an excellent estimate of a test statistic and therefore of g. Conversely, a level p (e.g., $p < .05$) gives a poorer estimate, because it would ordinarily be treated as exactly the p level given (e.g., $p < .01$ would be understood as $p = .01$). The mere statement that a finding is "significant" can be treated as $p = .05$ in studies that apparently use the conventional $p < .05$ rule for determining significance and indicate the direction of the effect, but the effect sizes estimated on this basis may be quite inaccurate (Ray & Shadish, 1996). Finally, reports often differ in whether a one-tailed or two-tailed probability level is reported; if no information is provided, the convention is that the study authors have used a two-tailed test.