*American Journal of Community Psychology, Vol. 19, No. 3, 1991*

*Methodology*

# A Practitioner's Guide to Meta-Analysis[1]

**Joseph A. Durlak**[2]
*Loyola University Chicago*

**Mark W. Lipsey**
*Claremont Graduate School*

*Describes the application to community issues of the meta-analytic research strategies increasingly used in many field of psychology. First, we highlight the potential value of meta-analysis to community research. Second, we describe six major steps involved in conducting an effective meta-analysis. These steps include formulating the initial research question(s), locating relevant studies, abstracting critical information from each study, and presenting, analyzing, and interpreting the resultant data. In this guide, the major aspects of meta-analysis are discussed with particular emphasis on the procedures that are most critical to the validity of its conclusions. Greater familiarity with the techniques, issues, and potential of meta-analysis may stimulate investigators to make more effective use of this powerful approach to integrating research in community psychology.*

Since the work of Glass (1976), meta-analysis has become commonplace within the behavioral and social sciences. By 1984, Kulik (1984) estimated that approximately 300 meta-analytic reviews had appeared, and we estimate that at least twice that number have been conducted since then.

The appearance of so many meta-analyses indicates the enthusiasm with which social scientists have embraced this research strategy. At the same time, the technique has engendered considerable controversy. Many

[1]The authors wish to thank Georg Matt, Emil Posavac, Ed Seidman, William Shadish, and two anonymous reviewers for helpful comments on this manuscript.
[2]All correspondence should be sent to Joseph A. Durlak, Department of Psychology, 6525 N. Sheridan Road, Loyola University, Chicago, Illinois 60626.



0091-0562/91/0600-0291$06.50/0 

researchers have objected to the conclusions and interpretations offered in specific reviews, or have emphasized the limitations and problems involved in applying meta-analytic techniques generally (e.g., Abrami, Cohen, & d'Apollonia, 1988; Bullock & Svyantek, 1985; Garfield, 1983; Guzzo, Jackson, & Katzell, 1987; Nurius & Yeaton, 1987; Strube, Gardner, & Hartmann, 1985; Wanous, Sullivan, & Malinak, 1989; Wilson, 1985).

Three additional issues contribute to potential confusion regarding meta-analysis. One, meta-analysis is not an easy technique to understand or to apply, and many researchers are unfamiliar with its procedures. Jackson (1984) found that more than two-thirds of a sample of *Psychological Bulletin* authors were not familiar with meta-analysis. Two, meta-analysis is an evolving research strategy with refinements and technical improvements being offered at a pace that makes it difficult to keep abreast of developments in the field. Third, and particularly disconcerting, is the fact that meta-analyses regularly offer different findings and conclusions than conventional reviews of the same research literature. At times, separate meta-analyses of similar literatures have also reached different conclusions. Readers who are not cognizant of meta-analytic goals and procedures may find it challenging to reconcile such divergent results.

This paper provides a practitioner's guide to meta-analysis to help researchers understand and apply this technique. We do not pretend to offer the last word on meta-analysis nor to advance its theoretical foundations. Rather, our intent is to provide an overview of the aspects of a meta-analytic review that are most important in practice and to apprise readers of the latest developments in the field. Our commentary is replete with references so that the reader can consult additional sources for further information.

## THE UTILITY OF META-ANALYSIS

Despite the widespread application of meta-analysis in many areas of behavioral and social science, at the time of this writing we are aware of only one meta-analysis that has appeared in the major publication outlets in community psychology (Okun & Stock, 1987). Community researchers are thus overlooking a potentially valuable and powerful analytic tool for understanding community issues. We begin, therefore, with a brief overview of what meta-analysis can accomplish.

Meta-analysis is one of several techniques that can be used to review a research literature (Cooper, 1984). It has the same general goals as most literature reviews: to summarize the results in a research domain, to investigate how findings might vary as a function of key characteristics of the

reviewed studies, to offer recommendations to improve future research, and to draw out the implications of research for policy and practice. The distinctive features of meta-analysis compared to other review strategies are the quantitative representation of key research findings in the studies reviewed and the statistical analysis of the distribution of findings across studies and the relationship of study features to those findings.

In comparison to conventional narrative reviews, there are several advantages to the statistical integration of study findings that is achieved in meta-analysis. First, meta-analysis makes full use of the outcome information from studies rather than reducing data to all-or-none categories such as positive or null results based upon statistical significance (often referred to as "vote counting" in reviews). Instead, meta-analysis examines both the direction and the magnitude of effects obtained in each study as well as the distribution of effects across studies. As a result, the meta-analyst can test explicitly the "groupwise" statistical significance of the set of study outcomes, a tactic not customarily employed in conventional reviews.

Second, quantitative coding of study characteristics permits a researcher to keep track of a large amount of potentially important information and then conduct a more detailed breakdown of this data than is easily managed using conventional review techniques. Meta-analysis can easily handle scores of variables from hundreds of studies. Most important, the database that results from coding study characteristics can be analyzed using multivariate techniques that make it possible for the reviewer to examine a wider range of relationships, interactions, and the like than less differentiated and more qualitative techniques permit.

Third, by treating the review process itself as a type of research study, meta-analysis can make very explicit and visible the sample of studies selected, the aspects of the studies attended to, and the basis for the reviewer's claims about the outcomes and relations found. In this way, research integration becomes subject to replication and to criticism regarding the assumptions, judgments, and procedures applied by a particular reviewer. A reader, therefore, is required to take much less on faith and guesswork when interpreting a meta-analysis than a conventional review.

Finally, it should be of particular interest to community psychology that many social policy implications can emanate from meta-analytic reviews (Cordray, 1990; Kiesler, 1985). For example, consider the dilemmas faced by policymakers regarding training, service, or research priorities in the area of prevention. Policymakers might want to know: How effective are preventive programs? What is the relative effectiveness of primary vs. secondary prevention in achieving different goals? What program components are responsible for different outcomes? Are there particularly effective or exemplary interventions that could serve as models for others? How do different

programs compare in terms of cost effectiveness? Finally, what gaps exist in current research and practice that would suggest priorities for future work? It is likely that any type of research review could produce some relevant information in response to the above questions but, as Kiesler (1985) observed, the statistical emphasis in meta-analytic strategies may yield the most compelling form of data. Numbers often impress. Those involved in formulating public policy might be more influenced by the careful procedures and replicable quantitative results of meta-analysis than by scientific writing that appears to be more subjective or inferential in nature.

At the same time, we note the possibility of abuse or misuse of meta-analysis. Poorly conducted meta-analyses can provide misleading information and even well-conducted ones are subject to misuse. Unless great care is taken in reporting and interpreting the results of a meta-analysis, erroneous conclusions might be formed and unjustified actions with far-reaching implications might be undertaken.

## TYPES OF META-ANALYSIS

Many forms of meta-analysis have been distinguished, varying in terms of purpose, unit of analysis, treatment of study variation, and products (Bangert-Drowns, 1986). For present purposes, we distinguish two major categories, each with two subcategories: group contrast meta-analysis, which can involve either treatment effectiveness or group differences issues, and correlational association meta-analysis, which involves either test validity or issues of covariation among other types of variables.

*Treatment Effectiveness Meta-Analysis.* Treatment effectiveness meta-analysis surveys research on a defined treatment domain to summarize the effects of treatment and the relationship of those effects to such factors as the nature of the subjects, the amount of treatment, and the specific treatment modality. The primary statistical indicator for this type of meta-analysis is characteristically a standardized effect size representing the magnitude of the difference between the treatment and control group mean on a relevant outcome measure.

A classic example of treatment effectiveness meta-analysis is the Smith, Glass, and Miller (1980) meta-analysis of psychotherapy outcome. Since then hundreds of others have been conducted, from cognitive therapy for depression (Dobson, 1989) to effects of examiner familiarity on children's test performance (Fuchs & Fuchs, 1989). In community psychology, treatment effectiveness meta-analysis might be used to assess the effects of prevention programs educational campaigns, services to special-need populations, alternatives to traditional treatment, and the like.

*Group Differences Meta-Analysis.* The second type of meta-analysis surveys research on differences between more-or-less naturally occurring groups, e.g., males and females, children with and without fathers in the home, and so forth (e.g., Jorm, 1987; Martocchio & O'Leary, 1989; Stevenson & Black, 1988). Like treatment effectiveness meta-analysis, this approach generally relies on a statistical indicator of effect size that represents the difference between the means of the groups on a variable of interest. Group differences meta-analysis in community psychology might be used to investigate such issues as the differences in use of mental health services by different ethnic groups, risk factors distinguishing abusive from non-abusive families, depression in various age groups, and so forth.

*Test Validity Meta-Analysis.* Where the above two types of meta-analysis revolve around a group contrast, the remaining two focus on correlational relationships. One common application involves research on test validity by investigating the correlation between a test or measure and a criterion variable. Meta-analysis provides an effective way to summarize a number of such studies and investigate the influence of differences among them. The primary statistical indicator for this type of meta-analysis is the product-moment correlation. While the most extensive use of this technique has been in the domain of personnel testing (Hunter, Schmidt, & Jackson, 1982), it is applicable to assessing the validity of any measurement procedure (e.g., Achenbach, McConaughy, & Howell, 1987; Okun & Stock, 1987; Skiba, 1989). In community psychology, this form of meta-analysis could be used to integrate validity evidence for any measures with a sufficient psychometric research base, e.g., social support, life stress scales, and environmental assessments.

*Variable Covariation Meta-Analysis.* A second type of meta-analysis using the correlation coefficient statistical index involves research on the covariation of two or more variables of interest (e.g., Loeber & Stouthamer-Loeber, 1986; Suls & Wan, 1989; Tait, Padgett, & Baldwin, 1989). The variables included in meta-analysis of this type may be solely of practical interest or may be constructs of theoretical importance. This approach to meta-analysis might be used in community psychology to investigate such issues as the relationship between social support and perceived stress, income and utilization of social services, locus-of-control and community activism, and AIDS awareness and sexual behavior.

## STEPS IN META-ANALYSIS

There are six major steps in a meta-analysis, and at each one important choices must be made by the researcher. In this paper, we enumerate

these steps and describe the implications of different procedural choices with particular emphasis on those decisions that have important implications for the validity of meta-analytic results. Since meta-analysis is a research technique, its strengths and limitations in different circumstances can be evaluated using familiar methodological criteria (Wortman, 1983). In particular, meta-analysis is analogous to survey research with the difference that the unit that is sampled and "interviewed" is a written research report rather than a living, talking human being.

Table I lists the six major steps in a meta-analysis and summarizes the issues most important at each step. To borrow a frequently-used metaphor, the steps involved in meta-analysis can be viewed as links in a chain, and the finished product is only as good as the weakest link. For example, well-formulated research questions derived from careful scrutiny of past studies are important to launch the meta-analysis, but this step can be subverted unless a representative and nonbiased literature search is conducted. Similarly, the virtues of an effective literature search are negated if coding procedures are insufficient to capture the essence of research reports or if inappropriate statistical techniques are applied to the data. Therefore, each and every aspect of the meta-analysis is important.

## Formulating the Research Question(s)

The first step in a meta-analysis is deciding its purpose. Most meta-analyses in the behavioral sciences have been exploratory in nature and descriptive in intent. In this approach, the primary interest is in summarizing the findings on the focal relationship of interest, e.g., impact of interventions, group differences, and covariation between key variables, and describing the general characteristics of the relevant population of studies.

Descriptive meta-analyses have limited explanatory value. In contrast, meta-analyses that are hypothesis- or issue-driven represent a more probing approach to a research literature. In this approach, the meta-analyst offers either specific hypotheses *a priori* or at least identifies the important theoretical, conceptual, or procedural questions to be addressed. In such cases, the meta-analyst should begin the review by specifying the major issues pertinent to the literature. What rival theories predominate in the literature? What variables are believed to be important in determining the outcome of past research? Have there been differences in theory or technique that may distinguish different investigations? What controversies or disagreements exist in prior studies among the data or their interpretation?

**Table I.** Major Steps in a Meta-Analysis and Important Issues Relevant at Each Step

1. Formulating the research question(s)
   (a) Are specific research questions, formal hypotheses or the major variables of importance made explicit? Do such formulations rest upon prior work in the area?
   (b) Is the literature to be reviewed fully defined? Does the definition capture the important literature in the field?
   (c) Are both inclusionary and exclusionary criteria presented and are such criteria reasonable?
2. Literature search
   (a) Is a representative and unbiased sample of studies identified?
   (b) Has potential publication bias been estimated by including a sample of unpublished studies?
   (c) Have several different methods of searching the literature been used?
   (d) Is the number of relevant but nonusable studies presented?
   (e) Is a fail-safe *n* calculated to assess the robustness of obtained findings vis-à-vis the likelihood that not all relevant studies have been obtained?
   (f) Are all the sampled studies listed or available from the author?
3. Coding procedures
   (a) Are problems in coding procedures described?
   (b) Is the coding system available on request and does it contain criteria for coding potentially confusing study features?
   (c) Have proper estimations of interrater agreement been conducted and reported?
4. Index of effect
   (a) Has the pooled standard deviation been used to calculate individual effects in group difference meta-analyses?
   (b) Are all methods of calculating effect sizes described?
   (c) Is the procedure for dealing with "nonsignificant findings" made explicit? Is incidence of such findings made known? Are the implications of nonsignificant findings related to the conclusions and generalizations of the meta-analysis?
   (d) Are comparison groups described and coded, if appropriate?
5. Statistical analyses
   (a) Are effects adjusted for small sample bias?
   (b) Are outliers identified and examined for their potential heuristic benefit?
   (c) Has an appropriate unit of analysis been used? Has one effect been calculated per construct per study per research question to avoid confounding important constructs in the analyses?
   (d) Are appropriate weighting procedures used throughout the analyses?
   (e) Has a systematic and defensible approach been taken toward analyzing differences in effect sizes?
   (f) Are *a priori* hypotheses used to explore obtained differences in effects across studies?
   (g) Is sufficient attention given to the potential influence of methodological features?
   (h) Has the meta-analyst proposed a statistical model that correctly specifies the obtained data?
6. Conclusions and interpretations
   (a) Does the meta-analyst relate the issue of power to statistical findings?
   (b) Are conclusions restricted to the literature reviewed?
   (c) Is a table presented describing the characteristics of reviewed stufies so that missing information in the reporting of certain important study features is apparent?
   (d) Are appropriate qualifications offered in line with the state of the research being reviewed?

Whatever the approach, we stress that investigators should have a good understanding of past research in the area before beginning a meta-analysis. As Nurius and Yeaton indicated (1987), inspection of the relevant literature prior to the meta-analysis can help the reviewer sharpen the research questions and anticipate problems in obtaining the necessary evidence from existing studies. As a result, some hypotheses may be altered, additional variables may be considered, and, in general, procedures can be developed so that the vital research questions become more "answerable." It is possible, indeed, it is likely, that the meta-analyst will find that some important research questions cannot be answered because prior research is deficient in either execution or reporting of procedures and findings. However, such a conclusion would itself have important policy implications for future research.

Formulating the initial research question(s) also involves providing a clear definition of the population of research studies to be evaluated. This definition should be accompanied by detailed inclusionary and exclusionary criteria. The definition and inclusionary/exclusionary criteria permit a reader to identify the specific domain of research surveyed and, more importantly, to recognize any bias or limitations that may qualify the conclusions.

*Published vs. Unpublished Studies.* Certain recurring issues in defining the relevant population of studies for a meta-analysis deserve comment because of their ubiquity and potential for influencing the results. First is the question of whether to include unpublished studies, e.g., dissertations, convention papers, and technical reports. Meta-analysts often restrict their survey to published reports on the rationale that these represent the highest quality work, or as a matter of convenience, or without offering any explanation. This is dubious practice on scientific grounds. Published and unpublished studies in the same research domain frequently show many differences, not the least of which are different orders of magnitude in the key effect sizes or correlations upon which the meta-analysis focuses. Generally, published studies show larger effects than unpublished studies (Smith, 1980), presumably as a result of the greater motivation of researchers, reviewers, and editors to publish findings that reach statistical significance than those that do not. A meta-analysis that samples only published studies, therefore, runs considerable risk of bias in the reporting and interpretation of effects.

Given the likely difference between published and unpublished studies, some investigators will argue that the latter are categorically inferior (else they would have been published) and thus should be dismissed on validity grounds. There are many reasons why a study might not be published, however, and quality is only one. Moreover, a wide range of quality

is found in behavioral science research among both published and unpublished research (Lipsey, 1988). If quality criteria are to be imposed, they should be specified and applied equally to both published and unpublished research.

*Methodological Quality.* Indeed, whether to include all studies or only those that meet certain methodological criteria is another recurring issue in defining the population of studies for meta-analysis. Glass and his colleagues (Glass, McGaw, & Smith, 1981; Smith et al., 1980) have argued for liberal methodological criteria, preferring to discover empirically what relations exist between study methods and findings. This approach often yields surprising findings. In their psychotherapy meta-analysis, there was no appreciable difference in the effect sizes for studies with random vs. nonrandom assignment to experimental conditions (Smith et al., 1980). In other situations, especially where significant political controversy is involved, meta-analysts have chosen to survey only studies meeting the highest methodological standards (Slavin, 1986; Wortman & Bryant, 1985).

Clearly this issue must be resolved in part according to the purpose of the meta-analysis—examination of the relations between methodological characteristics and study findings will not be of interest to all meta-analysts. Given the potential influence of varying methods on study outcome, however, they cannot be ignored. To the extent that variation in important method characteristics is not restricted by inclusion criteria, they should be coded in the meta-analysis. In this way, their role can be assessed and any bias can be discovered rather than simply being assumed.

It is important to keep in mind that there are no absolute standards available to judge design quality or methodological rigor. Different reviewers can and do employ different criteria. Whether methodological quality is a basis for exclusion from the meta-analysis or a basis for coding within the meta-analysis, therefore, it is vital that all the criteria involved in its assessment be specified. This permits a reader to judge both their appropriateness and their effect on the reported results. Schemes for rating methodological quality can be found in Chalmers et al. (1981) and Gibbs (1989).

It is also important to recognize that methodological quality is a multi-dimensional concept. A common practice is to assess it on the basis of only a few criteria, usually emphasizing ratings of internal validity or, perhaps, all four of the conventional validity categories (Cook & Campbell, 1979). In a given research domain, however, there are many particulars of sampling, design, measurement, and implementation that may have far more important influences on outcomes than broad quality ratings or categories. Potential interactions among the methodological criteria and study findings also bear consideration. For example, random vs. nonrandom as-

signment to conditions might not relate to outcome in small laboratory studies, but it may be important in field studies.

*Time Frame for Search.* Another issue involves the period of time covered in the literature search. Often meta-analysts chose the survey period arbitrarily. Instead, a convincing rationale for confining the search to a particular period should be presented. For example, reviewers may show that there are few eligible studies appearing before or after a particular date, or may argue that only a distinct period of time is relevant because of the research paradigms used, the social context, or the nature of the problem definition.

In summary, an important principle in defining the literature to be surveyed is that the criteria for including and excluding studies from a meta-analysis be made explicit and be justified. This practice assures that the studies surveyed are both appropriate and representative, as well as communicating clearly to readers the exact domain under investigation. Failure to define the relevant inclusionary and exclusionary criteria can easily lead to conflicting findings and unnecessary confusion among meta-analysts who are purporting to examine the same literature (see Abrami et al., 1988).

## Literature Search

Once a meta-analyst has developed a clear and complete definition of the research literature of interest, the next step involves identifying and obtaining relevant studies. At the outset, we must emphasize that a truly comprehensive review of all relevant investigations is probably impossible, particularly with regard to unpublished studies.

Thus, the major goal of any literature search should be to implement a search strategy that yields a representative and unbiased sample of relevant studies. This goal is not as easy to achieve as it might appear. Some studies are relatively simple to find, others quite difficult. Nonrepresentativeness most often occurs when the reviewer's efforts to obtain hard-to-find studies are insufficient to produce the same proportionate yield achieved for easy-to-find ones. The result is that a significant portion of relevant work is omitted which may well differ in important ways from that which is found.

*Search Strategies.* Experience suggests that multiple search methods must be used in locating relevant studies for meta-analysis. We recommend that a combination of four search strategies be used: manual journal searches, examination of reference lists from reviews and identified studies, computer and manual searches of abstracting and indexing databases, and

contact with persons or organizations likely to have produced or to know of studies.

A manual search of journal contents is time-consuming but seems essential to secure representative studies. The journals that are searched manually should be listed to permit the reader to assess whether the main publication outlets for research in that area have been examined. As an obvious example, a review of community research that omitted either *The American Journal of Community Psychology* or *The Journal of Community Psychology* would be a serious gaffe. Previous research reviews are an excellent starting point for identifying relevant studies. The reference list of each relevant study should, in turn, also be examined since it may provide useful citations.

Computerized searches, e.g., through the Dialog databases, are notoriously unreliable as a means of identifying relevant literature, especially if the research topic is broad, e.g., psychotherapy, early intervention, prevention. For instance, Weisz, Weiss, Alicke, and Klotz (1987) used 21 key word terms in their computerized search for child and adolescent psychotherapy studies and obtained 1324 citations. Ultimately, only 108 studies were included in the meta-analysis, however, and only a portion of these had been identified by the computer search procedure. We discovered (Durlak, Lampman, & Wells, 1991) that only one of every three entries appearing in our computer-generated study lists was relevant and approximately two-thirds of the relevant studies were not picked up via the computer search. In short, computer searches tend to produce high numbers of false positives (irrelevant studies found) and an untold number of false negatives (relevant studies missed) depending on the area reviewed.

Despite their limitations, however, computer database searches do identify studies that are likely to be missed by other procedures and thus must be included in any comprehensive search strategy. Indeed, one of the most important lessons regarding literature searches is the surprisingly modest overlap in the eligible studies identified by different procedures and with different databases (Glass et al., 1981). Computer searches, therefore, should not be restricted to one database, e.g., PsychInfo, but should cover all those of potential relevance. Moreover, some computer searches are usefully supplemented by manual searches of the corresponding hard copy versions of the abstracts of interest, e.g., *Psychological Abstracts.* Most volumes of abstracts are organized into sections that deal with different broad areas or topics. Perusal of the sections pertinent to a meta-analysis topic can turn up eligible studies not located by a keyword computer search of that same database. This is particularly true for topics that are not associated with an established set of standard descriptors.

The greatest search challenge to the meta-analyst is identifying and retrieving relevant unpublished studies. Of these, dissertations present the least difficulty. A very high proportion of dissertations from U.S. universities are abstracted in *Dissertation Abstracts International* and are available through interlibrary loan or University Microfilms.

Although some searchable archives are available for unpublished conference papers and technical reports, e.g., ERIC, many unpublished studies can only be pursued by identifying and contacting the authors who conducted them or the institutions that may receive, sponsor, or collect them. Generally, there will be one or more communities or networks of researchers whose work is known to each other. Once contact is made with some members in the network, they can be asked to identify others, and so on until no new leads turn up. A meta-analyst might begin, for example, by reviewing presentations scheduled at relevant conferences, then follow-up with successive waves of mail or telephone surveys to solicit both studies and names of other relevant researchers. A similar approach can be used to obtain unpublished studies from organizations, e.g., government agencies, nonprofit and private service organizations, companies, and research firms.

*Unavailable and Excluded Studies.* Inevitably, some studies that meet the meta-analyst's inclusionary criteria will not appear in the final meta-analysis. Some studies will be overlooked and others that are obtained will not report sufficient statistical information to permit estimation of an effect. In the latter case, the most conservative approach is to code the remaining study characteristics of interest for excluded studies and show the extent to which they differ, if at all, from the group included in the final meta-analysis. If excluded without coding, however, the reader should be informed of the number of studies involved so that the representativeness of the surveyed literature can be assessed.

Since it is virtually impossible to conduct a comprehensive review of all research on a topic, meta-analysts have developed formulas to indicate how many additional studies would be needed to reduce obtained results to a level of nonsignificance (Orwin, 1983; Rosenthal, 1979). Calculation of this "fail-safe *n*" is useful in assessing the robustness of findings, but it must be interpreted cautiously. Rosenthal (1979) suggested that if the fail-safe *n* is not more than five times the number of reviewed studies plus ten, obtained findings are probably robust. However, this "five plus ten rule" is quite arbitrary. The actual number of studies missing from a meta-analysis depends on the research area and the care taken in the literature search. As Shadish, Doherty, and Montgomery (1989) discussed, determining the actual number of studies conducted but never published on a research topic is very difficult. Excluding dissertations, they estimated that there were as

many unpublished ($n$ = 112) as published ($n$ = 115) randomized, controlled studies of family and marital psychotherapy. In one of our own meta-analyses (Durlak et al., 1991), we located over 350 published studies of child psychotherapy, but estimated there were probably twice as many unpublished dissertations conducted on the same topic.

A prospective meta-analyst should not underestimate the time and expense involved in making a comprehensive search. It is not unusual to retrieve and screen between six to ten times more studies than are ultimately included in the meta-analysis. Incidentally, those studies that are eventually included should be listed in an appendix to the final report or at least be available from the meta-analyst on request.

## Coding of Studies

*Which Variables to Code.* In general, the major intent should be to code studies for all those substantive and methodological characteristics that might potentially influence study findings. However, there is tremendous discretion in determining such features. Meta-analyses have varied from coding just a few variables to coding over a hundred variables per study. That is one reason why the initial research questions guiding the meta-analysis are so important. Based upon prior research reports, theory, and methodological insight, the meta-analyst should be able to enumerate those variables that appear most important for a particular research area and then attempt to code for each. Furthermore, these variables should figure prominently in the subsequent statistical analyses to preclude capitalizing on chance findings when many variables are being analyzed.

It is impossible to specify all the variables that should be coded in any meta-analysis. The importance of particular variables varies across research areas. However, coding for the following has proven useful in previous reviews: study context, e.g., year, country, source of publication, and author characteristics, methodological characteristics, e.g., details of the sampling, research design, type of control group, measurement of independent and dependent variables, and statistical analysis, subject/client characteristics, e.g., age, sex, ethnicity, cognitive level, and presenting problems, characteristics of tasks or interventions, e.g., theoretical orientation, setting, components of treatment or experimental tasks, mode of administration, duration, and experimenter or service provider characteristics, and effect sizes (the key statistical outcomes for the study and information about how they were calculated).

*Unreported or Uncertain Information.* Invariably, the reporting of studies is deficient on some variables of interest. Therefore, coding procedures

should routinely contain an "unknown" or "not ascertainable" category which should be reported in any summary of study characteristics. For example, Shapiro and Shapiro (1983) reported that many psychotherapy outcome studies lacked such basic descriptive information as characteristics of clients, the training and experience level of the therapists, and the components of treatment.

Even when studies present information on a variable to be coded, it is often more ambiguous than the meta-analyst would like. Orwin and Cordray (1985) suggested that, for major variables, the meta-analyst obtain a confidence rating from coders to indicate the amount of guesswork that went into the coding. Analyses can later be conducted to assess if findings are influenced by either coder judgment in cases of uncertainty or missing information for some variables.

*Coder Training and Documentation.* The range and variability of reporting in studies eligible for meta-analysis makes coding a difficult and time-consuming task under even the most favorable circumstances. Coding time can average several hours per study. Prospective meta-analysts should plan for sufficient time and resources to complete the coding process effectively.

We recommend that a detailed coding manual be prepared that specifies as completely as possible coding definitions and procedures for handling various coding circumstances. This manual should be made available to others on request. Moreover, it is important for the meta-analyst to "stay close to the data." That is, the coding manual should be based on the researcher's own experience in coding a sample of studies, and the meta-analyst should be alert to the usual need to revise initial coding definitions and procedures as problems are encountered and thus to recode some studies.

Coders should have sufficient background to make the necessary judgments and be trained to criterion levels of consistency in their ratings before their work is accepted into the meta-analysis. On large projects, periodic refresher training is also wise to control for possible coder drift that might occur over time as experienced coders increasingly make judgments without reference to the coding manual (Jackson, 1980).

As a final check on the quality of the coding, of course, the meta-analyst should determine and report the level of intercoder agreement attained in coding the studies (Stock et al., 1982). Since coding procedures often involve a combination of dichotomous, nominal, and continuous data, appropriate methods of estimating interjudge reliability should be used (see Hartmann, 1982).

It is also wise to examine interjudge agreement separately for individual variables and categories of variables in the meta-analysis. Some variables represent greater levels of specificity or subjectivity than others and

naturally attain less intercoder agreement. Some variables, e.g., effect size, are more important to the meta-analysis than others and, thus, low reliability is more consequential for them. Yeaton and Wortman (in press) provide a useful hierarchical scheme for conceptualizing and reporting intercoder agreement in meta-analysis.

## The Index of Effect Size

The key variable in meta-analysis, the hub around which the whole enterprise revolves, is the index of effect size. The various statistical outcomes in the studies of interest are converted into a common, quantitative metric, the index of effect, that can then be compared, aggregated, and analyzed across the studies in the sample. There are several issues that must be considered in order to define and code effect size indices appropriately.

*What Metric to Use?* In principle, there are many ways an index of effect can be defined. In practice, however, there are relatively few that can be readily calculated, represent study results in terms of both direction and magnitude of effect, and have sufficiently well-developed statistical theory to permit sophisticated analysis.

For studies concerned with correlational association, the product-moment correlation or its variants (biserial, point-biserial) is generally the metric of choice. Many such studies report their results directly in this form or in ways that can be readily converted into correlations. Comprehensive discussions of working with a correlational metric are found in Hedges and Olkin (1985), Hunter and Schmidt (1990), and Rosenthal (1984).

For studies concerned with group differences, there are several possibilities to be considered. If all the studies use exactly the same dependent measure, the most straightforward approach is to use the simple difference between group means as the effect size index (e.g., Devine & Cook, 1983; Hovell, 1982; Messick & Jungeblut, 1981). Practically speaking, however, this circumstance rarely occurs in the social and behavioral sciences.

More typically, studies use different measures even when addressing the same research question. The most fully-documented metric for this situation is the standardized difference between group means, i.e., the difference between the means divided by the common (or pooled) standard deviation of the scores within each group (Hedges, 1981). This metric thus represents each difference between means in standard deviation units, i.e., as a *z*-score, rather than in the metric of the original measures. By convention in treatment-control comparisons, these *z*-scores are given positive signs when treatment groups outperform control groups and negative signs

for the reverse outcome. Full discussion of this metric, variously called *d* (after Cohen, 1977), or *g* (after Glass, 1976), or simply ES (effect size), can be found in Glass et al. (1981), Hedges and Olkin (1985), Hunter and Schmidt (1990), and Wolf (1986). Although many early meta-analyses were conducted using only the standard deviation of the control group to calculate the difference between group means, it has now been shown that the pooled standard deviation from both the experimental and control conditions provides a better estimate of the true population effect (Hedges, 1981; Hedges & Olkin, 1985).

A third approach to group differences is to depict them as relationships between an independent variable (group membership) and a dependent variable (the variable on which the groups are compared). Viewed this way, group differences can be represented in strength of association terms, i.e., as a product-moment correlation (in particular, as a biserial or point biserial correlation). This approach is advocated by Hunter and Schmidt (1990) and Rosenthal (1984). While this metric carries the same information as the standardized difference between means (and can be algebraically transformed to that form), it is not obvious that it should be preferred to the latter. The correlational metric is somewhat easier to correct for measurement unreliability and other such artifacts (Hunter & Schmidt, 1990) and, where possible, might be used for that purpose. For more general purposes, however, it has the disadvantage of being less directly interpretable with regard to the magnitude of group differences than the standardized mean difference metric. Reporting that the average treatment vs. control group difference in a research domain is +.50 standard deviations on the dependent measures of interest is generally more easily understood than reporting that the correlation between categorical group membership and the dependent measure is .24.

*Estimation of Effect Sizes.* Effect sizes can be calculated directly when certain basic information is provided, e.g., the exact correlations or the means, standard deviations, and sample sizes per condition, etc. With distressing frequency, the studies one is attempting to code for meta-analysis fail to report these minimal data. In the long run, we can hope that authors and editors will adopt "full disclosure" of key statistics as standard policy in research reporting. Meanwhile, one can try contacting authors to obtain the missing information, although such attempts are not necessarily successful (Hyde, 1981). Usually, the meta-analyst must be prepared to estimate effect sizes from incomplete information and to lose some otherwise eligible studies for lack of adequate statistical information.

Depending on the circumstances, effect sizes can be estimated from less than optimal statistical information. Results presented in one fashion, such as *t* values or *p* levels, can be converted to the desired metric. Many

conversion or estimation techniques are available to deal with the wide range of possibilities encountered in research reports (see Curlette, 1987; Glass et al., 1981; Holmes, 1984; Rosenthal, 1984; & Wolf, 1986). Since the degree to which effect size must be estimated may systematically influence its magnitude, it is a good idea to separately code the method employed to calculate effect size and then assess whether different methods yield different findings.

Two aspects of effect size estimation occur so frequently that some mention is appropriate here. One involves studies within a research domain using both continuous and dichotomous dependent measures. In correlational meta-analysis, the correlations usually involve continuous measures. When they do not, biserial and point-biserial (or Phi coefficient) variants for the correlation are usually serviceable. In group differences meta-analysis, however, it is not so apparent how to convert dichotomous outcomes, e.g., success vs. failure or percent improved, into a standardized difference between means. Probit and logit transformations have been proposed (Glass et al., 1981). Unless these approaches are dictated by specific statistical assumptions about the nature of the underlying distribution being dichotomized, however, we recommend use of Cohen's (1977, 1988) arcsine transformation. It is easy to use and has the advantage of producing effect sizes for dichotomous outcomes that bear the same relationships to sample size and statistical power as analogous effect sizes using continuous outcome measures.

Another potentially troublesome case arises when investigators merely indicate that results were "nonsignificant" for a particular correlation or group difference and report no other data. The most common strategies for dealing with this situation are either to set the effect size at zero in all such cases, or to exclude these findings and their accompanying measures from all analyses. The former approach is conservatively biased because it assumes a zero effect size for results that most likely are not exactly zero. Most meta-analyses find that a majority of studies yield effects that are positive in sign although variable in magnitude. Therefore, setting all "nonsignificant" findings to zero generally underestimates true effect sizes.

Dropping nonsignificant findings, on the other hand, produces an overestimate of the true effect since findings too small to achieve customary levels of statistical significance are underrepresented in the analysis. Note that a reader cannot judge the extent of bias introduced by either strategy unless the meta-analyst indicates how frequently this situation occurs among the studies in question. As with all such instances of ambiguity, it is a good practice for the meta-analyst to bracket the range of potential distortion introduced into the results by reporting analyses based on alternative approaches. In this case, for example, the key meta-analytic findings

could be reported for the analysis with zero substituted for incalculable nonsignificant effect sizes *and* for the analysis in which these effect sizes are dropped out entirely. If the results are comparable, confidence in them is increased. If not, the order of magnitude of the difference can be assessed and its implications interpreted.

*What Effect Sizes to Code.* Few research domains are characterized by studies that report a single finding on a single relationship. More typical are studies reporting findings on a number of measures, sometimes taken at more than one point in time (pre, post, followup), for various subject groupings (total sample, subgroups). To handle this diversity systematically, the meta-analyst must first determine which sorts of findings are of interest. This, of course, will depend upon the purposes of the meta-analysis and the substantive issues relevant to the particular research domain.

Second, the meta-analyst must determine which types of findings are reported with sufficient frequency to yield a data set large enough to analyze. While the durability of experimental effects at successive follow-up points or the differential criterion validity of a measure for different sample subgroups may be of considerable interest, too few studies might report on such matters to justify coding and analysis. In group differences studies, particular attention should be paid to the range of measures on which groups are contrasted. Effect sizes on all such measures can be coded if the meta-analyst feels that it is meaningful to aggregate them into overall means, etc. If categorically different types of measures are represented, however, some may be used with such low frequency as to not justify the effort required to code them. In meta-analysis of educational interventions, for example, most studies will report results on academic performance, but only a handful may also report results on pupil attitudes, self-esteem, or the like. Consistent with the dictum of making the meta-analytic process explicit, the meta-analyst should identify the basis on which measures were selected for effect size coding and, especially, the basis for any exclusions. As Matt (1989) has shown, the decision rules used to select the effect sizes that enter a meta-analysis can have a major impact on the results.

*Contextual Information About Effect Sizes.* Since the magnitude of the effect size is dependent upon the particularities of its nature and coding, it is important to document any factor that may be necessary for interpretation of the results. The coding should, for example, include information about such features as what types of measures are represented in the effect size, relevant specifics of the treatment or research situation, and, for treatment meta-analyses, the nature of the group comparison itself. Usually, the comparison group is a control condition of some sort, but this can include no-treatment, waiting-list, or placebo controls as well as other variants. A recommended procedure is to aggregate studies according to the type of

control condition and test for possible differences in effect sizes before combining studies with differing controls. Some meta-analyses have found that different control conditions yield significantly different effect sizes (e.g., Dush, Hirt, & Schroeder, 1983; Kazdin & Bass, 1989) whereas others have not (e.g., Miller & Berman, 1983). Therefore, the comparability of controls is an empirical question that must be answered in each research area.

It is also possible to use an alternative treatment instead of a control as the comparison in the calculation of effect sizes. The alternative treatments must be comparable across studies, of course. In evaluating the effects of cognitive-behavioral treatment, it would make no sense to aggregate results based on behavior modification alternative treatments with those based on nondirective counseling. Also, treatment vs. treatment and treatment vs. control calculations of effect size should be distinguished in the analysis since the former usually yield lower effect sizes than the latter (Kazdin & Bass, 1989).

Another important piece of information about the effect size is the exact sample size represented in whatever index is used. For correlational meta-analysis, this refers to the total number of subjects over which each correlation is computed. For group comparison meta-analyses, the sample size for each group mean (separately) is at issue. Because of attrition, nonresponse, and other such missing data events, the $n$ may be different for each effect size coded from a study and, in some cases, may be substantially lower than the total number of subjects participating in the overall study. As we will discuss in the next section of this paper, these sample sizes play an important role in any statistical analysis involving effect sizes.

## Statistical Analysis

The statistical analysis of meta-analytic data focuses on the distribution of effect sizes and their covariation with the various descriptive variables coded on each study. The effect size index is thus the dependent variable in the analysis and variables describing the nature and circumstances of the studies, study methods, and the like are independent variables presumed capable of influencing the effect size magnitude.

Statistical theory for meta-analysis data has developed considerable sophistication over the last decade and now provides a relatively complete framework for handling most of the analysis issues with which an investigator might be concerned. We will review here the major features of this framework with emphasis on practical application rather than the statistical theory itself.

*Adjustments to Effect Size.* The raw effect size coded in meta-analysis, whether correlation or standardized mean difference, is generally not the appropriate datum for analysis. It first requires adjustment to correct various inherent inadequacies the index possesses as a statistical estimator.

One such problem is that the effect size is biased when estimated on small samples. It is therefore necessary to adjust the raw effect size for small sample bias and easy-to-use formulas are available for that purpose for both group difference and correlational effect sizes (Hedges, 1981, 1984; Hedges & Olkin, 1985). In practical terms, corrections are important for sample sizes less than 30 and are negligible with larger samples, but common practice is to make the corrections anyway since they can be quickly done by computer.

Effect size estimates are also subject to distortion when based on measures that are unreliable, invalid, restricted in range, or dichotomized (Hunter & Schmidt, 1990). More generally, magnitude-of-effect estimates as a class can be influenced by a variety of methodological, statistical, and psychometric factors that can make comparison among studies difficult (O'Grady, 1982; Sechrest & Yeaton, 1982). If possible, therefore, appropriate corrective adjustments to the effect size estimate should be applied. Statistical procedures are available for correcting the most widespread distortions, e.g., from unreliable study measures, for both correlational and standardized mean difference effect indices (Hedges & Olkin, 1985; Hunter & Schmidt, 1990). Unfortunately, meta-analysts seldom find that the studies of interest report sufficient information to permit all appropriate adjustments to individual effect sizes. For example, Lipsey et al. (1985) examined a sample of evaluation research publications across a wide range of program areas. Fewer than 30% of the studies reported any reliability information on any of the measures used and even fewer reported validity information. Under such circumstances, the missing reliability data could be estimated from the available data or from other published sources, but there is uncertainty regarding the wisdom of this practice. The most conservative approach would be to bracket the mean effect size range between the unadjusted value and the value adjusted on the basis of the incomplete reliability data.

*Outliers.* Statistical analysis of effect size data is vulnerable to considerable distortion by inclusion of even a few extreme values that are greatly discrepant from the remainder of the distribution. There are various ways to detect outliers (Hedges & Olkin, 1985, Chap. 12), although considering any effect size that is two or more standard deviations beyond the mean of its respective group as an outlier serves well in many situations. Once identified, outliers can be handled by excluding them from the statistical

analysis or, less wasteful of data, setting them to less extreme values, i.e., recoding them to designated floor or ceiling values.

While outliers reflecting grossly extreme values are most likely to be the result of error, it should be recognized that the high and low ends of the effect size distribution can offer clues to potentially important explanatory variables in the meta-analysis (Light & Pillemer, 1984). For example, suppose that in a review of 50 school-based programs to prevent drug abuse, several studies with strongly positive findings were discovered. Inspection of these studies indicates that they all involved pre-program training of teachers *and* efforts to insure that the intervention was implemented as planned. The relationship of these two factors to outcome could then be investigated more fully in the remaining studies.

*Unit of Analysis.* As noted earlier, an effect size can be calculated for each dependent measure in a study. Since the number of such measures will almost certainly vary across studies, the resulting data will represent as few as one effect size for some studies and considerably more, sometimes dozens, for other studies. A fundamental issue in meta-analysis is how to handle this very uneven relationship between individual studies, which constitute the sampling unit for meta-analysis, and the effect sizes, which constitute the data points for analysis.

Three major strategies have been applied to this problem (illustrated in Table II). Smith and Glass (1977) advocated using effect sizes from each outcome measure in each study regardless of how many outcome measures were involved per study. In their review of 375 therapy outcome studies, 833 effect size estimates were calculated and entered into the analysis. This approach, however, permits studies with more outcome measures to contribute differentially more data to the analysis. For example, when using each outcome measure as the unit of analysis (column three in Table II), Study A has six outcome measures whereas Study E has only one. Thus, Study A receives six times more weight in the meta-analysis than Study E even though on all other grounds they may be of equivalent importance. This grossly disproportionate weighting and the potential for statistical interdependencies among effect sizes from the same study have led many meta-analysts away from using each outcome measure as the unit of analysis.

A second approach is to use each study as the unit of analysis. Effect sizes from multiple dependent measures are averaged so that each study yields one mean effect size. Thus, the differential weighting and statistical dependency problems that occur in the strategy described above are avoided since each study contributes only one (mean) effect size to the analysis. This approach has been popular among meta-analysts, but it too has a major drawback. In studies where multiple outcomes do not all measure the same construct or relationship, the mean effect size per study ob-

Table II. Findings as a Function of Different Units of Analysis

| Study | Types of measures[a] | Calculated effect size | | |
|---|---|---|---|---|
| | | Per measure | Per study | Per construct |
| A | 1, 1, 2, 3, 4, 5 | 1 = 0.30 1 = 0.45<br>2 = 0.00<br>3 = 0.37<br>4 = 0.12<br>5 = 0.00 | 0.21 | 1 = 0.38<br>2 = 0.00<br>3 = 0.37<br>4 = 0.12<br>5 = 0.00 |
| B | 1, 3 | 1 = 0.98<br>3 = 0.60 | 0.79 | 1 = 0.98<br>3 = 0.60 |
| C | 4, 6, 6 | 4 = 0.24<br>6 = 0.89 6 = 1.25 | 0.79 | 4 = 0.24<br>6 = 1.07 |
| D | 1, 2, 2, 5 | 1 = 0.25<br>2 = 0.09 2 = 0.17<br>5 = 0.33 | 0.21 | 1 = 0.25<br>2 = 0.13<br>5 = 0.33 |
| E | 2 | 2 = 0.88 | 0.88 | 2 = 0.88 |
| Means | | 0.43 | 0.58 | 1 = 0.54<br>2 = 0.34<br>3 = 0.49<br>4 = 0.18<br>5 = 0.17<br>6 = 1.07 |

[a]1 = interpersonal problem-solving measures, 2 = self-reports of anxiety, self-esteem, 3 = checklist/ratings of classroom/home behavior, 4 = academic achievement, 5 = peer sociometrics, and 6 = behavior observational data.

scures the variation in effects that is likely to occur across different construct domains. For example, in Table II, note how Studies B and C achieve the same mean effect size (0.79), but this average results from a combination of totally different types of outcome measures in each study. Whereas the mean effect in Study B is derived from outcome data assessing interpersonal problem-solving skills and parental report of the child's behavior at home, the findings in Study C come from effects averaged over two behavioral observations of the child's classroom behavior and a measure of academic achievement. Thus, Studies B and C achieve the same results numerically, but not conceptually. Using each study as the unit of analysis thus provides data on the general magnitude of effect, but it often fails to provide meaningful information about the specific nature of the effect that has occurred.

A third strategy is to calculate mean effect sizes for each distinct construct domain represented in a study, and to keep these effect sizes separate in subsequent analyses. This procedure involves calculating a separate effect size for each separate outcome dimension. This approach is portrayed in the last column of Table II. Study A yields four separate effect sizes,

one for each outcome measure that assesses a different aspect of adjustment (problem-solving skills, personality self-reports, behavior checklists, and academic measures). Each study contributes one effect size per construct, however, as the effect sizes for the two problem-solving measures are averaged in Study A. Subsequent analyses are then done separately for the effect sizes in each construct domain, e.g., one analysis examining effects on problem-solving skills, one examining personality self-report, and so forth (see the bottom of Table II). If intercorrelations among the outcome variables are available (which is rarely the case), multivariate analysis could be conducted instead of separate analyses for each construct (Raudenbush, Becker, & Kalaian, 1988).

Keeping the construct domains separated permits more specific assessment of study outcomes which, in turn, is generally more useful for practice, research, and social policy. The results for the five studies in Table II indicate that these programs have been most successful in changing specific behaviors (mean ES = 1.07), least successful in modifying the target groups' academic performance (ES = 0.18) or sociometric status (ES = 0.17), and have achieved relatively intermediate success on the other outcome dimensions (mean effect sizes between 0.34–0.54).

One drawback of this third strategy is that averaging effect sizes within construct domains for each study may obscure some differentiation among the specific measure of a construct that the meta-analyst wishes to carry forward in the analysis. For example, in Table II, the two problem-solving measures averaged in Study A may represent one standardized measure and one *ad hoc* measure developed by the researcher. If the meta-analyst wished to investigate whether this difference had any systematic influence on effect sizes across all the studies, averaging the two types of measures would obscure the distinction. The solution in this case would be to create additional outcome categories that reflected this finer-grained differentiation. If subsequent analysis showed that the distinction was not important, the categories could then be aggregated.

Of the three general strategies regarding the unit of analysis, the third usually represents the soundest and simplest procedure for most applications. It is often vital to keep different constructs unconfounded in the analyses (e.g., Chambless, 1989), and we recommend that approach whenever possible.

*Appropriate Weighting Procedures.* Virtually all meta-analyses examine the characteristics of the distribution of obtained effect sizes. The meta-analyst is often interested in whether a mean effect differs significantly from zero, and whether there are significant between-study differences in the effect sizes.

An important feature of the statistical analysis of such issues is the necessity of differentially weighting each effect size in any computations or statistical tests. It will rarely be correct to combine unweighted effect sizes across a set of studies. In such a procedure, each effect size has an equal contribution on the group mean irrespective of the number of subjects upon which it is based. Thus, an effect size from a study using ten subjects is given the same weight as an effect size from a study using 1000 subjects even though the former is clearly a much less reliable estimate than the latter.

Effect size data from large *n* studies must therefore be weighted more heavily in any statistical calculations than effect size data from smaller *n* studies. One obvious approach is to weight each effect size in an analysis by the sample size upon which it is based (Hunter & Schmidt, 1990). This is a workable procedure that has the advantages of being straightforward and easily implemented. Because of these advantages, we would recommend sample size weighting for all but the most specialized meta-analytic applications. It represents a considerable improvement over the common practice of constructing unweighted effect size means, and it is very close to statistically optimal for homogeneous effect sizes (more on this to follow), that is, those that differ from each other only by sampling error (Hunter & Schmidt, 1990). Moreover, as a practical matter, sample size weighting can in most cases be expected to produce nearly the same results as more sophisticated weighting schemes.

Two caveats about sample size weighting are in order, however. First, if there is great disparity in sample sizes, the extremely large *n* studies will dominate the rest in any weighted analysis. This may not be a problem if the effect sizes and other characteristics of the larger studies are not deviant, but it can distort the analysis if they are. One approach to this situation is to analyze the data with and without the large *n* studies to assess their influence on the results. Another reasonable approach is to set very large *n*s equal to some specific limiting value, that is, one less discrepant from the rest of the distribution, to minimize the distortion they might produce.

Second, sample size weighting can be used in situations of heterogeneous effect sizes only if they are not widely disparate. Hedges and Olkin (1985) and Hunter and Schmidt (1990) have shown that the optimal weights for disparate effect sizes combine sample size and population effect size values into a coefficient representing the inverse of sampling error. Population effect size values, of course, are not usually known and thus must be estimated. Hedges and Olkin (1985) proposed that they be estimated by observed effect sizes even though some of these may be very unreliable; Hunter and Schmidt (1990) suggested that they be estimated by the mean

of homogeneous groupings of effect sizes. Neither approach, however, will produce results that are likely to differ meaningfully from sample size weighting unless the range of effect sizes being weighted is greater than about 0–2.0 for standardized mean differences or 0–1.0 for correlations.

To summarize, we wish to emphasize four points with respect to weighting. First, unless all the effect sizes in a meta-analysis are based upon very similar sample sizes, unweighted analysis will be incorrect and should not be done. Second, simple sample size weighting is quite defensible and will yield acceptable results unless it is assumed that the population effect sizes that underlie the sample data being analyzed are widely disparate. Third, if such an assumption is relevant, the inverse variance weighting scheme of Hedges and Olkin (1985) should be used. Fourth, whatever the specific form of weights chosen, all statistical analyses of effect sizes in the meta-analysis should use them, including computation of means, confidence intervals, heterogeneity tests, ANOVA, multiple regression, and the like.

*Testing the Heterogeneity of Effect Sizes.* Potentially the most interesting work a meta-analyst performs is to examine the relationships between effect sizes and various study characteristics. For example, the analysis may investigate whether effect sizes are larger for studies that use one type of participant vs. another, for one intervention vs. another, for one type of control group vs. another, and so forth. The basic challenge for analyses of this sort involves dealing appropriately with the heterogeneity of the observed effect sizes and the differences among the studies upon which those effect sizes are based. On the one hand, the meta-analyst must recognize that even if studies were virtual replicates of each other, all estimating the same identical population effect size, sampling error would result in variability in the observed effect sizes. In this case, the meta-analyst should avoid interpreting mere sampling error as real differences between study outcomes (Hunter & Schmidt, 1990).

On the other hand, care should be taken to avoid reporting mean effects for aggregates of studies that yield genuinely dissimilar effect sizes. The mean effect size, for example, would be a poor representation of a bimodal distribution comprised of one group of studies with strong negative effects and another group with strong positive effects. Moreover, confidence intervals around that mean would be estimated on the basis of an observed effect size variance that reflected not only sampling error but the real variation in effects. Confidence intervals would thus be wider than sampling error alone would yield, and statistically significant differences from zero would be more difficult to demonstrate.

It is important, therefore, to examine the distribution of observed effect sizes to check whether it is plausible that they estimate a single population value differing only by sampling error or are likely to represent real

differences among studies along with sampling error. Unfortunately, in many circumstances, there is no definitive way to conduct this check. One approach that should be used routinely is to apply a homogeneity test. Hedges (1982a, 1984; Hedges & Okin, 1985) describes the homogeneity test for standardized mean differences; Rosenthal and Rubin (1982a) and Hunter and Schmidt (1990) describe it for correlations. If the distribution of effect sizes is homogeneous, the weighted mean and confidence interval can be interpreted as estimating and bracketing, respectively, a single population effect size. If, however, the distribution of observed effect sizes fails the homogeneity test, it may be assumed that there are real differences among the effect sizes, and the analyst may wish to search for moderator variables that explain those differences.

The problem with sole reliance upon the homogeneity test is one of statistical power. The *n* for this test is the number of studies, not the number of subjects in the studies. A meta-analysis with a modest number of studies, e.g., 100, may fail to reject homogeneity when there are, in fact, real differences among effects (Hunter & Schmidt, 1990). A very large meta-analysis, on the other hand, may reject homogeneity even when differences among effects are trivial.

An alternate approach is to examine the effect size variation directly, partitioning it into the sampling error component and the residual variance. The sampling error component can be estimated from information about the number of subjects upon which each effect size is based (Hunter & Schmidt, 1990). The residual component is obtained by subtraction from the total effect size variance and includes both true variation among effect sizes and any error that has not already been adjusted for in the effect size estimates, e.g., from unreliable measures, reporting and coding errors, and so forth. If residual variance is small relative to sampling error, a homogeneous distribution of effect sizes can be assumed. Hunter and Schmidt (1990) suggested a "75% rule"—if the sampling error is 75% or more of the total variance, the residual should not be assumed to represent meaningful variation among effect sizes.

Monte Carlo studies of the performance of various approaches to assessing effect size variation (e.g., Sackett, Harris, & Orr, 1986; Spector & Levine, 1987) provide some guidance to the meta-analyst on this issue. They have shown that both the homogeneity test and the "75% rule" perform well in cases of at least 100 independent effect sizes each based on 75 subjects (study *n*). With as few as ten independent effect sizes each based on fewer than 50 subjects, neither procedure performs well. At intermediate levels, the "75% rule" has more statistical power (is more likely to detect effect size variability if it is present) while the homogeneity test

has lower Type I error rates (is less likely to show effect size variability when none is present).

If the distribution of effect sizes is judged homogeneous (that is, consistent with the assumption of a single "fixed" population effect size), the weighted mean and confidence interval for that mean can be constructed using formulas for the "fixed model" (Hedges & Olkin, 1985; Hunter & Schmidt, 1990). Further breakdowns of effect size, e.g., comparing different treatment modalities, research methods, etc., would not be warranted given the evidence of the homogeneity assessment that there were no meaningful differences among the effect sizes.

If, however, the distribution of obtained effect sizes is judged heterogeneous, the meta-analyst must come to some decision about the source of the effect size variance that exceeds sampling error. If it is assumed to stem from essentially random (or otherwise unaccountable, unknown, or unmeasured) factors, the meta-analyst may choose to adopt the "random model" of effect size variance and offer no further analysis of it. Procedures for constructing confidence intervals about the weighted mean effect size under the random model are provided by Hedges (1983) and Hedges and Olkin (1985). Those confidence intervals will generally be different from, and wider than those for the fixed model—a consequence of the additional variability assumed in the population distribution of effect sizes under the random model.

Alternatively, the meta-analyst may assume that the effect size heterogeneity results from systematic, observable differences among the studies (rather than random effects). In this case, the challenge to the analyst is to identify the study characteristics responsible for the effect size heterogeneity and demonstrate that when the variability attributable to differences in those characteristics is accounted for, the residual heterogeneity is no greater than would be expected from sampling error under the fixed effect model. Since identifying moderator variables (that is, variables that account for heterogeneity in study outcomes) is often a major purpose of meta-analysis, this aspect of statistical analysis is discussed next.

*Searching for Moderators.* One straightforward way to determine the factors that account for differences in effect sizes is to compare the mean effect sizes for studies grouped according to variables of interest (Hedges, 1982b, 1984). For example, a meta-analyst might suspect that one type of treatment is more effective for a particular problem than another or that certain client characteristics affect responsiveness to intervention. Study effects are then categorized accordingly and compared.

How studies should be grouped for analysis merits some discussion since it is one of the more controversial aspects of a meta-analysis. Critics frequently cite the "apples and oranges" problem in meta-analysis, that is,

the creation of overly broad categories of studies that obscure important differences among studies within each category. The problem is that, although studies may be subdivided on one variable of interest, e.g., treatment, they are likely to also differ simultaneously on a host of other potentially important variables that are confounded with the one selected for attention. The multiplicity and possible interdependence of variables make it difficult to determine which one(s) may account for between-study differences.

An example illustrates the typical situation. Imagine a meta-analysis of preventive mental health programs for children in which the coded study characteristics are type of preventive program (primary vs. secondary), type of intervention (behavioral vs. nonbehavioral), number and type of therapists, various child characteristics, e.g., age and gender, and general design features, e.g., assignment to conditions and type of controls used. The prevention studies vary on each of the above characteristics. The meta-analyst who has already determined that there is significant variation among the effect sizes is confronted with the question of how to subdivide the studies to permit meaningful between-study comparisons.

Suppose the meta-analyst believed that the general type of preventive program and that the type of treatment administered were important determinants of effect size. He or she could divide the studies according to the type of prevention attempted (primary vs. secondary) and then compare the mean effect sizes, but this division would probably result in an uneven distribution of type of treatment within each prevention category, potentially biasing the comparison. Alternately, he/she could divide the studies into nonbehavioral and behavioral types of treatment but this, of course, might yield an uneven distribution of type of prevention within the treatment categories and also result in a biased (confounded) comparison. If sufficient studies were available, type of treatment and type of program could be crossed to produce four cells in a 2 × 2 comparison. However, each of these cells is unlikely to be balanced with respect to the other study characteristics such as type of therapist, age and gender of children, and the like. Furthermore, the latter variables might have stronger influences on effect sizes than the ones selected for priority attention.

The worst approach for an analyst to take in this situation is to fish around for data partitions that show high contrast. Such a procedure overemphasizes chance differences, especially in meta-analyses based on a limited number of studies. As mentioned earlier, it is far better to begin a meta-analysis with specific *a priori* hypotheses regarding which study characteristics are expected to influence effect sizes. These hypotheses then guide the partitioning and comparison of effect sizes and provide a context

for judging the importance of any bias introduced by confounds among the independent variables of interest.

Hedges and Olkin (1985) have developed useful statistical procedures for evaluating hypothesized differences among effect sizes (see also Hedges, 1982b, 1984). They demonstrate that the homogeneity statistic (generally called the $Q$ statistic) is distributed as chi square, tests effect size homogeneity (as described earlier), and can be partitioned into "between groups" and "within groups" components. This procedure is thus analogous to an ANOVA in the sense that calculations are made for total heterogeneity across all studies ($Q$-total), heterogeneity between groups of studies ($Q$-between), and heterogeneity within study groups ($Q$-within). The statistical significance of each component can be separately tested.

This approach allows statistical analyses to be conducted to examine the viability of any conceptual grouping of studies hypothesized by the meta-analyst. If $Q$-total is significant (heterogeneity is present), then $Q$-between tests the significance of the differences among the groups of interest, e.g., different types of treatment. $Q$-within tests whether there is still more variability within groups than can be attributed to sampling error. However, before testing $Q$-between, it is important to show that $Q$-within is nonsignificant. If $Q$-within is *significant,* the meta-analyst's hypothesis of which variables accounted for the heterogeneity is found to be insufficient to account for all the effect size variance. Other variables might then be added to create more differentiated groupings or a completely different basis for grouping might be tried. The implications of failing to reduce $Q$-within to nonsignificance are discussed later.

If homogeneity is achieved within each group of studies, the meta-analyst's hypothesis regarding the importance of the grouping variables is supported, each group of studies can be assumed to represent a population with a single fixed effect size, and the $Q$-between statistic provides the test for the between-group differences. The meta-analyst is then on stronger ground in concluding that the grouping variables were responsible for the obtained differences in effect size. Given the confounding among variables, however, it is still possible that a different grouping scheme could produce equally defensible results. The choice among such alternatives often hinges more on the researcher's hypothesis than on the effect size data themselves.

*Multiple Regression.* Simple grouping variables may not be sufficient to account fully for between-study heterogeneity in many meta-analyses. Or, the relevant study features may be best expressed as continuous rather than categorical variables. In these situations, a multiple regression format (Hedges, 1982c, 1984) can be used instead of the ANOVA-type format described above to search for moderators to explain between-study differences. In this approach, variables describing various study features are used

as independent variables to "predict" effect size. The analyst then attempts to develop a regression model in which the predictor variables representing study characteristics account for significant portions of the effect size heterogeneity found between studies. The model is successful if it can be shown that the residual variance in effect size is homogeneous as reflected by the homogeneity test.

As noted earlier, the regression analysis appropriate to meta-analysis is not the usual unweighted version familiar to most researchers. Meta-analysis must use weighted multiple regression, weighting each case entered in the analysis by a coefficient that reflects the sample size (or sampling error variance) upon which the corresponding effect size is based. An example of this approach is found in Lipsey (1991).

Multiple regression or ANOVA analysis with weighted cases is generally straightforward with modern computer programs, e.g., SPSS, SAS. Prospective meta-analysts should not be inhibited by the necessity of conducting case-weighted analysis in situations where simple partitioning of the $Q$-statistic is not capable of doing justice to their hypotheses, e.g., when there are numerous independent variables of interest, when interactions among independent variables are at issue, or when independent variables are continuous rather than categorical. One caveat is in order, however. While adding the appropriate weighting function to conventional analysis schemes should produce correct parameter estimates, case-weighting may have the effect of misleading the computer program about the size of the samples at issue with associated distortion of standard errors and reported levels of statistical significance. Where accurate significance testing is desired, it can generally be done by adjusting the reported standard errors so that they incorporate the correct sample size (Hedges & Olkin, 1985). The specifics of this procedure, of course, will vary according to the type of analysis and the particulars of the computer program used.

*Confounding of Method and Substantive Variables.* When analyzing effect size data, it is particularly important that meta-analysts attend to the possibility that differences among studies on methodological variables will distort the meta-analytic results on substantive variables. When systematically examined, method differences among studies may show more influence on effect sizes than substantive variables (e.g., Lipsey, 1991). In fact, they may be so confounded with substantive variables that great care must be taken to ensure that what appear to be substantive differences are not equally attributable to method artifacts.

Table III provides an example. Section A shows mean effect sizes from a fictional meta-analysis contrasting primary prevention intervention with secondary prevention intervention. Secondary prevention appears to be more effective. Section B, however, breaks those same categories down

Table III. Illustration of Confounding of Methodological and Substantive Variables

A. Comparison of mean effect size for two treatment types

| | Mean ES | *N* of studies |
|---|---|---|
| Primary prevention | .51 | 260 |
| Secondary prevention | .58 | 240 |

B. Treatment types diisaggregated by type of research design

| | Design | |
|---|---|---|
| | Comparison group | Pre-post comparison |
| Primary prevention | .43 (200) | .76 (60) |
| Secondary prevention | .36 (100) | .74 (140) |

C. Treatment types with comparison group designs by type of measure

| | Outcome measure | | | | | |
|---|---|---|---|---|---|---|
| | Problem solving | Self-report | Behav. checklist | Acad. ach. | Peer socio. | Behav. obs. |
| Primary | .33 (25) | .47 (25) | .52 (40) | .23 (25) | .15 (25) | .59 (60) |
| Secondary | .34 (16) | .48 (14) | .50 (16) | .20 (20) | .16 (20) | .60 (14) |

according to research design. Here we see that pretest–posttest designs yield much larger effect sizes than comparison/control group designs (a very common finding in meta-analysis). Within both design categories, however, we now see that primary prevention produces larger effect sizes than secondary prevention, thus reversing the conclusion of Section A. This happened because a higher proportion of the secondary prevention programs were being evaluated using pre–post designs which, as noted, tended to yield larger effect sizes.

In Section C of Table III we see a further breakdown of effect sizes by type of outcome measure for comparison group designs only. Here we see that those types of measures that show the largest effects were more frequently used in primary prevention studies. When we compare primary and secondary prevention within each measurement category, we find that their effect sizes are virtually identical. Thus, neither primary nor secondary prevention appears superior.

The point of this example should be clear. A meta-analyst cannot draw a valid conclusion about the relative efficacy of primary vs. secondary prevention treatments without first ascertaining that major methodological confoundings have been identified and examined for any possible distortion they might introduce into the findings. Unfortunately, methodological confoundings are very common in meta-analysis and, even more unfortunate, rarely examined. The result is that many of the substantive conclusions re-

ported in the meta-analysis literature on the basis of simple comparisons may well be invalid. We recommend, therefore, that meta-analysts pay close attention to the methodological variation among their studies and make the effort to show either that it is not significant or, if significant, is not confounded with the substantive differences among studies that the meta-analyst wishes to interpret.

*Model Failure.* As described above, a meta-analyst with a set of observed study effect sizes should assess their homogeneity and come to some judgment about whether the differences among them are greater than would be expected from sampling error alone. If not, further analysis of their differences is inappropriate. If, however, the assessment suggests that there are real differences, then the next step is to fit some statistical model to the data. When information about study characteristics that might account for the effect size differences is not available, or the differences are assumed to be inherently random, the analyst might work with the "random effects" model for effect size variance mentioned earlier. In this model, the effect size variance can be partitioned into a component estimating sampling error and one estimating real differences among the study effects, but no further account of the latter is offered.

If an attempt is made to account for differences in study effects using observed study characteristics as independent or predictor variables along the lines discussed in the immediately preceding sections of this paper, the question of the adequacy, or goodness of fit of the resulting model must be faced. Such a model can be defended as adequate (though not uniquely so) if it can be shown that the variability in effect sizes *not* accounted for by the model is no greater than would be expected from sampling error. In ANOVA style models with categorical independent variables, the variance unaccounted for is that within the cells defined by the predictors, e.g., $Q$-within. In multiple regression models with continuous independent variables, the variance unaccounted for is that of the residuals after the predicted values have been subtracted.

If the variance unaccounted for by the model that is fit to the effect size data is too great to attribute to sampling error, the meta-analyst has a problem. By definition, other factors are necessary to account fully for the differences among effect sizes. Those other factors, if included in the model, might well change the strength of the relationship between the factors already in the model and the effect size, much as adding information about research design in Table III, Section B, changed the relationship between effect size and type of treatment that appeared in Section A. The problem described here is one of potential model misspecification and, when it occurs, it makes interpretation of the model very uncertain.

The best protection against model misspecification, of course, is good hypotheses about the sources of effect size differences and good independent variables with which to test those hypotheses. If, however, the model fails, i.e., has residual variance greater than expected from sampling error, the meta-analyst must offer cautions about placing too much faith in the effect size correlates identified in that model. It may be that the overlarge residual variance in effect sizes represents random or irrelevant variation rather than some systematic factor(s) that might alter the relationships in the model, but such a circumstance is difficult to confirm. Model fit, therefore, should be explicitly tested in meta-analysis and, if inadequate, the implications for interpreting relationships between independent variables and study effects should be discussed.

## Drawing Appropriate Conclusions

Although it seems elemental that meta-analysts should not over-interpret their findings or over-generalize their conclusions, the sheer scope of most meta-analytic reviews presents unusual temptations. In addition to the cautions just advanced concerning model misspecification, we can offer three important guidelines to assist researchers in exercising proper restraint.

*Explaining Nonsignificance.* First, meta-analytic results testing important hypotheses that fail to reach significance, should be closely examined. Some findings from meta-analyses have received notoriety because of their *non*significance, i.e., that the type of treatment provided has no differential effect on psychotherapeutic outcome (Smith & Glass, 1977). Such null results might accurately reflect the true state of affairs, but they can also be artifactual. To begin with, there is the inherent problem of affirming the null hypothesis; in cases of ambiguity the null hypothesis stands by default, not because it has been confirmed. Also, there may be confounds, e.g., with method variables as noted earlier, that act to suppress real differences on substantive variables. It might also happen that there would be so much error in the effect size data from unreliable measures in the studies, coder guesswork, or other uncorrected artifacts (Hunter & Schmidt, 1990) that effects for moderator variables are obscured. Finally, there is the issue of statistical power which is often overlooked and which we wish to emphasize here.

The question of power is particularly relevant to between-study comparisons in meta-analysis. The operative *n*s for such comparisons are the numbers of studies in the respective groups, not the (much larger) number of subjects represented in the studies. When meta-analysts divide the total group of studies into subgroups to investigate certain variables of interest, rather small numbers per cell may result. Missing data often causes a further

loss of studies for particular comparisons. We have seen meta-analysts reach conclusions based on cell sizes of less than 12 without attending to the very limited power of those comparisons. Therefore, it is important in all analyses to present clearly the sample sizes for each comparison group and conduct a power analysis to assess the ability of the statistical tests to detect effects. At the least, results of such comparisons should be reported as confidence intervals, which are very broad in low power situations, rather than as simple significant vs. nonsignificant dichotomies. Kazdin and Bass (1989) provide a good example of power considerations in relation to meta-analytic findings. It is up to the meta-analyst to demonstrate that the failure to find significant effects on important issues is not merely a question of low power.

*Restricting Conclusions.* A second element helpful in guiding the interpretations of meta-analytic findings is that any conclusions that are offered should be restricted specifically to the literature reviewed. Once again, the original delineation of the literature to be searched and the attendant inclusionary and exclusionary criteria should make clear the class of studies to which the results apply. For example, meta-analysis of preventive interventions focusing on primary prevention should emphasize that current results do not apply to secondary prevention programs. Since the parameters of the literature examined often get "lost" during a complicated or lengthy meta-analysis, these features should be re-emphasized when authors are interpreting their findings.

*Limitations in the Database.* Third, the meta-analyst must qualify findings in relation to the limitations of the available studies, particularly with respect to restricted or diminished data on certain variables of interest. For example, it would be misleading to conclude that length of treatment has no effect on outcomes if the majority of studies involve short-term interventions of ten sessions or less. We recommend that a table describing the major characteristics of the studies included in the meta-analysis be routinely presented to minimize ambiguity on this point. Often, meta-analysts must acknowledge that the absence of sufficient data precluded effective analyses of some variables of interest. For example, reports rarely contain information on the racial/ethnic characteristics of the sample or present outcome data separately for males and females. An analysis of ethnic or gender effects based on comparatively few unrepresentative studies containing such information is not likely to be accurate or meaningful.

## TRENDS TO BE ENCOURAGED

We indicated at the beginning of this article that the strategies employed in meta-analysis have evolved since their introduction into the social

and behavioral sciences. When a research strategy is evolving rapidly, it is difficult to predict exactly its future course. Nevertheless, we feel there are a few developing trends to highlight and, indeed, to encourage as the field of meta-analysis matures.

## Replications of Prior Reviews

Our first point follows directly from the notion that advancements have occurred in the concepts, the methods, and especially the statistical aspects of meta-analysis. These developments suggest the need to re-examine the findings of many prior meta-analytic studies. For example, meta-analyses conducted on small samples, those involving no unpublished studies, those not attending to concerns regarding the appropriate grouping of studies, and those lacking statistical analysis of effect size variability may yield findings that are not tenable. We are not saying that the results of previous meta-analyses are necessarily inaccurate. As any field advances, however, previous findings should be re-examined using sounder and more probing research techniques before they are accepted as current and valid. Therefore, we think readers can expect to see many attempted replications of previous meta-analyses, especially those that have strongly influenced perspectives and practices in certain research domains.

## Practical and Clinical Significance

Although a meta-analysis provides information on the statistical magnitude of an effect, it does not necessarily provide information on the practical significance of that effect. Does a mean correlation of .20 between social support and homelessness identify a situation that should be of concern to researchers or policymakers? Does a standardized mean difference of .35 between clients receiving therapy and those in a control group represent a meaningful change? The answers to such questions are essential to understanding the implications of meta-analytic results for practice, policy, and, often, even for theory.

Although assessment of the practical magnitude of effect sizes in meta-analysis is currently rare, notable progress has been made in identifying various techniques by which statistical effect size information can be translated or interpreted into more meaningful form. One approach is the simple conversion of correlational or standardized mean difference effect sizes into another index more readily understood. Rosenthal and Rubin (1982b), for example, have advocated the use of what they call the "binomial effect size display"—a representation of effect sizes in terms of dif-

ferential success rates (in percentage terms) between two groups. A correlation of "only" .20, for instance, translates into a difference of 20 percentage points in the success rate of, say, a treatment group and a control group, e.g., 60% vs. 40% success. A variety of similar translations, based on the notion of overlapping distributions, has been offered by Cohen (1988).

For treatment effects, some meta-analysts have conducted normative comparisons to assess the clinical significance of outcomes (Durlak, Fuhrman, & Lampman, in press; Robinson, Berman, & Neimeyer, 1990; Trull, Nietzel, & Main, 1988). For example, Durlak et al. (in press) calculated a "normative effect size" statistic by dividing the difference between the means of a treatment and a normal group by the standard deviation of the normal group. Calculated before and after treatment, this index assesses the magnitude of the initial gap between clients and a normal population and the extent to which treatment subsequently closes that gap.

More generally, a variety of approaches have been proposed for indexing a given effect size to the magnitude of effect found in a relatively well understood criterion situation (e.g., Abelson, 1985; Lipsey, 1990). Carver (1975), for example, showed how differences in mean achievement scores between students at different grade levels could be used to index the magnitude of educational effects.

Another approach to interpreting the magnitude of obtained effect sizes is to compare them to the range of effects found in other studies. Cohen (1977, 1988) reported that over a wide range of social and behavioral science research, standardized mean differences of about .20 were small, while .50 was medium and .80 was large. The analogous values for correlations are .10, .24, and .37, respectively. Cordray and Orwin (1983) and Lipsey (1990) have extended this actuarial approach by reporting reference distributions specifically from treatment effectiveness meta-analysis research.

Additional approaches to effect size interpretation are available in some situations. Sechrest and Yeaton (1981) provided a useful review of strategies potentially applicable to meta-analysis results. They suggested, for example, that many of the relationships indexed by effect sizes may have overt behavioral consequences. Sechrest and Yeaton proposed that for these effects the researcher establish the magnitude of a "just noticeable difference," e.g., behavior change that family or friends would spontaneously notice, and use that to calibrate effect size data from research.

Whatever the approach adopted, it is clear that meta-analysis will not realize its potential for influencing policy and practice until the results are routinely reported in terms that make their practical significance clear. We

encourage continued efforts to translate the inherently statistical results of meta-analysis into practical information.

## Cost/Benefit Analyses

One way to judge the practical significance of effect size data is to apply cost analysis (Sechrest & Yeaton, 1981). The ratio of the monetary gain associated with an effect to the cost of producing that effect gives one indicator of practical value. In addition, cost/benefit analyses are extremely important in terms of public policy implications, especially in the domain of intervention and social programs (Yates, 1985). Basically, such analyses evaluate whether the returns from a program are worthwhile in terms of the expenditures required to establish and maintain the program. Many permutations of relative costs and benefits are possible, and it is important to stress that it is the relation between costs and benefits, not the absolute levels of either that are critical. For instance, highly effective programs may be too costly to implement on a wide scale and thus fare poorly in cost/benefit terms, whereas interventions with more modest effects that are also less expensive may justify broad dissemination and adoption.

Cost/benefit analyses will be a challenge for meta-analysts since few research reports collect or present data directly relevant to such determinations. A study by Levin, Glass, and Meister (1984), however, shows the potential inherent in combining cost analysis and meta-analysis. They examined the cost-effectiveness of four educational interventions using mean effect sizes from meta-analysis to assess relative program effectiveness. As this example illustrates, meta-analysts will need to apply techniques from several specialized fields in order to construct an appropriate cost-benefit schema. The benefits of greater utility for policy and program decisions, however, should justify the effort entailed.

## Explanatory Meta-Analysis

To date, meta-analysis has been largely applied in a descriptive mode. It has been used to characterize the magnitude and variability of the effects found in a research domain and to provide some limited analysis of the correlates of those effects. While this descriptive information is often quite valuable, meta-analysis has made little contribution to theory development in the domains to which it has been applied. Since primary research itself often has a theoretical aspect, and meta-analysis builds upon primary research (and itself should be hypothesis driven, as noted earlier), there is

no intrinsic reason why meta-analysis cannot deal with theory and explanation as well as description.

Meta-analysis becomes explanatory rather than descriptive when it begins to integrate evidence on multiple, interconnected relationships rather than focusing on a single key relationship. For example, meta-analysis can be used effectively to summarize evidence on the effects of a social intervention. But the outcome of such an intervention can be thought of as the end result of a causal chain that involves a series of intermediate interactions among environmental, participant, and change agent variables. If we charted a proposed causal diagram for the intervention process, we would find that each link posited particular relationships among relevant variables. For each such relationship, there is likely a research literature and a potential meta-analysis that could integrate evidence on the strength of the relationship. The coordination of such meta-analytic results would bear directly on the theory embodied in the causal diagram.

This broader application of meta-analysis is still in its infancy, but holds great potential as it matures. Noteworthy work in this direction has been sponsored by the Russell Sage Foundation's meta-analysis project (Russell Sage Foundation, 1991).

## CONCLUSION

Meta-analysis can be a powerful and useful technique. In the short term, the results of meta-analysis can increase understanding and influence the future endeavors of researchers interested in a particular field. Over the long term, meta-analysis provides a vehicle by which large bodies of research can be integrated and focused on public policy issues.

Administrative structures currently exist to foster the relationship between research findings and public policy. As one example, Saxe (1986) has described the activities of the Office of Technology Assessment (OTA) established by Congress to provide lawmakers with scientific information in such a way as to be useful for policy decision-making. OTA is commissioned by Congress to provide assessments of scientific findings, and meta-analyses are an ideal resource in this regard. Saxe (1986) notes "In almost all cases, the emphasis is on making sense of already collected data, thus literature reviews and methods for synthesizing research results are relied upon" (pp. 61-62).

Whether meta-analysis actually plays an important role in public policy depends on many factors, not the least of which is the care with which meta-analyses are conducted. Hopefully, our comments here will help in-

vestigators conduct methodologically rigorous meta-analyses of research relevant to the field of community psychology.

## REFERENCES


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

Abrami, P. C., Cohen, P. A., & d'Apollonia, S. (1988). Implementation problems in meta-analysis. *Review of Educational Research, 58,* 151-179.

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101,* 213-232.

Bangert-Drowns, R. L. (1986). Review of developments in meta-analytic method. *Psychological Bulletin, 99,* 388-399.

Bullock, R. J., & Svyantek, D. J. (1985). Analyzing meta-analysis: Potential problems, an unsuccessful replication, and evaluation criteria. *Journal of Applied Psychology, 70,* 108-115.

Carver, R. P. (1975). The Coleman Report: Using inappropriately designed achievement tests. *American Educational Research Journal, 12,* 77-86.

Chalmers, T. C., Smith, H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D., & Ambroz, A. (1981). A method for assessing the quality of a randomized control trial. *Controlled Clinical Trials, 2,* 31-49.

Chambless, D. L. (1989). Measurement effects on outcome of treatment for agoraphobia. *Behavior Therapy, 20,* 465-466.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.

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

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

Cooper, H. (1984). *The integrative research review: A systematic approach.* Beverly Hills, CA: Sage.

Cordray, D. S. (1990). Strengthening causal interpretations of non-experimental data: The role of meta-analysis. In L. Sechrest, E. Perrin, & J. Bunker (Eds.), *Research methodology: Strengthening causal interpretations of nonexperimental data* (pp. 151-172). Washington, D.C.: U.S. Department of Health and Human Services, Agency for Health Care Policy and Research.

Cordray, D. S., & Orwin, R. G. (1983). Improving the quality of evidence: Interconnections among primary evaluation, secondary analysis, and quantitative synthesis. In R. J. Light (Ed.), *Evaluation studies review annual* (Vol. 8, pp. 91-119). Beverly Hills, CA: Sage.

Curlette, W. L. (1987). The meta-analysis effect size calculator: A BASIC program for reconstructing unbiased effect sizes. *Educational and Psychological Measurement, 47,* 107-109.

Devine, E. C., & Cook, T. D. (1983). Effects of psychoeducational interventions on length of post-surgical hospital stay: A meta-analytic review of thirty-four studies. In R. J. Light (Ed.), *Evaluation studies review annual* (Vol. 8). Beverly Hills, CA: Sage.

Dobson, K. S. (1989). A meta-analysis of the efficacy of cognitive therapy for depression. *Journal of Consulting and Clinical Psychology, 57,* 414-419.

Durlak, J. A., Lampman, C., & Wells, A. (1991). *Effectiveness of child psychotherapy: A re-evaluation.* Manuscript submitted for publication.

Durlak, J. A., Fuhrman, T., & Lampman, C. (in press). Effectiveness of cognitive behavior therapy for maladapting children: A meta-analysis. *Psychological Bulletin.*

Dush, D. M., Hirt, M. L., & Schroeder, H. E. (1983). Self-statement modification with adults: A meta-analysis. *Psychological Bulletin, 94,* 408-422.

Fuchs, D., & Fuchs, L. S. (1989). Effects of examiner familiarity on Black, Caucasian, and Hispanic children: A meta-analysis. *Exceptional Children, 55,* 303-308.

Garfield, S. L. (Ed.). (1983). Meta-analysis and psychotherapy (special section). *Journal of Consulting and Clinical Psychology, 51*(1).

Gibbs, L. E. (1989). Quality of study rating form: An instrument for synthesizing evaluation studies: *Journal of Social Work Education, 25,* 55-67.

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

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

Guzzo, R. A., Jackson, S. E., & Katzell, R. A. (1987). Meta-analysis analysis. *Research in Organizational Behavior, 9,* 407-442.

Hartmann, D. P. (Ed.). (1982). *Using observers to study behavior. New directions for methodology of social and behavioral sciences.* San Francisco: Jossey-Bass.

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

Hedges, L. V. (1982a). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92,* 490-499.

Hedges, L. V. (1982b). Fitting categorical models to effect sizes from a series of experiments. *Journal of Educational Statistics, 7,* 119-137.

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

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin, 93,* 388-395.

Hedges, L. V. (1984). Advances in statistical methods for meta-analysis. *New Directions for Program Evaluation, 24,* 25-42.

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

Holmes, C. T. (1984). Effect size estimation in meta-analysis. *Journal of Experimental Education, 52,* 106-109.

Hovell, M. F. (1982). The experimental evidence for weight-loss treatment of essential hypertension: A critical review. *American Journal of Public Health, 72,* 359-368.

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., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies.* Beverly Hills, CA: Sage.

Hyde, J. S. (1981). How large are cognitive gender differences: A meta-analysis using omega and *d. American Psychologist, 36,* 892-901.

Jackson, G. B. (1980). Methods for integrative reviews. *Review of Educational Research, 50,* 438-460.

Jackson, S. E. (1984, August). *Can meta-analysis be used for theory development in organizational psychology?* Paper presented at the Meeting of the American Psychological Association, Toronto, Canada.

Jorm, A. F. (1987). Sex differences in neuroticism: A quantitative synthesis of published research. *Australian and New Zealand Journal of Psychiatry, 21,* 501-506.

Kazdin, A. E., & Bass, D. (1989). Power to detect differences between alternative treatments in comparative psychotherapy outcome research. *Journal of Consulting and Clinical Psychology, 57,* 138-147.

Kiesler, C. A. (1985). Meta-analysis, clinical psychology, and social policy. *Clinical Psychology Review, 5,* 3-12.

Kulik, J. A. (1984, April). *The uses and misuses of meta-analysis.* Paper presented at the meeting of the American Educational Research Association, New Orleans.

Levin, H. M., Glass, G. V., & Meister, G. R. (1984). *Cost-effectiveness of four education interventions.* Stanford, CA: Stanford University.

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

Lipsey, M. W. (1988). Practice and malpractice in evaluation research. *Evaluation Practice, 9,* 5-24.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research.* Newbury Park, CA: Sage.

Lipsey, M. W. (1991). Juvenile delinquency treatment: A meta-analytic inquiry into the variability of effects. In *Meta-analysis for explanation: A casebook.* New York: Russell Sage Foundation.

Lipsey, M. W., Crosse, S., Dunkle, J., Pollard, J., & Stobart, G. (1985). Evaluation: The state of the art and the sorry state of the science. *New Directions for Program Evaluation, 27,* 7-28.

Loeber, R., & Stouthamer-Loeber, M. (1986). Family factors as correlates and predictors of juvenile conduct problems and delinquency. In M. Tonry & N. Morris (Eds.), *Crime and justice: An annual review of research* (Vol. 7). Chicago, IL: University of Chicago Press.

Martocchio, J. J., & O'Leary, A. M. (1989). Sex differences in occupational stress: A meta-analytic review. *Journal of Applied Psychology, 74,* 495-501.

Matt, G. E. (1989). Decision rules for selecting effect sizes in meta-analysis: A review and reanalysis of psychotherapy outcome studies. *Psychological Bulletin, 105,* 106-115.

Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin, 89,* 191-216.

Miller, R. C., & Berman, J. S. (1983). The efficacy of cognitive behavior therapies: A quantitative review of the research evidence. *Psychological Bulletin, 94,* 39-53.

Nurius, P. S., & Yeaton, W. H. (1987). Research synthesis reviews: An illustrated critique of "hidden" judgments, choices, and compromises. *Clinical Psychology Review, 7,* 695-714.

O'Grady, K. E. (1982). Measures of explained variance: Cautions and limitations. *Psychological Bulletin, 92,* 766-777.

Okun, M. A., & Stock, W. A. (1987). The construct validity of subjective well-being measures: An assessment via quantitative research synthesis. *Journal of Community Psychology, 15,* 481-492.

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

Orwin, R. G., & Cordray, D. S. (1985). Effects of deficient reporting on meta-analysis: A conceptual framework and reanalysis. *Psychological Bulletin, 97,* 134-147.

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

Robinson, L. A., Berman, J. S., & Neimeyer, R. A. (1990). Psychotherapy for the treatment of depression: A comprehensive review of the controlled outcome research. *Psychological Bulletin, 108,* 30-49.

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

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

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

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

Russell Sage Foundation (1991). *Meta-analysis for explanation: A casebook.* New York: Russell Sage Foundation.

Sackett, P. R., Harris, M. M., & Orr, J. M. (1986). On seeking moderator variables in the meta-analysis of correlational data: A Monte Carlo investigation of statistical power and resistance to Type I error. *Journal of Applied Psychology, 71,* 302-310.

Saxe, L. (1986). Policymakers' use of social science research: Technology assessment in the U.S. Congress. *Knowledge: Creation, Diffusion, Utilization, 8,* 59-78.

Sechrest, L., & Yeaton, W. H. (1981). Empirical bases for estimating effect size. In R. F. Boruch, P. M. Wortman, D. S. Cordray, & Assocs. (Eds.), *Reanalyzing program evaluations: Policies and practices for secondary analysis of social and educational programs* (pp. 212-224). San Francisco, CA: Jossey-Bass.

Sechrest, L., & Yeaton, W. H. (1982). Magnitudes of experimental effects in social science research. *Evaluation Review, 6,* 579-600.

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

Shapiro, D. A., & Shapiro, D. (1983). Comparative therapy outcome research: Methodological implications of meta-analysis. *Journal of Consulting and Clinical Psychology, 51,* 42-53.

Skiba, R. J. (1989). The importance of construct validity: Alternative models for the assessment of behavioral disorders. *Behavioral Disorders, 14,* 175-185.

Slavin, R. E. (1986). Best evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher, 15,* 5-11.

Smith, M. L. (1980). Publication bias and meta-analysis. *Evaluation and Education, 4,* 22-24.

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

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

Spector, P. E., & Levine, E. L. (1987). Meta-analysis for integrating study outcomes: A Monte-Carlo study of its susceptibility to Type I and Type II errors. *Journal of Applied Psychology, 72,* 3-9.

Stevenson, M. R., & Black, K. N. (1988). Paternal absence and sex role development: A meta-analysis. *Child Development, 59,* 793-814.

Stock, W. A., Okun, M. A., Haring, M. J., Miller, W., Kinney, C., & Seurvorst, R. W. (1982). Rigor in data synthesis: A case study of reliability in meta-analysis. *Educational Researcher, 11,* 10-14.

Strube, M. J., Gardner, W., & Hartmann, D. P. (1985). Limitations, liabilities, and obstacles in reviews of the literature: The current status of meta-analysis. *Clinical Psychology Review, 5,* 63-78.

Suls, J., & Wan, C. K. (1989). The relation between Type A behavior and chronic emotional distress: A meta-analysis. *Journal of Personality and Social Psychology, 57,* 503-512.

Tait, M., Padgett, M. Y., & Baldwin, T. T. (1989). Job and life satisfaction: A re-evaluation of the strength of the relationship and gender effects as a function of the date of the study. *Journal of Applied Psychology, 74,* 502-507.

Trull, T. J., Nietzel, M. T., & Main, A. (1988). The use of meta-analysis to assess the clinical significance of behavior therapy for agoraphobia. *Behavior Therapy, 19,* 527-538.

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

Weisz, J. R., Weiss, B., Alicke, M. D., & Klotz, M. L. (1987). Effectiveness of psychotherapy with children and adolescents: A meta-analysis for clinicians. *Journal of Consulting and Clinical Psychology, 55,* 542-549.

Wilson, G. T. (1985). Limitations of meta-analysis in the evaluation of the effects of psychological therapy. *Clinical Psychology Review, 5,* 35-47.

Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis,* Beverly Hills, CA: Sage.

Wortman, P. M. (1983). Evaluation research: A methodological perspective. *Annual Review of Psychology, 34,* 223-260.

Wortman, P. M., & Bryant, F. B. (1985). School desegregation and black student achievement: An integrative review. *Sociological Methods and Research, 13,* 289-324.

Yates, B. T. (1985). Cost-effectiveness analysis and cost-benefit analysis: An introduction. *Behavioral Assessment, 7,* 207-234.

Yeaton, W. H., & Wortman, P. M. (in press). On the reliability of meta-analytic reviews: The role of intercoder agreement.