Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.
Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.
Thibaut, J. W., & Kelley, H. H. (1959). *The social psychology of groups*. New York: Wiley.

CHAPTER 15

# *Quantitative Research Synthesis*

## *Examining Study Outcomes Over Samples, Settings, and Time*

WENDY WOOD
*Texas A&M University*

P. NIELS CHRISTENSEN
*San Diego State University*

Research synthesis is a technique for reviewing research literatures that provides quantitative estimates of the size and pattern of results across studies. In many ways, this technique is analogous to other scientific research methods. Research synthesists generate hypotheses about a phenomenon, collect data, analyze the data with meta-analytic statistical techniques, and draw conclusions. A unique feature of research synthesis is that the study, and not the individual participant, is the primary unit of analysis. The set of studies evaluated in a synthesis likely represents the efforts of multiple researchers who tested different samples of participants using a variety of operations of variables in varying experimental contexts at a variety of points in time. Thus, syntheses can provide estimates of effects across studies, measures, research contexts, and time periods. This chapter describes the process of conducting a research synthesis.

Research synthesis is a quantitative alternative to narrative reviews (e.g., term papers), which provide qualitative summaries of research findings. A hallmark of research synthesis is the use of systematic decision rules and procedures (Lipsey & Wilson, 2001). Narrative reviews can be conducted systematically but often are not

AUTHORS' NOTE: Preparation of this chapter was supported by a grant from the National Institute of Mental Health (1R01MH619000-01). The authors thank Alice H. Eagly, Blair Johnson, and Jeanne Twenge for their thoughtful comments on an earlier draft of the chapter.

(see Bushman & Wells, 2001; Cooper & Rosenthal, 1980). Following a systematic process does not ensure that everyone agrees with the conclusions, but it does make each step of a synthesis explicit and available to others' scrutiny. This use of systematic procedures to identify, code, and analyze study findings becomes increasingly important with greater numbers and greater complexity of the studies included in the review. For example, the classic Smith and Glass (1977) synthesis of psychotherapy outcomes included more than 300 studies—an accomplishment that would have been difficult with narrative reviewing techniques.

Another benefit of quantitative reviews is the use of effect size estimates. Effect sizes represent the direction and magnitude of an effect apart from its statistical significance. Narrative reviews often base conclusions on the statistical significance of individual study findings in the reviewed literature. This can be misleading, especially when the reviewed studies used small sample sizes that may not have provided sufficient power to detect effects. Then reviews can wrongly conclude that an effect is not present when in fact it is (Schmidt, 1996). In addition, when exact effect sizes are estimated for each study, the overall variability can be calculated across the study effects. Low variability might indicate that the reviewed studies converge on a common conclusion, whereas high variability might indicate that study findings are influenced by factors that vary across studies (e.g., attributes of the research participants). In general, understanding the variability in a phenomenon can be just as important as understanding its overall size.

In this chapter, we draw on our own experiences in conducting research syntheses to develop general introductory guidelines for how to conduct such reviews. First, we give examples from two of our own syntheses that illustrate the various purposes for conducting reviews. Then we describe the procedures involved in conducting a research synthesis, emphasize the key decisions at each stage of the process, and consider some of the potential pitfalls associated with quantitative reviewing. We conclude with a discussion of how to interpret synthesis results.

## USES FOR QUANTITATIVE RESEARCH SYNTHESIS

### *Evaluating Existing Theories*

The question of whether media violence affects aggression might have an obvious answer to anyone who has spent a Saturday morning watching cartoons with young viewers. However, both scientists and the general public have debated the existence of media violence effects. A large body of empirical research has tested this effect, and a number of literature reviews have compiled the research findings. Although some early reviews failed to detect much of a relationship (e.g., Freedman, 1988; McGuire, 1986), most subsequent reviews concluded that media violence increases aggression (e.g., Felson, 1996; Geen, 1998; Heath, Bresolin, & Rinaldi, 1989; Paik & Comstock, 1994).

Given the practical and theoretical importance of media violence effects, these conclusions deserve critical scrutiny. In general, a research synthesis is only as good as the studies on which it is based. If the primary studies being reviewed are poorly constructed or share a common flaw, then the review conclusions may suffer from a "garbage-in, garbage-out" problem (Cook & Leviton, 1980; Sharpe, 1997). With respect to media violence, many studies have used correlational designs. Syntheses of such studies, like the individual research reports, cannot demonstrate a causal effect of media violence on viewer aggression. In addition, media violence effects as studied in the laboratory may differ from the real world settings in which people usually watch and react to violent programs. Thus, the external validity of many laboratory experiments—and syntheses of these experimental findings—may not be appropriate for drawing conclusions about media exposure and aggression in everyday life.

To provide a targeted answer to the question of whether media violence has a causal impact on aggressive responses in everyday contexts, we (Christensen & Wood, in press; Wood, Wong, & Chachere, 1991) conducted a synthesis of 24 experiments. The studies in our review ensured the validity of causal inferences by randomly assigning participants to be exposed to violent or nonviolent media. Furthermore, participants' aggressive responses following media exposure were assessed during naturally occurring social interaction with peers or strangers. By limiting the synthesis to reports that satisfied these constraints, we were able to draw conclusions about the causal effect of media violence in everyday situations. The results indicated that exposure to media violence does increase aggression: Approximately 20% more participants assigned to the media violence conditions displayed greater than average aggression, when compared with the participants assigned to control conditions.

The central question addressed in research syntheses on media violence also was the focal question in most of the primary research reviewed. This synergy between syntheses and primary research emerges when topics of current interest to the field generate considerable primary research, and this interest, along with the available data base, then enables quantitative research reviews. Although aggregated findings often can be anticipated from the findings of the individual studies in a review, some reviews have yielded surprising results. For example, in a synthesis concerned with domestic violence, Archer (2000) challenged the common notion that such physical aggression is perpetrated primarily by men. Although men were found to engage in more intense aggression against their partners than did women—and thus men inflicted greater physical harm—women were found to engage in more frequent acts of aggression against their partners than did men (although see Frieze, 2000; O'Leary, 2000). Follow-up syntheses are also useful when primary research has yielded an inconsistent pattern of effects. By identifying key moderators of an effect, a synthesis can potentially explain the inconsistency. For example, a number of syntheses have addressed the elusive phenomenon of mental telepathy, and the most recent review obtained reliable evidence of telepathy primarily in research that followed a standard paradigm (Bem, Palmer, & Broughton, 2001; see also Bem & Honorton, 1994; Milton & Wiseman, 1999).

### *Testing Novel Hypotheses*

Meta-analytic syntheses in social psychology sometimes test hypotheses that were not the focus of the primary reports and can be addressed only when findings are aggregated across studies. Synthesists can examine whether an effect depends on the features of the studies that vary in the reviewed literature. For example, some of the research syntheses on sex differences have evaluated primary studies that included only one sex, allowing the synthesist to compare findings from male-only studies with those from female-only ones (see Eagly & Wood, 1991). Also, syntheses have effectively evaluated changes in social phenomena over time by examining whether study findings depend on the year when data were collected or a report was published (see Twenge, 2002).

To illustrate how research syntheses test new hypotheses not addressed in primary research, we describe Ouellette and Wood's (1998) synthesis of the effects of intentions on behavior. These researchers were interested in the fact that people intend to act in certain ways, such as to adopt a healthy lifestyle, but

these intentions don't always translate into actions. There is even a holiday based on this phenomenon, the New Year's resolution. For a few days or weeks in January, most people can stick to resolutions such as exercising or eating healthy foods, but then most revert back to their old behavior patterns.

Why is it so difficult for people to make changes in many of the behaviors they perform every day? One explanation is that behavior is guided by multiple systems, only some of which are conscious and involve explicit intentions. When behavior is well practiced, so that it becomes habitual, it can be cued automatically by features of the environment with minimal awareness. Ouellette and Wood (1998) hypothesized that habits can explain why people do not always act on their conscious intentions. Over time, when explicit intentions such as New Year's resolutions conflict with automatic habitual responses, the habits often will win out.

Meta-analytic techniques were ideal to test this hypothesis. Even though no primary studies had examined the different systems guiding behavior, a broad literature exists on behavior prediction. Ouellette and Wood divided 64 prediction studies into groups based on whether they investigated a behavior that was likely to be guided by habits or by intentions. Behaviors such as drinking coffee and wearing seatbelts are ones for which people can establish habits because they have sufficient opportunity to perform the behavior in stable contexts. For these behaviors, the synthesis revealed that people followed their established routines and their behavior was not influenced strongly by their stated intentions. In contrast, it is difficult to establish habits for behaviors such as getting a flu shot or donating blood, which occur only once or a few times a year. For these behaviors, the synthesis revealed that people were highly likely to carry out their intentions. In general, then, the synthesis provided evidence for a new hypothesis that was not tested by the individual studies in the review but that could be evaluated when the studies were aggregated.

## PROCEDURES IN CONDUCTING A RESEARCH SYNTHESIS

### *Determining if You Have Enough Studies*

A research synthesis requires the existence of a number of primary studies that are relevant to the research hypothesis. The question of how many studies are needed is a question of the power of statistical tests. That is, does the synthesis have sufficient power to detect the overall effect and the variability in effect estimates? Although research syntheses are often reputed to have high statistical power, this is not always the case (Hedges & Pigott, 2001). A statistical rule of thumb is that larger numbers of studies are required when the effect of interest is small and highly variable and when the reviewed studies are underpowered (see Cohen, 1988, 1992). From a statistical perspective, only two studies would be needed to detect an effect typical of the size found in social psychology (Hedges & Pigott, 2001).[1] However, almost 30 studies would be required for sufficient power to detect a medium degree of variability in the effect estimates (Hedges & Pigott, 2001). From a practical standpoint, the required number of studies is jointly determined by the constraints of power and the meaningfulness of the results once synthesized.

A typical assumption in meta-analysis is that the studies to be aggregated are replications of each other. However, this is rarely true in practice. Researchers often build on each other's work by varying aspects of an earlier research design to determine boundary conditions or to examine the variety of processes that generate an effect. For this reason, research literatures typically include numbers



of studies that vary in their specific features or operations but test a common theoretical or conceptual idea. Critics of the technique have argued that aggregating across studies with diverse operations and participant samples is tantamount to combining "apples and oranges" (Cook & Leviton, 1980; Sharpe, 1997). As we explain below, meta-analytic statistical techniques that estimate the extent of variability across study findings provide a way to assess whether it is reasonable to aggregate across a given set of studies and to treat them as tests of a common hypothesis.

### *Defining the Problem, Variables, and Sample*

The first step in a synthesis is developing a clear research question. Typically, a synthesis evaluates evidence for the relation between two variables, and this relation often involves the influence of an independent variable on a dependent variable. To frame the research question appropriately, reviewers will need to evaluate the history of the research problem in the literature.

Theoretical explanations for the effect can help to identify the appropriate ways to define the variables of interest and the sample of studies to be included in the review. Because individual studies can differ widely in the operations they use to measure and manipulate variables, synthesists have to develop their own definitions of variables that can be applied systematically to the reviewed studies. These definitions need not include all available operations of constructs in the literature. For example, Ouellette and Wood (1998) restricted their definition of habits to frequency of past behavior and did not include potentially less reliable measures, such as people's reports of whether they performed a behavior out of habit.

The research question directing the synthesis also determines how the sample of studies is defined and how the boundary conditions are set for including and excluding research. In our experience, the decision about what studies to include is fluid and often changes as the review progresses. It is not uncommon for reviewers to modify the research question, the definition of variables, and the boundary conditions for the sample as they discover the full range of variables and designs that have been used to test the research hypothesis.

To generate an optimal sample definition, synthesists need detailed knowledge of the research question and the relevant literature. When samples are defined too broadly, they include heterogeneous sets of studies that tap diverse psychological processes. Such a high level of variability can obscure the effect of interest. For example, if Christensen and Wood (in press) had included studies of exposure to print media in addition to video, the less vivid written presentation might have reduced media effects or limited them only to participants who attended to and comprehended the printed material. Of course, if a reviewer were interested in comparing media effects across modality, it would be appropriate to include print and video presentations in the sample of studies and to conduct analyses to compare them.

The opposite problem occurs when samples are defined too narrowly. In that case, reviews may generate a limited conclusion that does not make a significant contribution beyond the individual studies themselves. One example of a narrow sample definition occurs when researchers aggregate findings across multiple studies within a single research report. Single-report syntheses provide estimates of an effect size in a particular experimental context. Because the cumulative sample size across studies is larger than that of individual studies, these syntheses provide greater power for tests of statistical significance. However, the conclusions of such a synthesis are limited to the specific research participants, stimuli, and setting of the initial studies.

Sample definitions sometimes specify that only high-quality studies are eligible to be included. However, including studies with lower methodological quality is not likely to bias the review's conclusion if each study has its own unique flaws and the overall effect aggregates across the individual study biases. Another reason to include studies regardless of quality is that scientists differ in their evaluations of quality, and decisions about what methodologies to include and exclude may be hard to defend. Yet, restrictive methodological criteria have the advantage of yielding results from the most credible studies. In our example at the beginning of this chapter, Christensen and Wood (in press) focused on studies with high internal and external validity in order to provide clear evidence of the causal direction of the media effect and its size in everyday social interaction. Selecting studies for certain criteria also makes sense when systematic study biases can be identified in the sample. For example, if small-sample studies get published primarily when their effects are statistically significant, then including such studies will likely inflate the overall estimate of effect size (Johnson, Carey, & Muellerleile, 1997; Kraemer, Gardner, Brooks, & Yesavage, 1998). Although there is no all-purpose solution to the problem of study quality, we recommend adopting lenient selection criteria for methodologies while coding each study for a wide range of methodological and procedural features. This allows researchers to evaluate the potential impact of the various methodological variations.

### *Locating Relevant Studies*

The standard goal of a research synthesis is to describe either the universe of studies that has examined a particular relation or an unbiased sample of this universe. For this reason, it is important to include all relevant research in the study sample. When a research literature is very large, one solution is to randomly select a subset of studies to evaluate.

Given the goal of describing the universe of studies, both published and unpublished studies are eligible for inclusion. Including unpublished studies is especially important when there is reason to expect publication bias, especially the tendency for studies with small samples to be published only when their effects achieve statistical significance. Publication bias might seem less of a concern when the relation tested in the meta-analysis was not examined in the original studies, as occurred with Ouellette and Wood's (1998) finding that behavioral intentions are weak predictors when habits exist to direct actions. But even then, we strongly advise including unpublished research to ensure that the sample of studies in the review validly represents the sample of studies conducted on the issue.

A variety of strategies can be used to locate relevant studies. Because each search strategy is likely to provide some unique information, reviewers should use all the strategies available. The most common technique for locating studies is to conduct a search of computerized databases (e.g., PsycInfo, WorldCat). These can be used to identify articles that mentioned certain terms in their titles or abstracts (e.g., the variables of interest in the review) and articles that were classified in the database as being relevant to general index terms. Unpublished studies often can be located through searches of Dissertation Abstracts or ERIC (Educational Resources Information Center). Another strategy is the descendancy approach, in which reviewers identify an important early article and then search a database for subsequent articles that cited the initial work. The result of these computerized database searches typically is a list of titles and abstracts, which needs to be evaluated further for eligibility for the review. Typically, reviewers will need to locate and read the actual articles to determine whether they meet the study inclusion criteria.

Another standard strategy is to use the ancestry approach, searching the reference lists of reviews and articles. Additional articles also can be identified through hand searches of relevant journals. Sometimes reviewers also use an "invisible college" strategy and contact active researchers in an area and others who might have relevant research that is unpublished or in the process of being published. In general, the process of locating studies in research synthesis bears some similarity to the procedures of sampling respondents in survey methodologies; like participant sampling, it is important that study identification proceed systematically. Synthesists should document the search strategies that they used, along with the key terms used to search computerized databases. When these are reported in sufficient detail, other researchers can replicate the search process.



### *Forming the Meta-Analytic Database*

To form the meta-analytic data set, coders read through eligible studies to extract the relevant information and record it on a standard coding form. The database is compiled from these forms. It consists of codes of the potentially important study attributes along with effect sizes, or quantitative estimates of each study's findings.

#### *Coding Study Features*

What information should be coded from each study? In general, synthesists code features of the studies, such as their methodology, along with the theoretically identified conditions that might reduce or strengthen the effect.

Because study methodology can affect study findings, codes should be formed to reflect features of each study's methods. For example, Christensen and Wood (in press) coded whether the media violence studies in their sample were conducted in laboratory settings or in everyday contexts (e.g., schools, playgrounds). Laboratory studies generally reported larger effects of exposure to media violence, presumably because they provided greater control over extraneous variables and less random variation. Other aspects of study methodology that often are evaluated in meta-analyses include the reliability of the measures (e.g., the number of items for each measure); aspects of the procedure and design, including indicators of treatment strength (e.g., length of film, extremity of violence depicted); and attributes of the research participants, such as sex and age (see review by Wilson & Lipsey, 2001).

A variety of additional study attributes could influence study findings, including the year the study was completed, the nationality of the research participants, and the identity of the study authors. For example, a number of syntheses have reported that the original researchers who identify a phenomenon tend to obtain stronger evidence of the predicted effect than subsequent or unrelated researchers (e.g., Johnson & Eagly, 1989; Wood, Lundgren, Ouellette, Busceme, & Blackstone, 1994). These effects of researcher identity could reflect a variety of factors, including the original researchers' superior understanding of the phenomenon as well as the tendency for subsequent research efforts to examine the boundary conditions for an effect by identifying when it appears and when it does not.

The moderating factors that theoretically are expected to affect the size of the relation of interest are another important source of coded variables. For example, Ouellette and Wood (1998) evaluated the relation between people's intentions and behavior by coding whether the studied behaviors were ones for which habits were likely to operate. In addition, synthesists sometimes derive codes from

sources other than the original articles. For example, Costa, Terracciano, and McCrae's (2001) meta-analysis of cross-cultural sex differences in personality included codes representing the status of women compared with men for each culture studied in the reviewed research. They were then able to evaluate whether personality sex differences varied with men's and women's status in societies. In Eagly and Steffen's (1986) synthesis on sex differences in aggression, undergraduate students rated the reviewed studies for how much harm the students would expect to experience in each experimental setting. The researchers then predicted the sex differences in aggression from male and female students' perceptions of harm. Thus, syntheses can test a variety of theories using coded variables from the studies themselves and from external sources.

Study codes typically are recorded on standard forms (see Stock, 1994, for an example of a coding form). Because it is important that coding be done accurately, multiple coders should individually complete the coding forms, and their responses should be compared. Wide disagreement among coders suggests that variables need to be more clearly defined and that the coding procedure needs to be implemented again. The agreement or interrater reliability between coders will need to be calculated (e.g., Cohen's kappa) and reported along with the synthesis results (see Orwin, 1994, for reliability statistics).

### *Selecting Computer Programs to Calculate and Analyze Effect Sizes*

The goal of effect size calculations is to convert the outcome information provided by each study into a common metric. This allows synthesists to aggregate and compare the outcomes of independent studies that may have used different research designs and operations of variables. This chapter does not address the technical details of the statistical computations involved in meta-analysis. A number of statistical computing packages have been developed that can perform the two central meta-analytic computations: first calculating effect size estimates and then analyzing the estimates generated (e.g., Borenstein & Rothstein's, 1999, *Comprehensive Meta-Analysis*; Johnson's, 1993, *DSTAT 1.10*; Lipsey & Wilson's, 2001, *MS Excel* Effect Size Computation Program and *SPSS* Macros for Meta-Analysis; Shadish, Robinson, & Lu's, 1999, *ES*; Wang & Bushman's, 1999, *SAS* Macros for Meta-Analysis). As we explain in the section on analysis techniques, the unique properties of meta-analytic data make it inappropriate to use standard statistical packages.

Each of the available programs has specific strengths and weaknesses. For example, *DSTAT* and *ES* provide highly comprehensive facilities to calculate effect sizes from complex experimental designs. *Comprehensive Meta-Analysis*, the *SPSS Macros*, and the *SAS Macros* provide a useful range of analytic procedures once effect sizes have been calculated. Yet, to use any of these programs effectively, analysts need to understand how to select the data from a research report, design a meta-analytic database, and select an analysis strategy. The present chapter provides a guide for making these decisions.

### *Calculating Effect Sizes*

A number of statistics can be used to represent the size of an effect. Regardless of the particular statistic, the effect size is given a positive or negative sign to indicate the direction of the relation between the two variables of interest. The signs are applied so that studies with opposite outcomes have opposing signs. The sign is given in a way that ensures readers' easy interpretation. A positive sign often is used to signify a relation in the predicted direction, and a negative sign is used to signify an unexpected outcome. However, this strategy does not always convey a clear meaning. In some cases,



such as when inverse relations are predicted, interpretation would be clearer if expected outcomes had a negative sign.

One commonly used effect size is the standardized mean difference, which represents the size of the relation of interest in terms of standard deviation units. It is calculated as

$$g = \frac{M_E - M_C}{SD_{pooled}},$$

where $M_E$ is the sample mean in the experimental group, $M_C$ is the sample mean in the control group, and *SD* is the pooled standard deviation for the groups. Because this formula overestimates population effect sizes, especially when they are based on small samples, the estimate typically is corrected for this bias, $d = J(m)g$, where $d$ is an unbiased estimator of the population effect size and $J(m)$ is the correction,

$$J(m) = 1 - \frac{3}{4m - 1},$$

where $m$ is the degrees of freedom, or $n_E + n_C - 2$. Researchers often label the uncorrected effect size estimate as $g$ and the corrected estimate as $d$.

The standardized mean difference is commonly used as the effect size metric when the majority of reviewed studies report comparisons between two groups, such as an experimental group and a control group. Thus, this effect size often has been used in syntheses of experimental research in social psychology. Unfortunately, the statistic $d$ rarely can be taken directly from the research reports, and it typically needs to be calculated by the analyst.

To calculate $d$ for a between-subjects experimental design, the study's means, standard deviations, and sample sizes can be inserted into the formulas above. The estimate of the pooled standard deviation, *SD*, represents the square root of the pooled variances of the two groups and is the same variance estimate as that used to calculate an *F* or *t* test difference between two groups. The formula is computed as

$$SD = \sqrt{\frac{(n_E - 1)\,SD_E^2 + (n_C - 1)\,SD_C^2}{n_E + n_C - 2}},$$

where $n_E$ and $n_C$ are the number of observations in the experimental and control groups, respectively, and $SD_E$ and $SD_C$ are the standard deviations for the experimental and control groups, respectively. When the groups to be compared are from a within-subjects design, the correct variance estimate is the standard deviation of the differences between paired observations. Also, when the standard deviation for the overall sample is given, it needs to be converted to the pooled within-cell standard deviation. This is accomplished by removing from the overall estimate the variance that is due to the difference between the experimental and control conditions being compared (Hedges & Olkin, 1985).

It is also possible to calculate effect sizes from the probability values associated with statistical tests. The $p$ value of the comparison of interest, the sample size, and the direction of the effects (i.e., whether the experimental or control condition obtained the higher score) can be entered into most meta-analytic computer programs to yield an effect size estimate. Yet, probabilities often do not allow for exact estimates. They may be presented as levels (e.g., $p < .05$, $p < .01$), or a study report might simply note that a finding is statistically significant and thus imply that $p < .05$. In this case, the most reasonable assumption is that $p$ is equal to the specified level. Obviously, this yields only an approximate estimate of the true effect size,

and alternate information should be used when available.

Most meta-analytic computer programs also can calculate effect sizes from chi-square tests of association or from proportions of study participants in experimental groups that meet a criterion value on some measure (e.g., the percentage of media viewers who engaged in aggressive acts). However, if chi-square values have more than one degree of freedom, they cannot be converted directly into effect sizes. Instead, the frequency tables need to be converted into comparisons between two groups, perhaps by aggregating across the data that are presented.

Another frequently used effect size estimate is the correlation coefficient, *r*. It is also corrected for overestimation bias that occurs especially with small samples,

$$G_r = r + \frac{r(1 - r^2)}{2(n - 3)},$$

where $G_r$ is the approximation of the population effect size and $n$ is the sample size. Because the sampling distribution of the correlation coefficient does not follow a normal curve, it is conventional to use Fisher's *r*-to-*Z* transform and to perform meta-analytic calculations on the *Z* values. Correlation coefficients typically are used in syntheses if most of the reviewed studies report relations between two continuous variables. But because *r* can be transformed easily into *d* and vice versa, the choice of an effect size metric for meta-analysis is somewhat arbitrary. A simple rule of thumb that holds for small and moderate-sized effects is that *r* values are numerically about half (or slightly more than half) the size of *d* values, so that $r = .20$ corresponds to $d = 0.40$. In general, the convention is to use the metric that most closely represents the way the majority of study findings were originally reported.

An advantage to using the correlation coefficient as an index of effect size is that this statistic often can be extracted directly from study reports. However, *r* statistics are not all interchangeable, and relations reported in the form of a point-biserial *r* will need to be converted to a product-moment *r* (see Rosenthal, 1991). In addition, when regression models are reported, standardized beta weights can be interpreted as correlation coefficients only when the model has a single predictor. In regressions that include multiple predictors, the regression weight for the variable of interest is adjusted for the other predictors. Thus, the results of regressions with multiple predictors cannot be used to represent the simple relation between the two variables of interest (although see Becker & Schram, 1994, for more complex solutions).

In general, calculating effect sizes involves many complex decisions. To ensure reliability, we recommend that two people independently complete these calculations and compare their results.

## PROBLEMS (AND SOLUTIONS) WHEN CALCULATING EFFECT SIZES

### *Independence of Observations*

Single studies often report multiple outcomes that could be included in the data set. For example, a study might report media effects on both verbal and physical aggression. Including multiple outcomes from a single study violates the assumption of most meta-analytic techniques that each outcome is independent of the others. One solution is to generate a single effect size estimate for each study, perhaps by computing an average or some other measure of central tendency (see Gleser & Olkin, 1994). Another example of multiple outcomes occurs when studies provide more than one form of results that could be used to calculate an effect size. For example, a study might provide means and standard deviations along with a *t* value. When both sources yield similar information, analysts can compute effect size estimates from both and take their average.When the estimates are highly discrepant, the analyst could select the one that appears most valid or could exclude the study from the review because the findings are too ambiguous.



Although independence of effects is a goal in forming the meta-analytic database, it is not always possible. For example, a synthesist might not be interested solely in the overall effect of exposure to media violence but also in whether the effect varies with sex of viewer. One strategy is to conduct the analysis in stages with two data sets. The primary data set would evaluate media effects across all study participants by calculating a single effect size for each study. In this data set, all effects would be independent. Then, in a second stage to test for sex effects, a subset of studies could be selected that reported the findings separately for the sexes. For this subset, each study would yield an effect size for males and an effect size for females, and analyses could then compare the sexes. Although the data in this second stage are not independent and run the risk of biasing meta-analytic test statistics (see Gleser & Olkin, 1994), they provide information about the moderating variable of interest.

### *Complex Primary Study Designs*

Calculation of effect size estimates is not straightforward in studies with experimental designs that varied factors in addition to the relation of interest. The effect sizes in these studies need to be comparable to those in the other studies in the review. First, select the appropriate experimental condition(s) to compare to the control condition(s). Because experiments may have multiple experimental conditions and multiple controls, the experimental versus control comparison selected in each study should afford clear inferences about the treatment of interest. For example, a study of viewer aggression might have included experimental conditions in which participants were angered before exposure and other conditions in which they were not. The experimental condition(s) to use in the effect size calculations depend on the control condition(s) in the study. When control participants were not angered, the no-anger treatment is the appropriate treatment comparison. When controls were angered, then the anger-plus-media condition is appropriate to detect media exposure effects. Of course, a study code would need to be added to the synthesis to indicate whether participants for a particular study were angered; then analyses can be conducted to compare the media effect in studies in which participants were angered versus those in which they were not.

Another concern is how to represent the error term when computing effect sizes from analysis of variance designs that varied multiple factors. When the irrelevant factors in a design are individual difference variables (e.g., measures of viewers' dispositional level of aggressiveness), a one-way experimental design can be approximated to compare the two groups of interest by recalculating the error terms of the analysis of variance to include the irrelevant factors. The composite error term can then be used in the effect size calculations (see Johnson & Eagly, 2000; Morris & DeShon, 1997). However, when other factors in the design represent powerful experimental manipulations that ordinarily do not coexist with the relation of interest, they may not be appropriate to include in the error term. The general rule for error terms is that all effect sizes in the review should be based on the same sources of variability.

These complex issues indicate how important it is that reviewers understand a study's experimental design before estimating the size

of its effects. A useful strategy for reviewers, suggested by Johnson and Eagly (2000), is to generate a packet of descriptions of the sources of variance in common experimental designs in the reviewed literature (for summaries of designs, see Maxwell & Delaney, 1999; Myers & Well, 1991). A number of articles provide additional advice on the statistical issues to consider when aggregating data from different types of experimental designs (e.g., Dunlap, Cortina, Vaslow, & Burke, 1996; Morris & DeShon, 1997, 2002).

### *Corrections for Effect Size Bias*

We noted that the effect size statistics calculated from individual studies are often corrected because they are biased estimators of the population effect size. A variety of other corrections are available to adjust individual effect sizes for bias, artifact, or error. Some of the most elaborate adjustments have been proposed by Hunter and Schmidt (1994). They recommend adjusting for unreliability of measurement, artificial dichotomization of a continuous variable, imperfect construct validity, and range restriction. The goal of these corrections is to generate an idealized estimate of the magnitude of the true population effect size. Thus, the corrections are useful to demonstrate how large the true relation would be if it were not contaminated by these artifacts. The corrections are not appropriate to generate an estimate of the relation that was obtained in practice in a given research literature. Unfortunately, such adjustments typically are not an option in analyses of social psychological research because the original studies usually do not provide the information necessary to calculate adjustments.

### *Strategies for Nonreported Results*

A number of studies may not have reported sufficient statistical information to calculate effect sizes for the relation of interest. Several strategies exist to handle the resulting problem of missing data.

Sometimes reviewers fill in or impute values for missing effects. For example, an effect size of 0.00 could be entered when studies report that findings are not statistically significant. However, this procedure may be highly inaccurate—even large effects may not be significant when a study's sample size is small. It is also possible to replace missing effect sizes with a mean or some other estimated value (see Pigott, 1994). An alternate strategy, using a "vote-counting procedure," is possible when studies do not provide sufficient information to calculate an exact effect size but they do report the direction of the effect (Bushman, 1994; Bushman & Wang, 1996). This procedure calculates effect sizes from a tally of how many studies obtained a result in a given direction or how many studies obtained a statistically significant result. Regardless of the specific strategy used to address missing data, analysts should conduct "sensitivity" analyses to compare the findings from only the studies that generated an exact effect size with the findings from analyses that also include imputed values or analyses that are based on direction of effect.

## ANALYZING META-ANALYTIC DATA

The first step in analyzing meta-analytic data is to develop an understanding of the attributes of the data set. A visual display of the findings can be helpful. Some meta-analytic computer programs provide these automatically in forest plots and other displays. Stem and leaf plots are useful to convey the "big picture" from review findings. For example, Christensen (2003) used this technique to display the effect size findings from his synthesis of research relating people's self-esteem to their identification with various social groups (see Table 15.1). Each number on the right of the vertical bar signifies a study finding. The graph is read from left to right, with the number on the left of the bar representing the first numeral of the effect size and the number on the right representing the second numeral. Thus, this plot indicates that most study findings cluster around $r = .10$ and .20; people have higher self-esteem when they identify more strongly with political, religious, school, and racial and ethnic groups.

**Table 15.1** Distribution of Effect Sizes (rs) on a Stem and Leaf Plot from Christensen (2003)

| *Stem* | *Leaf* |
|---|---|
| +.7 | 0 |
| +.6 | |
| +.5 | 23 |
| +.4 | 00447789 |
| +.3 | 01111222245567 |
| +.2 | 000000001122333344557778999 9 |
| +.1 | 000011122224455566677777899 99 |
| +.0 | 123333444445556667789 99 |
| −.0 | 974332 |
| −.1 | 74 |
| −.2 | 90 |
| −.3 | 7 |

SOURCE: Adapted from Christensen (2003).

NOTE: Each correlation is composed of one stem and one leaf. The numbers on the left side are the stems or the first numeral in the correlation. The numbers on the right side are the leaves or the second numeral in the correlation. Each numeral on the right side is one leaf and indicates a unique correlation. Positive numbers suggest that higher levels of social identification are associated with higher levels of self-esteem.

The shape of the distribution of effects in the plot can reveal much about the sample of studies (Greenhouse & Iyengar, 1994; Light, Singer, & Willett, 1994). Note that the overall pattern in Table 15.1 generally follows a normal distribution. If small effect sizes did not appear in the distribution—in particular, if no small effects were obtained from studies with small sample sizes—then this could indicate that the selection of studies is biased to represent statistically significant findings. Such a bias could emerge if statistically significant results were more likely to be published. Of course, when reviewers have adequately sampled sources of unpublished data and included unpublished findings in the data set, publication pressures are less likely to distort the results of the review.

Displaying the distribution of effect sizes also can be informative about outlying, extreme effects that are unrepresentative of the sample of studies and may even be spurious. Such extreme values require close scrutiny because they have a disproportionate influence on the means, variances, and other statistics. With careful inspection, reviewers may be able to identify how studies with deviant findings differ from the rest of the sample. Some meta-analysts advocate removing the most extreme outliers or adjusting them to more moderate values prior to conducting additional analyses (Hedges & Olkin, 1985; see Johnson, 1993, for a computer application).

Special statistical procedures and computer programs are required to analyze meta-analytic data. Effect sizes differ from data points in primary research because each effect is associated with its own unique variance. In addition, meta-analytic statistical analyses provide useful information that is not generated by conventional statistics with primary data. That is, meta-analytic techniques provide estimates of the consistency or homogeneity of effect sizes across studies.

### *Step 1: Choosing a Model*

Synthesists have a choice about whether to use fixed-effects or random-effects procedures in the analysis. Although many published meta-analyses have assumed fixed-effects models, and these are computationally simpler than random-effects models, new computing programs are available that offer both procedures (e.g., Borenstein & Rothstein, 1999;

Lipsey & Wilson's, 2001, adaptation of *SPSS*; Wang & Bushman, 1999).[2]

From the perspective of statistical inference, the choice of model is similar to the choice of whether to use fixed-effects or random-effects models in analysis of variance. Fixed-effects models are appropriate if meta-analysts wish to make inferences about the effect-size parameters in the reviewed studies or about an identical set (Hedges & Vevea, 1998). In these models, the study effects estimate the population effect with the only error being from the random sampling of participants within the studies. Random-effects models allow inferences that generalize beyond the specific set of reviewed studies to a broader population. These models are appropriate when random differences are likely to exist between studies, and these differences include more than just participant-level sampling error (e.g., random variations in experimental procedures and settings).

From a practical standpoint, a review's conclusions can be affected by the decision to use a fixed- or random-effects model. Fixed-effects models usually have greater power to detect effects and yield smaller confidence intervals, whereas random-effects models tend to be more conservative. However, neither model appears especially robust when its assumptions are violated (Field, 2001; Overton, 1998). Thus, it is important to select the appropriate model for the data and the research question. In general, we recommend conducting sensitivity analyses, which involve calculating both models and comparing their results (Lipsey & Wilson, 2001).

### *Step 2: Estimating Means and Variability*

An initial step in meta-analysis is to aggregate effect sizes across studies to determine the overall strength of the relation between variables. To illustrate the specific procedures involved in aggregating and analyzing effects, we will present a simple fixed-effects analysis on a standardized mean difference statistic (Hedges & Olkin, 1985).

Study outcomes are combined by averaging the $d$ values across the number of studies, $k$. Each study outcome, $d_i$, is weighted by the reciprocal of its variance, and a mean weighted effect size, $d.$, is computed as a weighted average of the individual studies' effect sizes. Typically, this is calculated as

$$d. = \frac{\sum_{i=1}^{k} w_i d_i}{\sum_{i=1}^{k} w_i},$$

where the weights for each study outcome, $w_i$, are defined as

$$w_i = \frac{1}{v_i} = \frac{2(n_{iE} + n_{iC})\, n_{iE}\, n_{iC}}{2(n_{iE} + n_{iC})^2 + n_{iE}\, n_{iC}\, d_i^2},$$

in which $v_i$ is equal to the variance of the effect size estimate. This weighting procedure gives greater influence to the studies with more reliably estimated outcomes, which in general are those with the larger sample sizes. Similar weighting procedures are used when aggregating correlation effect size estimates (Hedges & Olkin, 1985; Rosenthal, 1991).

To help interpret the effect size, a confidence interval can be computed around the mean:

$$d. \pm 1.96 \sqrt{\text{variance}},$$

where 1.96 is the unit-normal value for a 95% confidence interval and the variance is defined as the reciprocal of the sum of the weights (as defined in the equation for weights given above). If the confidence interval includes zero, then it can be concluded that the relation of interest is not statistically significant.



Meta-analytic statistics also provide an estimate of the homogeneity or consistency of effect sizes. The test statistic, $Q$, evaluates the hypothesis that the effect sizes are consistent (Hedges & Olkin, 1985). This statistic is calculated as

$$Q = \sum_{i=1}^{k} w_i (d_i - d.)^2,$$

$Q$ has an approximate $\chi^2$ distribution with $k - 1$ degrees of freedom. A significant $Q$ rejects the null hypothesis that the study outcomes differ only by unsystematic sampling error. If the $Q$ statistic is not significant, it is likely that the study outcomes all represent a common population parameter. However, nonsignificant tests also could be due to statistical power that is inadequate to detect variability when the number of effects is small or when they come from studies with small samples. Thus, a large but nonsignificant $Q$ statistic can suggest variability in study outcomes (Johnson & Turco, 1992).

### *Step 3: Investigating Possible Moderators*

A large or significant $Q$ statistic suggests that the variability in effect sizes is more than just sampling error. Researchers may believe that this variability is systematic and that the relation of interest increases or decreases systematically with a third, moderating variable. For example, the relation between media violence and viewer aggression might be stronger in countries like the United States, in which a history of high exposure to media violence could make people more accepting of violence, than in countries with less exposure.

To evaluate moderating effects, synthesists can use an analog to analysis of variance in order to examine moderators that are assessed categorically (e.g., nationality of study participants) or an analog to multiple regression to examine moderators assessed on a continuous scale (e.g., year of publication, mean age of study participants). Categorical models divide studies into groups based on the moderator of interest and evaluate whether differences between groups can account for the variability in study outcomes apparent in the significant $Q$ statistic. Continuous models use regression procedures to estimate whether some third, continuous variable is associated with the size of the relation between the two variables of interest. Both of these approaches yield a test of the systematic variance associated with each moderator (i.e., $Q_B$ in categorical models, a test based on the regression coefficient in continuous models) and a test of the remaining, unexplained variability ($Q_W$ in categorical models, $Q_E$ in continuous models; see Hedges & Olkin, 1985). Significant remaining variability in moderator analyses violates the fixed-effects assumption that only random participant-level variability will remain after accounting for moderators. In this case, it is probably appropriate to use random-effects models that include a term to represent the random variability associated with studies. A relatively straightforward adaptation of such an approach is a mixed-effects model, which first accounts for the systematic effects of moderators through fixed-effects analyses and then uses random-effects procedures to estimate the remaining unmeasured random effect and the participant-level sampling error (Shadish & Haddock, 1994).

Although research syntheses is ideally suited to testing moderating relations, it generally is less effective at testing for mediators, or the extent to which one factor affects another factor through some intervening process (see Hoyle & Robinson, Chapter 10, this volume, for a discussion of mediating and moderating relations). When the studies in a research literature did not assess a common mediating process, it is not possible

to evaluate mediation at an aggregate level. However, studies sometimes vary moderators that provide insight into mediating processes (see Shadish, 1996, for strategies to test meta-analytic mediation).

### *Step 4: Reporting Findings*

To have maximal impact, research syntheses need to provide clear and coherent answers to research questions that are compelling to a broad audience. It can be challenging for synthesists to present findings in a simple, clear fashion after they have conducted complex data analyses involving myriad study details. Yet, some of the most useful information about a research literature is simply descriptive, such as graphs displaying study outcomes and tables systematically describing study attributes. Also useful are tables reporting central tendencies of the aggregated effects and estimates of their variability, including (a) unweighted mean effect sizes, (b) weighted mean effect sizes and variability estimates for fixed- and random-effects models, (c) confidence intervals, and (d) sample sizes (see Rosenthal, 1995). In general, reports of findings should be organized to provide clear answers to the questions that initially motivated the synthesis.

## DRAWING CONCLUSIONS FROM META-ANALYSES

### *Interpreting Effect Size Statistics*

In science as in life, size matters. Large effects are sometimes thought to be important ones. However, variability of the effects also matters. A mean effect that is based on highly variable study outcomes may not converge on the truth as much as obscure it. When variability is marked, understanding of the factors that moderate the size of an effect will likely be more meaningful than interpreting the size of the effect.

A much-respected statistician, Jacob Cohen (1992), provided informal guidelines to interpret the size of effects. He believed that small effects, $d$s < 0.20 or $r$s < .10, are typical of findings in personality, social, and clinical psychology. He termed large effects, $d$s > 0.80 or $r$s > .50, more typical of sociology, economics, and experimental and physiological psychology; these fields apparently investigate potent variables or use methods with strong experimental control. This observation makes the important point that the size of an effect is dependent on both the impact of treatment variables and the amount of experimental control. Finally, Cohen described medium-sized effects, $d$s = 0.50 or $r$s = .25, as being of sufficient magnitude that they are likely to be apparent to a careful observer in daily life. Although these guidelines are widely cited, Cohen cautioned that they represent subjective assessments that should be used only in the absence of a better basis for interpretation.

One common way of interpreting effect magnitude is to calculate effect size in the form of a squared correlation coefficient. The $r^2$ statistic is interpreted as the percentage of variability that is explained by an effect of a given magnitude. Thus, an effect size of $d$ = 0.40 or $r$ = .20 accounts for about 4% of the variance. However, it may not be appropriate to apply this method to interpret an aggregated correlation across a sample of studies. When some of the studies reported a positive relation and others a negative relation, the aggregated effect will necessarily underestimate the extent to which the two variables are related.

Empirical evidence of the effect sizes typical of social psychology was provided by Richard, Bond, and Stokes-Zoota (2001, in press) in their review of 322 meta-analyses conducted on social psychological topics. Aggregating across the 474 effect sizes provided by the reviews, the mean effect size in social psychology proved to be $r$ = .21. The largest magnitude effects emerged from group discussion, in that hearing others' arguments caused group members to shift their attitudes to be more extreme ($r$ = .75, $k$ = 12), and from assessments of the reliability of measures, in that measures tend to consistently yield similar scores ($r$ = .75, $k$ = 154).

Even if social psychologists can typically expect to obtain small-to-medium-sized effects, this does not mean that social psychological phenomena are inconsequential. Small effects can be impressive. For example, an effect that is small for any one observation can magnify over time as effects cumulate. Abelson (1985) made this point in the context of baseball, by demonstrating that a batter's skill has only a small impact on what happens at any single time at bat: Even for the best batters, the typical outcome is to make an out. However, batters' skill has important effects on team performance when cumulated across a game and a season. Similarly, small discriminatory biases against female employees may have little impact on any one salary or promotion decision, but across a whole career of such decisions, women may end up with substantially lower salaries and lower status positions than men (Martell, Lane, & Emrich, 1996). Small effects also can be meaningful when aggregated across large numbers of people. For example, Bushman and Anderson (2001) noted that if media violence increased aggressiveness in only 1% of viewers, the substantial numbers of viewers in the United States means that large numbers of people would be incited to act more aggressively.

Another approach to interpreting effect size values is to convert them into a more intuitively meaningful metric. For example, Rosenthal and Rubin's (1983) binomial effect size display translates correlation of effect sizes into an indicator of "success rate" in an experimental versus a control group. Success rate is meant to be defined broadly and in a given synthesis might represent, for example, the incidence of viewer aggressiveness. To understand this approach, imagine that a "success threshold" is set at the median of the distribution of scores on a dependent variable for both treatment and control groups. Then the groups are separated, and the proportion in each group is calculated that is above the overall success threshold. Interestingly, the success differential between the two groups is always equal to the correlation effect size between the two groups. Thus, if a correlation is $r$ = .2 between an outcome measure (e.g., aggressive behavior) and whether people were in an experimental group (e.g., watched a violent video) or control group (watched a nonviolent video), then the experimental group's "success" rate of acting aggressively will be 20% higher than the control group's. This demonstration of treatment impact illustrates the importance of effects that might otherwise be interpreted as relatively small.

Finally, another way to interpret the magnitude of effects is to compare them with effect sizes in similar domains. For example, Bushman and Anderson (2001) provided a framework for readers to understand the practical value of media violence effects by noting that they are larger than the relation between (a) condom use and sexually transmitted HIV, (b) passive smoking at work and lung cancer, (c) exposure to lead and IQ scores in children, (d) calcium intake and bone mass, and (e) homework and academic achievement. These contexts provide a framework to better understand the impact of exposure to media violence.

### *The Impact of Syntheses Findings and the Future of Research Syntheses*

Although research syntheses are sometimes treated as simply summaries of past research, they also are likely to spark additional investigation in several ways (Eagly & Wood, 1994). For example, a synthesis that provides a

summary evaluation of existing knowledge can be an initial step in generating novel hypotheses about the mechanisms through which the effect emerges and the boundary conditions for the effect. Syntheses also spur additional investigation by identifying new phenomena that can be evaluated only across studies (e.g., effects of year of publication).

Another reason for additional investigation is to validate synthesis conclusions when the findings of the synthesis are correlational. Findings are correlational when the original studies used correlational designs or when the synthesis used study-level variables (e.g., year of study publication) as moderators of the effect of interest. Using study-level variables as moderators can be highly informative but, like all correlational strategies, raises questions about how to interpret the results. For example, to demonstrate that people act on their intentions primarily when habits have not developed, Ouellette and Wood (1998) divided studies into groups reflecting whether habits would be likely to develop with the behaviors investigated in the studies. Give that the studies in each group may have varied on a number of dimensions from those in other groups, it is not clear how to interpret any between-group differences. Thus, the researchers conducted an additional primary study to demonstrate that intention effects emerge primarily in the absence of habit. Because of the importance of the primary research findings to interpreting the synthesis results, Ouellette and Wood reported the two studies in a single article. In this way, the validity of synthesis conclusions can be enhanced through pairing with the results of other research methods with complementary strengths.

In general, the future of research synthesis as a methodological technique is ensured by the sheer amount of research data being produced each year. As the amount of information increases, so does the need for systematic procedures to distill this information. Advances in information technology are likely to further facilitate use of research syntheses, including the development of electronic depositories to store and codify research findings (e.g., the Cochrane Collaboration's, 2002, medical research archive). Given the power of the technique and the availability of new computer technologies to support it, research syntheses is likely to remain a popular tool for summarizing existing findings and testing novel hypotheses.

## NOTES

1. This calculation is based on a mean effect size of $r = .21$ and a median study sample size of $n = 127$. These estimates were obtained from 322 syntheses in social psychology, which yielded 474 effect sizes (Richard, Bond, & Stokes-Zooka, 2001, in press; also Charles Bond, personal communication, November, 2002). Note further that the analysis assumes a fixed-effects model, which is described in the section "Analyzing Meta-Analytic Data."

2. Lipsey and Wilson (2001) have made their computer programs available on the Internet. Copies of the program to calculate effect sizes (using Microsoft's Excel spreadsheet program) and the program to analyze effect sizes (using SPSS/Win Version 6.1) can be downloaded at the following URL address: http://mason.gmu.edu/~dwilsonb/home.html



## REFERENCES


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

Archer, J. (2000). Sex differences in aggression between heterosexual partners: A meta-analytic review. *Psychological Bulletin, 126*, 651-680.

Becker, B. J., & Schram, C. M. (1994). Examining explanatory models through research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 357-381). New York: Russell Sage.

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

Bem, D. J., Palmer, J., & Broughton, R. S. (2001). Updating the ganzfeld database: A victim of its own success? *Journal of Parapsychology, 65*, 207-218.

Borenstein, M., & Rothstein, H. (1999). *Comprehensive meta-analysis: A computer program for research synthesis*. Englewood Cliffs, NJ: Biostat.

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

Bushman, B. J., & Anderson, C. A. (2001). Media violence and the American public: Scientific facts versus media misinformation. *American Psychologist, 56*, 477-489.

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.

Bushman, B. J., & Wells, G. L. (2001). Narrative impressions of literature: The availability bias and the corrective properties of meta-analytic approaches. *Personality and Social Psychology Bulletin, 27*, 1123-1130.

Christensen, P. N. (2003). *Motivational connections between the group and the self: A review with meta-analytic support for the relationship between social identification and self-esteem*. Unpublished manuscript, San Diego State University, San Diego, CA.

Christensen, P. N., & Wood, W. (in press). Effects of media violence on viewers' aggression in unconstrained social interaction: An updated meta-analysis. In R. Preiss, B. Gayle, N. Burrell, M. Allen, & J. Bryant (Eds.), *Mass media effects research: Advances through meta-analysis*. Hillsdale, NJ: Lawrence Erlbaum.

Cochrane Collaboration (2002). Retrieved December 11, 2002, from http://www.update-software.com/collaboration/

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

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

Cook, T. D., & Leviton, L. C. (1980). Reviewing the literature: A comparison of traditional methods with meta-analysis. *Journal of Personality, 48*, 449-472.

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

Costa, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology, 81*, 322-331.

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

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

Eagly, A. H., & Wood, W. (1991). Explaining sex differences in social behavior: A meta-analytic perspective. *Personality and Social Psychology Bulletin, 17,* 306-315.

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

Felson, R. B. (1996). Mass media effects on violent behavior. *Annual Review of Sociology, 22,* 103-128.

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

Freedman, J. L. (1988). Television violence and aggression: What the evidence shows. In S. Oskamp (Ed.), *Applied social psychology annual: Television as a social issue* (Vol. 8, pp. 144-162). Newbury Park, CA: Sage.

Frieze, I. H. (2000). Violence in close relationships–development of a research area: Comment on Archer. *Psychological Bulletin, 126,* 681-684.

Geen, R. G. (1998). Aggression and antisocial behavior. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 2, pp. 317-356). Boston: McGraw-Hill.

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

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

Heath, L., Bresolin, L. B., & Rinaldi, R. C. (1989). Effects of media violence on children. *Archives of General Psychiatry, 46,* 376-379.

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

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

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

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

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

Johnson, B. T., Carey, M. P., & Muellerleile, P. A. (1997, February 5). Large trials versus meta-analysis of smaller trials. *Journal of the American Medical Association, 277,* p. 377.

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

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

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

Kraemer, H. C., Gardner, C., Brooks, J. O., III, & Yesavage, J. A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods, 3,* 23-31.



Light, R. J., Singer, J. D., & Willett, J. B. (1994). The visual presentation and interpretation of meta-analysis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 439-453). New York: Russell Sage.

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

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis. Applied social research methods series* (Vol. 49). Thousand Oaks, CA: Sage.

Martell, R. F., Lane, D. M., & Emrich, C. (1996). Male-female differences: A computer simulation. *American Psychologist, 51,* 157-158.

Maxwell, S. E., & Delaney, H. D. (1999). *Designing experiments and analyzing data: A model comparison perspective.* Mahwah, NJ: Lawrence Erlbaum.

McGuire, W. J. (1986). The myth of massive media impact: Savagings and salvagings. In G. Comstock (Ed.), *Public communication and behavior* (Vol. 1, pp. 173-257). San Diego: Academic Press.

Milton, J., & Wiseman, R. (1999). Does psi exist? Lack of replication of an anomalous process of information transfer. *Psychological Bulletin, 125,* 387-391.

Morris, S. B., & DeShon, R. P. (1997). Correcting effect sizes computed from factorial ANOVA 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 designs. *Psychological Methods, 7,* 105-125.

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

O'Leary, K. D. (2000). Are women really more aggressive than men in intimate relationships? Comment on Archer (2000). *Psychological Bulletin, 126,* 685-689.

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

Ouellette, J. A., & Wood, W. (1998). Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychological Bulletin, 124,* 54-74.

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

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

Pigott, T. D. (1994). Methods for handling missing data in research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 163-176). New York: Russell Sage.

Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2001). That's completely obvious . . . and important: Lay judgments of social psychological findings. *Personality and Social Psychology Bulletin, 27,* 497-505.

Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (in press). One hundred years of social psychology quantitatively described. *Review of General Psychology.*

Rosenthal, R. (1991). *Meta-analytic procedures for social research: Applied social research methods series* (Vol. 6). Thousand Oaks, CA: Sage.

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

Rosenthal, R., & Rubin, D. B. (1983). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74,* 166-169.

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

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., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261-281). New York: Russell Sage.

Shadish, W. R., Robinson, L., & Lu, C. (1999). *ES: Effect size calculator.* St. Paul, MN: Assessment Systems Corp.

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

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

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

Twenge, J. M. (2002). Birth cohort, social change, and personality: The interplay of dysphoria and individualism in the 20th century. In D. Cervone & W. Mischel (Eds.), *Advances in personality science* (pp. 196-218). New York: Guilford.

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

Wilson, D. B., & Lipsey, M. W. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. *Psychological Methods, 6,* 413-429.

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

Wood, W., Wong, F. Y., & Chachere, J. G. (1991). Effects of media violence on viewers' aggression in unconstrained social interaction. *Psychological Bulletin, 109,* 371-383.

# Part IV

# Emerging Interdisciplinary Approaches: The Integration of Social Psychology and Other Disciplines