Some Limiting Factors in Meta-Analysis

Robert L. Bangert-Drowns

In first explicating the notion of quantitative literature review for the social sciences, Glass (1976) argued that knowledge is not built from any individual study, but from the integration of findings from many studies. Individual studies do not so much yield knowledge as evidence with which knowledge can be built. Knowledge is socially constructed. To overemphasize a single study's findings or integrate research only impressionistically leaves researchers knowing less than the evidence offers, insufficiently exploiting the wealth of data scattered in separate studies.

Quantitative research integration, or meta-analysis, has a history in both the physical and social sciences that precedes Glass' formulation (Bangert-Drowns 1986; Hedges 1987). Most generally, meta-analysis is a perspective rather than a method, a recognition that research findings can be interpreted probabilistically in the context of collections of studies. The meta-analytic perspective is consistent with, and perhaps newly empowers, communal and cumulative activities of science in refining method and transforming data into knowledge (Schmidt 1992).

A number of writers initially responded with skepticism or even overt hostility to this apparently new method of inquiry (e.g., Eysenck 1978). It is hard now to find critics opposed to meta-analysis in principle (Wachter 1988). However, two kinds of concerns are still expressed about meta-analysis. The first suggests that quantitative review communicates an appearance of precision and comprehension which is in fact unreal and thus misleading. The second concern is that meta-analysis is not doing what it claimed it could do: settle important theoretical and practical questions in the midst of contradictory research findings.

These concerns arise from the fact that there is plenty of room for subjectivity and imprecision in meta-analysis (Guzzo et al. 1987; L'Hommedieu et al. 1988; Wanous et al. 1989). In spite of advances in meta-analytic method that are meant to increase the precision of literature review, meta-analysis is still, in many ways, a very human enterprise. Though in principle meta-analysis offers simple means for

rendering primary research more useful, meta-analysts disagree about appropriate method (Bangert-Drowns 1986), implementation of method (Carlberg et al. 1984; Slavin 1984), and interpretation of findings (Clark 1985). Implementations of meta-analyses vary in quality and must be read with the same scrutiny afforded primary research. Primary research itself presents vagaries and biases to the reviewer that surely confound precise conclusions about underlying parameters.

Meta-analysis promises to simplify complex literatures, but will be indelibly marked with the many human decisions that shaped the original data and then integrated it in new ways. Consumers of meta-analytic products therefore must carefully review meta-analytic findings. This chapter will alert readers to critical strengths and limitations of meta-analysis for policy, theory, and practice.

COMMON CRITICISMS OF META-ANALYTIC METHOD

Meta-analytic method consists of six phases: formulation of a purpose, retrieval of studies, coding of study characteristics, calculation of effect sizes, analysis of central tendency and variation in effect sizes, and interpretation and publication of findings. Meta-analysts hear many criticisms of this process, but most criticisms target specific phases of meta-analytic implementations rather than meta-analysis in principle.

Apples and Oranges

Some critics argue that meta-analysis, in its effort to be comprehensive, necessarily mixes elements that are too dissimilar to warrant integration. Meta-analysts have been said to use "overly broad categories" which in fact confuse rather than clarify important distinctions in the literature (Gallo 1978; Presby 1978).

This apples-and-oranges problem can affect both dependent and independent variables at the levels of constructs and operationalizations of constructs. Most readers would not be concerned if a meta-analyst mixed different operationalizations of the same construct, for example, finding an average attitude toward personal drug use by aggregating standardized outcome measures (effect sizes) associated with the different attitude toward drug use instruments. However, a meta-analyst could also aggregate across constructs, combining, for example, measures of knowledge, attitude,

and behavior to study a more generalized construct, effect of substance abuse education. A meta-analyst can define treatment or outcome constructs and operationalizations narrowly or broadly, and critics can complain about the breadth of such definitions.

Most importantly, however, meta-analysts control the scope of the constructs and operationalizations they wish to review. How meta-analysts formulate their purposes for review, and, secondarily, how they code study characteristics and calculate effect sizes, determine the breadth of categories they employ. Colleagues may complain that a construct is too broad to be interpretable or practical, or too narrow to provide an overview of a literature. But meta-analysts, not meta-analytic method, determine whether apples and oranges are mixed in overly broad categories.

Garbage In, Garbage Out

Another common criticism of meta-analysis (e.g., Eysenck 1978) concerns the quality of the primary research included in reviews. It has been claimed that meta-analysis is too inclusive and too willing to accept data from poorly designed studies in an effort to be comprehensive. Would it not be better to highlight the findings from a handful of well-designed studies than to give equal attention to the results of good and bad studies alike?

In principle, exclusivity has some merit, but reviewers invariably disagree about what constitutes good quality research. Glass (1976; Glass et al. 1981) argued that excluding studies a priori may lose data needlessly if quality of research has no relation with study outcomes. Glass' empirical response was to code threats to validity as independent variables and test their relation to treatment effects. If no relations exist, studies can be combined regardless of quality.

Glass' response is not an entirely satisfactory one. Good and poor studies may not differ in mean effect size, but in distribution. Differential distributions related to study quality could add considerable imprecision to average effect sizes, especially when categorizing studies into smaller groups according to study features. One also needs to consider the meta-analysis' credibility. Some studies are so notoriously or obviously flawed that to include them would cast doubt on overall findings.

No reviewer can escape issues of inclusion. Even the most inclusive meta-analysts exclude some studies from their reviews, perhaps case

studies or pre-post designs. In all cases, meta-analysts should report inclusion criteria explicitly so that readers can determine how the sample of studies was formed and if adequate attention was paid to study quality.

The "garbage in, garbage out" complaint reflects concern with the study retrieval phase of meta-analysis. Like the complaint about apples and oranges, it is directed more at implementation than at meta-analysis itself. Meta-analysts may attend insufficiently to study quality, but nothing about meta-analytic method necessitates such attention or inattention.

Oversimplification of Research

It is tempting to see meta-analysis' walk-away message in terms of main effects, and results of meta-analyses are sometimes cited solely for their average findings (Bloom 1984; Niemiec et al. 1986). Critics have complained that meta-analysis collapses complex and subtle scholarship into single numerical representations (Cook and Leviton 1980). Such oversimplification does gross injustice to hard-fought debates in a field.

Historical accident may have fostered the idea that average effects are meta-analyses' most important products. Some early meta-analyses emphasized average results and only secondarily examined effect size variation (Cooper 1979; Rosenthal 1976). Early meta-analyses that studied effect size variation often defined their constructs broadly and thus appeared to oversimplify the reviewed literature (Smith and Glass 1977).

Ironically, meta-analyses also may appear to oversimplify a literature when they suggest a resolution to confusion in findings. For example, excitement about using simple computer applications as instructional tools for improving student achievement has not been justified by meta-analyses (Bangert-Drowns 1993; Hembree and Dessart 1986; Russell 1991). For researchers and practitioners who have committed considerable resources to such issues, or policymakers who publicly advocated some side of a debate, reviews that yield such convincing evidence may seem too simple.

Certainly meta-analysis is a method of data reduction, but it does not oversimplify a literature necessarily. In fact, most current meta-analyses examine variation in study outcomes and thus describe not just overall effect magnitude, but relations among variables. A

particular meta-analysis could be criticized for defining its domain or its constructs too broadly, analyzing data in an overly simplistic way, or only emphasizing measures of central tendency in the findings. When valid, these criticisms reflect problematic implementation rather than a fault of meta-analytic method per se.

LIMITATIONS OF META-ANALYSIS

Given that common criticisms of meta-analysis more often describe problematic implementations than the method itself, does this mean that meta-analysis is limited only by the ingenuity of the reviewer? In spite of apparent objectivity and precision in systematic, quantitative review, two fundamental factors independent of statistical issues determine the validity and replicability of meta-analytic findings. First, the conclusions of a meta-analysis reflect the many judgments of a meta-analyst as much as the reviewed literature. Second, meta-analysis depends on characteristics of the reviewed literature.

Empirical Examinations of Human Judgments and Literature Characteristics in Meta-Analysis

Several investigators looked at ways in which human judgment and literature characteristics affect the process and outcomes of meta-analysis. Steiner and colleagues (1991), for example, found 35 meta-analyses in the literature on organizational behavior and human resources management. They coded these reviews on 10 variables: degree to which the review is theory based, method for locating studies, attention to potentially unretrieved studies ("file drawer problem"), elimination of studies, assumption of independent effect sizes, control for artifacts, type of meta-analysis used, method for locating moderators, quality of data presentation, and subtlety of interpretation. Steiner and colleagues then analyzed trends among the coded features of the 35 meta-analyses.

Most of the meta-analyses did not test theoretical propositions but averaged effects for different relations under different conditions. The meta-analysts showed insufficient sensitivity to the limits of their data, making causal claims from correlational findings or claiming generalizations on the basis of small data sets. Steiner and colleagues noted time trends in meta-analytic methods. Meta-analysts combined probabilities less frequently and used methods recommended by Hunter and Schmidt (Hunter et al. 1982; Hunter and Schmidt 1990) more frequently. Meta-analysts also more regularly took one effect

size from each study to maintain the independence of their data points.

Wanous and colleagues (1989) located four pairs of meta-analyses, each pair reviewing identical topics in organizational psychology and behavior. The authors divided meta-analytic method into 11 subtasks: defining the domain, establishing inclusion criteria, searching for studies, selecting studies, extracting data, coding for independent variables, deciding whether to group independent and dependent variables, determining the mean and variance of effect sizes, deciding whether to search for moderators, selecting potential moderators, and determining means and variances for effect sizes of subgroups. According to the authors, all of these tasks except those based on numerical calculation (determining means and variances for effect sizes and deciding whether to search for moderator variables) are acts of human judgment. The authors attempted to isolate the causes of discrepant findings within each pair in terms of the 11 subtasks.

The Wanous study is a conservative test of the effects of human judgment on meta-analytic findings. Pairs were selected for conceptual similarity, so they could not differ on step 1 (defining the domain). All pairs used the same meta-analytic techniques (Hunter et al. 1982) and their overall conclusions, not the analyses of moderators, were the products that primarily were compared. In short, pairs were selected and analyzed on criteria that favored similarity to simplify comparison.

Despite the conservative features of the Wanous study, human judgment did affect meta-analytic findings. In the early phases of these meta-analyses (e.g., determining inclusion criteria, locating studies, selecting studies), reviewers created different collections of effect sizes to analyze, and these differences explained most discrepancies in findings. Some discrepancies in findings resulted from minor judgment differences, the inclusion of a single unpublished study in one case. Fortunately, the explicit nature of meta-analysis allowed Wanous and colleagues to identify the specific sources of discrepancies within pairs.

Abrami and colleagues (1988) compared six meta-analyses of the validity of student ratings of instructional effectiveness to determine causes for their discrepant conclusions. They resolved meta-analysis into five subtasks: specifying inclusion criteria, locating studies, coding study features, calculating individual study outcomes, and data analysis.

The reviews differed greatly on each subtask, even though one author produced three of the six meta-analyses. The reviewers agreed on five inclusion criteria, but irregularly employed another seven. Evaluated against an independent exhaustive search of the literature, reviews differed greatly in comprehensiveness (ranging from 20 percent to 88 percent) and in the number of studies incorrectly included. Only one meta-analysis looked for relations between study features and study outcomes. There was only 47 percent agreement among the six meta-analyses regarding which effects to include and their estimates of magnitude. Finally, the reviewers differed in the ways they analyzed the effect sizes, some using weighting, others not, some using conventional statistical tests, others checking for variance attributable to sampling error.

Matt (1989) examined one facet of one feature checked by Abrami and colleagues (1988), scrutinizing a single decision point: How does one decide which effect sizes to include when several can be obtained from one study? Matt recoded 25 studies used in Smith and Glass' (1977) psychotherapy meta-analysis, applied Smith and Glass' original decision rule (the conceptual redundancy rule), and compared it to three other decision rules (the coder agreement, outcome reliability, and outlier truncation rules). The author and two other coders independently calculated effect sizes for the 25 studies and compared them to Smith and Glass' findings.

In terms of number of effect sizes and their magnitudes, all the raters showed considerable differences; and the differences were even greater when the raters compared their results to those of Smith and Glass. This single decision point made considerable differences among the raters' outcomes. The author concluded: "Point estimates of an intervention effect have particularly captured the attention of consumers of meta-analyses. Unfortunately, such point estimates are particularly affected by variation in the mostly implicit rules regarding the selection of effect sizes within studies, and it will often be desirable to present a range of defensible and appropriate estimates based on a number of different techniques, all imperfect but with different weaknesses" (Matt 1989, p. 113).

Dependence on Human Judgment

At each phase of meta-analysis, reviewers must make significant judgments guided by common sense and informed personal preference. These decisions can affect meta-analytic outcomes and deserve careful consideration.

Formulation of the Problem. The most fundamental decisions in a review—the domain to be reviewed, the nature and breadth of the constructs and operationalizations to be considered, and the specific questions to be addressed—are all products of human judgment. These most fundamental decisions constrain all subsequent phases of a meta-analysis.

Retrieval of Studies. This stage includes three important substeps. First, a reviewer must decide the comprehensiveness of the search. Studies can be located from various sources and with varying completeness. Some reviewers limit the extent of their searches to published research, research cited in previous prominent reviews, or studies conducted after a certain date.

Once potentially useful studies are identified, they must be obtained. Actually obtaining copies of identified documents is not always possible, but this is typically a logistical problem, not an issue of reviewer judgment.

Human judgment enters this phase of review most significantly after documents are obtained. A reviewer must determine which studies to include in the review. Decisions to exclude studies are sometimes quite easy, as in cases when an obviously irrelevant study was obtained erroneously. Other inclusion criteria, such as those based on quality of research, may be more unreliable and personal. By clearly and explicitly describing search strategies and inclusion criteria, meta-analysts at least open these decisions to public scrutiny and evaluation, but such explicitness does not mitigate the effects of meta-analysts' judgments.

Coding of Study Characteristics. At least two kinds of judgment operate in the coding of study features. First, meta-analysts must choose which study characteristics will receive detailed examination. They choose these variables for many reasons. Theory or practice suggests relations between some treatment variables and effect size. Reviewers test methodological variables to see if they are confounded by study outcomes. Other variables describe the range of settings and subjects represented in the studies. Because choice of study features is the result of personal insight and preference, scholars may disagree about the most important features to select.

After features are selected, coding itself reflects many acts of judgment. Reports often lack detail or clarity and require some

guesswork to reconstruct the most probable research scenario. Some variables require personal judgment even with the clearest reports. Variables that estimate treatment intensity or qualities of interpersonal relations, for example, are difficult to code but likely to be influential in social science phenomena.

Calculation of Effect Sizes. Meta-analysts translate measures in studies to a common metric of treatment effect or relation between variables. Usually, the common measure is either the correlation coefficient or the standardized difference between two group means (i.e., the difference between group means divided by the pooled standard deviation).

Though meta-analysts agree about how to calculate effect sizes (Glass et al. 1981), meta-analysts must exercise personal judgment in deciding when to calculate them. Imagine, for example, an evaluation of a substance abuse education program that employed three measures of knowledge. One is a more reliable instrument than the others, the second provides more comprehensive coverage of the program's content, and the third is a locally developed measure and thus most likely to be sensitive to local context. Should the meta-analyst select one dependent measure that somehow provides the "best" representation of treatment effects on knowledge, average the effects measured on all three tests, or include them all in the meta-analysis? Alternatively, the reviewer could calculate effect sizes for all three and meta-analyze the dependent measures separately: a meta-analysis for most reliable measures, a meta-analysis for most comprehensive measures, and a meta-analysis for local tests.

Internal contradictions and apparent reporting errors, research biases, selective presentation of only significant findings, or extremely positive or negative scores indicate potential problems for calculation of effect sizes. The careful meta-analyst must develop consistent and reasonable strategies for treatment of each kind of problem.

Investigation of Central Tendency and Variation in Effect Sizes. If the effect sizes obtained from a group of studies were identical, there would be no need for a literature review. Generally speaking, there are two approaches to analyzing effect size variation. One can consider each effect size as an irreducible data point and treat variation among effect sizes as analogous to variation among independent subjects in primary research (Glass et al. 1981; Kulik and Kulik 1989). Reviewers who take this view tend to use conventional statistical tests for research integration. Alternatively, meta-analysts

can examine variation among effect sizes in light of the variation that one might expect from sampling error within each study (Hedges and Olkin 1985; Hunter and Schmidt 1990). Some of these researchers advocate the use of tests of homogeneity. In either case, the meta-analyst seeks to find relations between the coded study features and study effects.

Meta-analysts continue to debate the appropriateness of various analytic strategies. Assumptions of conventional statistical tests are often not met in research integration. However, meta-analytic approaches accounting for sampling error favorably weight studies with larger samples regardless of their quality. Tests of homogeneity overvalue statistical significance; statistically significant heterogeneity may be practically unimportant, and nonsignificance does not disprove heterogeneity. Some authors criticize any univariate analyses in research integration as overly simplistic and advocate multivariate analysis techniques.

At present, it is impossible to identify any one analytic strategy as trouble free. Selection of analytic method is a decision that balances the quality of available data with the various risks of alternative methods. A multimethod approach only postpones the decision. If the results of such a multimethod approach are contradictory, the reviewer then must decide which conclusions are most accurately descriptive of the literature.

Interpretation and Publication of Findings. Publication of findings requires significant decisions on the part of the reviewer. The reviewer must interpret the results of data analysis in light of the initial problem statement. Though quantitative analysis may indicate the statistical significance of relations among variables, the meta-analyst must decide which relations are practically significant for theoretical, practical, or policy implications. When several variables are significantly related to study outcomes, the meta-analyst must attempt to explain how these variables are interrelated.

Given constraints on publication space, meta-analysts cannot report many of their decisions. The meta-analyst must balance thorough and explicit exposition with conciseness and select which aspects of method will be reported. Judgments regarding publication link with another series of judgments that also determine the effectiveness of a review: the judgments of readers. The meta-analyst not only aims for accurate and valid integration, but for presentation that is both

convincing and useful for the intended audience, whether researcher, policymaker, or practitioner.

Dependence on Primary Research

Obviously, human subjectivity and judgment interject into the meta-analytic process in many ways and with significant impact. This is not to disparage meta-analysis. In spite of its efforts to be precise, comprehensive, and objective, meta-analysis is not a technical feat, but demands as much subtle expertise as any other act of scholarship.

In addition to its dependence on judgment, meta-analysis is also fundamentally dependent on the primary research it integrates. Though an obvious observation, there are a number of less obvious implications that constrain interpretations that are possible from meta-analysis.

Meta-Analysis as a Particular Form of Literature Review. At least four types of literature review can be distinguished (Cooper 1982; Jackson 1980). Meta-analysis is a quantitative form of integrative review. Integrative reviews summarize findings from numerous studies that obtain apparently contradictory results, although the studies use a consistent research design to ask the same fundamental question. The integration of many such literatures in the social sciences is an important scholarly effort. But the comprehensive, statistical integration of contradictory empirical findings, the chief purpose of meta-analysis, is not the only goal of literature review.

Reviews can have at least three other purposes. Some may highlight pioneering methodological developments or theoretical formulations, examining only preliminary research at the cutting edge. Other reviews integrate concepts that appear in disparate literatures, drawing parallels among constructs previously considered distinct or connecting distinct constructs in larger theoretical formulations. Other reviews examine evidence to confirm or refute particular theories. These types of review might benefit from statistical analysis, but they rely primarily on conceptual analysis and do not strive to resolve contradictory findings through comprehensive integration of consistent studies.

Constraints on Questions That a Meta-Analysis Can Ask. Only their resources and creativity constrain primary researchers in the kinds of theoretical, practical, or policy-related questions they can investigate. Certainly good primary research builds on relevant work that precedes it, but the researcher is relatively unfettered in developing hypotheses,

operationalizing constructs, and determining the complexity of research design.

Meta-analysts are far more constrained in their work. Meta-analysts must frame their inquiry in terms that permit the inclusion of a reasonably sized sample of studies. Meta-analysts typically frame their questions in terms of constructs frequently used in the literature of interest, and labels for these constructs and their most common operationalizations become the keywords in the search for useful studies.

Meta-analysis has some independence from primary research. Meta-analysts, for example, can integrate different literatures if some underlying construct unites them (e.g., combining teenage pregnancy prevention, smoking prevention, alcohol education, and drug prevention interventions to answer questions about public health prevention programs). Also, meta-analysts ask questions that can only be answered in a multistudy context. For example, only integrative research can ask, "Have public health prevention programs been more effective under different federal administrations?"

However, the meta-analyst cannot answer questions from literature that does not provide necessary data, and, because meta-analysis is a statistical analysis, the data must be drawn from a number of studies. Primary researchers must describe treatment and setting characteristics in sufficient detail to permit reviewers to code their salient features. Does substance abuse education affect males and females differently? The meta-analyst would be helpless to answer unless primary researchers distinguish their findings by gender.

Meta-analysts commonly conceive of effect magnitude in terms of relations between two variables, partly for the sake of simplicity of interpretation, but also because, if an effect size measures an interaction within a large group of variables, few studies will measure that same interaction. Meta-analysis then tends to favor simpler research designs that highlight comparisons between two variables at a time.

Biases in the Literature. Whole collections of studies sometimes can reflect biases that may or may not be readily detectable. Even if detectable, correction or interpretation of such biases is not always straightforward.

Meta-analysts often test for publication bias to see if study findings are related to their source. The average effect size from unpublished studies is commonly different from, and often smaller than, the average effect from published studies. It is not clear, however, why published research would yield different findings from unpublished research. Perhaps journal editors prefer statistically significant findings, thereby inadvertently elevating published treatment effects. Such a claim, though plausible, casts doubt on all scholarly publication. Alternatively, doctoral students and researchers with limited methodological experience may produce the bulk of unpublished research, while more experienced researchers publish their work. It is not possible to interpret confidently this common meta-analytic finding.

Publication bias is but one example of biases that can permeate a group of studies. Primary researchers do not research topics at random, but select ones likely to attract funding, employ constructs developed by previous successful researchers, produce statistically significant findings, and finally find publication. Such a researcher preference bias will determine whether or not there are sufficient studies to do an integrative review on a given question. Meta-analysis, and literature review in general, is by definition retrospective and therefore reflects what has been done rather than what could be done. The retrospective bias of meta-analysis may significantly misrepresent phenomena that experience rapid innovation, such as computer-based instruction.

Finally, conventions within a domain may bias findings and leave the meta-analyst helpless to correct it. For example, those who research drunk driving generally agree that rearrest rate is a problematic measure of rehabilitation effectiveness. Localities differ considerably in enforcement intensity and strategy and in the severity with which offenders are prosecuted. Even with rigorous enforcement the likelihood of arrest for driving while intoxicated is quite low. Researchers admit that rearrest rates are too insensitive to accurately measure the effects of rehabilitation programs, but rearrest is still the most commonly used measure of treatment effectiveness, primarily because it is such an attractive bottom-line measure for policymakers. It is impossible for the meta-analyst to substitute a more sensitive measure of rehabilitation effectiveness because the meta-analyst is dependent on the primary research.

Small Samples. Data drawn from the same study are not truly independent. The resources, settings, implementations, and personal

impact of the researcher leave an indelible mark on all the subjects in a particular study. Though findings might come from numerous effect sizes and thousand of subjects, the number of studies, which roughly correlates to the number of research settings, is a better indicant of the comprehensiveness of a meta-analysis.

Given this standard of comprehensiveness, most meta-analyses are based on relatively small samples. For example, the median number of studies included in 35 meta-analyses reviewed by Steiner and colleagues (1991) was 43. It is not unusual to examine hundreds of documents, but finally settle on a sample of well less than 100 studies.

Nonexperimental Design. Reviewers do not randomly sample studies or randomly assign them to conditions for comparison; they take study conditions as delivered by the primary researcher. Meta-analysis is nonexperimental correlational research, defining important relations among variables but rarely able to determine causal links. The meta-analyst may only speculate about causal relations and triangulate evidence to bolster causal claims. For example, between-study comparisons might relate peer leadership to higher effect sizes in substance education programs. If within-study treatment comparisons show the same pattern, a reviewer could claim more confidently that the nature of program leadership influences program effectiveness.

SOME CRITERIA FOR JUDGING THE QUALITY OF A META-ANALYSIS

As a quantitative integrative review, meta-analysis possesses limited aims: the integration of studies with similar research goals and methods but contradictory results. The quality of a meta-analysis is defined in part by statistical adequacy, but, perhaps even more by the reviewer's craft knowledge and constraints imposed on that craft by the available literature.

Given the importance of craft, how does one determine the quality of a quantitative review? There are some general features by which readers can evaluate the quality of a quantitative literature review.

Comprehensiveness

How was the sample of studies gathered? Some reviews limit searches to specific sources, and this should be explicitly stated in the review. How- ever, other factors being equal, readers should give greater

weight to more exhaustive reviews. Exhaustive reviews are not necessarily those with large numbers of effect sizes or subjects, but reviews that include virtually all the published and unpublished research reasonably available on the defined question. Such reviews take into account conclusions from the largest number of researchers drawn from the largest number of settings.

A comprehensive search strategy should locate studies in relevant databases and institutional clearinghouses as well as previous prominent literature reviews on the topic of interest. Inclusion criteria should be stated explicitly and reflect a balance between attention to the internal validity of the studies as well as the external validity and comprehensive-ness of the review. Exclusion of large bodies of research that may bias the outcomes of the review should be carefully evaluated.

Calculation of Effect Sizes

Effect sizes must be calculated correctly. Many meta-analyses report names, major features, and effect sizes of included studies, and readers can scan these lists for unusual outliers or noticeable errors. Authors should define explicitly how they calculated effect sizes. When the effect size is the standardized mean difference (such as Cohen's ‘d', Glass' ‘ES', or Hedges' ‘g'), effect sizes all must be standardized by raw score variation rather than variation corrected for covariance. Effect sizes calculated from corrected variances are incomparable with effect sizes from uncorrected variance. Corrections reduce raw score variance; effect sizes calculated with reduced variances will appear larger and thus spuriously appear to represent superior treatments.

Studies often offer more than one effect size either from multiple criteria or from various subdivisions of the sample. How does the meta-analyst handle multiple effect sizes? The reader should check first for the apples-and-oranges problem. A meta-analyst may define broad constructs to investigate, but combining some operationalizations, especially dependent variables, may not be defensible. For example, no common construct underlies measures of knowledge, attitude, and behavior, and an average effect across measures is difficult to interpret. Such an average might suggest that substance abuse education is highly successful when in fact it may only be successful with knowledge outcomes but not with attitude and behavior.

Readers also should check the ratio of the number of effect sizes to the number of studies. For analysis of any criterion, it is best to have nearly a one-to-one correspondence between studies and effects. If a study contributes more than one effect to analysis, those effect sizes cannot be considered independent, and studies contributing the most effect sizes are overrepresented in the calculation of averages. Occasional violations of one-to-one correspondence are permissible, but as the ratio of effects to studies increases, it becomes more difficult to interpret the analysis of effect sizes.

In some meta-analyses, effect sizes are weighted by study features such as sample size, sampling error, or quality. Such weighting strategies complicate the interpretation of meta-analytic findings. An advantage of effect size over other statistics such as ‘t' and ‘F' is precisely that it is independent of sample size; weighting by sample size or sampling error (including strategies for testing homogeneity) gives greater importance to studies with large samples regardless of the quality of their design or implementation. Weighting by quality introduces other problems. Scholars differ about how to define quality of research, but even if there were agreement in definition, there certainly would be disagreement about the appropriate weights for different qualities. In general, if weighted effect sizes are analyzed, these results should be compared to analyses of unweighted effect sizes to check if differences are meaningful or artifactual.

Analysis of Effect Size Variation

A good meta-analysis not only calculates effect sizes and their average, but attempts to identify variables that explain variation in study findings. Analysis of effect size variation poses several problems. First among these is sample size. The more variables involved in an effort to explain variation, the larger the number of effect sizes (and thus of studies) needed. Overall there should be a large ratio of effect sizes (studies) to variables examined, and categorical variables should have respectable numbers of effects in each level. Other things being equal, reviews examining the larger number of studies are better suited to investigating effect size variation.

A second problem with analysis of effect size variation is the disagreement among meta-analysts about appropriate methods. Visual methods, conventional statistical tests, tests of homogeneity, consideration of sampling error and variation due to artifacts without significance testing, and multivariate and path analytic techniques

have been recommended. Any statistical procedure could potentially be applied to research integration, so the reader must keep informed about alternate methods and judge whether a particular implementation is convincing and competent.

Interpretation of Findings

The meta-analyst must not conclude more than the data suggest. With small samples, nonrandom assignment of studies to conditions, the vagaries of human judgment, and the limitations of the literature, conclusions of meta-analysis are largely speculative, "best guesses" of treatment effects and relations among variables. Meta-analysts need to avoid making causal claims on the basis of correlational data, unless such claims are explicitly tentative or unless there are within-study comparisons that support the between-study findings.

A meta-analysis rarely completes the research in a domain and, in fact, often can raise new questions about methodology or relations among variables. An important part of the interpretative portion of a meta-analysis identifies remaining questions or new questions that require additional research.

A META-ANALYTIC VIEW OF META-ANALYTIC FINDINGS

Given the many ways in which human judgment and limitations of the literature can determine the findings of a meta-analysis, it is best to keep a meta-analytic attitude toward meta-analytic findings. That is, a careful reader should compare the findings of any given meta-analysis to the conclusions of other reviews on the same topic, looking for consistencies and inconsistencies among them. Consistencies among reviews, especially when they were independently developed or used different techniques, contribute to the confidence one can place in the findings. The reader should attempt to locate reasons for inconsistencies in findings and either resolve the inconsistency or leave the debate for further primary research.

REFERENCES

Abrami, P.C.; Cohen, P.A.; and d'Apollonia, S. Implementation problems in meta-analysis. *Rev Educ Res* 58(2):151-179, 1988.

Bangert-Drowns, R.L. Review of developments in meta-analytic method. *Psychol Bull* 99(3):388-399, 1986.

Bangert-Drowns, R.L. The word processor as an instructional tool: A meta-analysis of word processing in writing instruction. *Rev Educ Res* 63(1):69-93, 1993.

Bloom, B.S. The 2 sigma problem: The search for methods of instruction as effective as one-to-one tutoring. *Educ Res* 13(6):4-16, 1984.

Carlberg, C.G.; Johnson, D.W.; Johnson, R.; Maruyama, G.; Kavale, K.; Kulik, C.-L.C.; Kulik, J.A.; Lysokowski, R.S.; Pflaum, S.W.; and Walberg, H.J. Meta-analysis in education: A reply to Slavin. *Educ Res* 13(8):16-23, 1984.

Clark, R.E. Confounding in educational computing research. *J Educ Comput Res* 1(2):137-148, 1985.

Cook, T., and Leviton, L. Reviewing the literature: A comparison of traditional methods with meta-analysis. *J Pers* 48(4): 449-472, 1980.

Cooper, H.M. Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *J Pers Soc Psychol* 37(1):131-146, 1979.

Cooper, H.M. Scientific guidelines for conducting integrative reviews. *Rev Educ Res* 52(2):291-302, 1982.

Eysenck, H.J. An exercise in mega-silliness. *Am Psychol* 33(5):517, 1978.

Gallo, P. Meta-analysis—mixed meta-phor? *Am Psychol* 3(5):515-517, 1978.

Glass, G.V. Primary, secondary, and meta-analysis of research. *Educ Res* 10(5):3-8, 1976.

Glass, G.V.; McGaw, B.; and Smith, M.L. *Meta-Analysis in Social Research*. Beverly Hills, CA: Sage Publications, 1981.

Guzzo, R.A.; Jackson, S.E.; and Katzell, R.A. Meta-analysis. In: Cummings, L.L., and Staw, B.M., eds. *Research in Organizational Behavior*. Greenwich, CT: JAI Press, 1987. pp. 407-422.

Hedges, L.V. How hard is hard science, how soft is soft science? *Am Psychol* 42(5):443-455, 1987.

Hedges, L.V., and Olkin, I. *Statistical Methods for Meta-Analysis*. San Diego, CA: Academic Press, 1985.

Hembree, R., and Dessart, D.J. Effects of hand-held calculators in precollege mathematics education: A meta-analysis. *J Res Math Educ* 17(2):83-99, 1986.

Hunter, J.E., and Schmidt, F.L. *Methods of Meta-Analysis*. Newbury Park, CA: Sage Publications, 1990.

Hunter, J.E.; Schmidt, F.L.; and Jackson, G.B. *Meta-Analysis: Cumulating Research Findings Across Studies*. Beverly Hills, CA: Sage Publications, 1982.

Jackson, G.B. Methods for integrative reviews. *Rev Educ Res* 50(3):438-460, 1980.

Kulik, J.A., and Kulik, C.-L.C. Meta-analysis in education. *Int J Educ Res* 13(3):221-340, 1989.

L'Hommedieu, R.; Menges, R.J.; and Brinko, K.T. Validity issues in meta-analysis: Suggestions for research and policy. *Higher Educ Res Dev* 7(2):119-130, 1988.
Matt, G.E. Decision rules for selecting effect sizes in meta-analysis: A review and reanalysis of psychotherapy outcome studies. *Psychol Bull* 105(1):106-115, 1989.
Niemiec, R.P.; Blackwell, M.C.; and Walberg, H.J. CAI can be doubly effective. *Phi Delta Kappan* 67(10):750-751, 1986.
Presby, S. Overly broad categories obscure important differences between therapies. *Am Psychol* 33(5):514-515, 1978.
Rosenthal, R. Interpersonal expectancy effects: A follow-up. In: Rosenthal, R., ed. *Experimental Effects in Behavioral Research.* New York: Irvington, 1976. pp. 440-471.
Russell, R.G. "A Meta-Analysis of Wordprocessing and Attitudes and the Impact on the Quality of Writing." Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1991.
Schmidt, F. What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *Am Psychol* 47(10):1173-1181, 1992.
Slavin, R.E. Meta-analysis in education: How has it been used? *Educ Res* 13(8):6-15, 1984.
Smith, M.L., and Glass, G.V. Meta-analysis of psychotherapy outcome studies. *Am Psychol* 32(9):752-760, 1977.
Steiner, D.D.; Lane, I.M.; Dobbins, G.H.; Schnur, A.; and McConnell, S. A review of meta-analyses in organizational behavior and human resources management: An empirical assessment. *Educ Psychol Meas* 51(3):609-626, 1991.
Wachter, K.W. Disturbed by meta-analysis? *Science* 241(4872):1407-1408, 1988.
Wanous, J.P.; Sullivan, S.E.; and Malinak, J. The role of judgment calls in meta-analysis. *J Appl Psychol* 74(2):259-264, 1989.

AUTHOR

Robert L. Bangert-Drowns, Ph.D.
Associate Professor
State University of New York at Albany
School of Education
1400 Washington Avenue
Albany, NY 12222

Click here to go to page 253