Best-Evidence Synthesis: An Alternative to Meta-Analytic and Traditional Reviews

ROBERT E. SLAVIN
Johns Hopkins University

ABSTRACT: This paper proposes an alternative to both meta-analytic and traditional reviews. The method, "best-evidence synthesis," combines the quantification of effect sizes and systematic study selection procedures of quantitative syntheses with the attention to individual studies and methodological and substantive issues typical of the best narrative reviews. Best-evidence syntheses focus on the "best evidence" in a field, the studies highest in internal and external validity, using well-specified and defended a priori inclusion criteria, and use effect size data as an adjunct to a full discussion of the literature being reviewed.

n the decade since Glass (1976) introduced the concept of meta-analysis as a means of combining results of different investigations on a related topic, the practice and theory of literature synthesis has been dramatically transformed. Scores of meta-analyses relating to educational practice and policy have appeared, and the number of articles using or discussing meta-analysis in education has approximately doubled each year from 1979 to 1983 (S. Jackson, 1984). Several thoughtful guides to the proper conduct of meta-analyses have been recently published (see, e.g., Cooper, 1984; Glass, McGaw, & Smith, 1981; Hunter, Schmidt, & Jackson, 1982; Light & Pillemer, 1984; Rosenthal, 1984).

Robert E. Slavin is Director, Elementary School Program, Center for Research on Elementary and Middle Schools, Johns Hopkins University, Baltimore, MD 21218. His specializations are cooperative learning, school and classroom organization, field research methods, and research review.

An earlier version of this paper was presented at the 1985 annual meeting of the American Educational Research Association, Chicago. This paper was written under grants from the National Institute of Education (No. NIE-G-83- 0002) and the Office of Educational Research and Improvement (No. OERI-G-86-0006). However, the opinions expressed do not necessarily represent Department of Education policy. I would like to thank Harris Cooper, Gary Gottfredson, Nancy Madden, Robert Stevens, and Noreen Webb for their helpful comments on earlier drafts of this paper.

Ever since it was introduced, meta-analysis has been vigorously criticized, and equally vigorously defended. In considering arguments for and against this procedure in the abstract, there is much validity to both sides. Proponents of quantitative synthesis (e.g., Cooper, 1984; Glass et al., 1981; G. Jackson, 1980; Light & Pillemer, 1984) are certainly correct to criticize traditional reviews for using unsystematic and poorly specified criteria for including studies and for using statistical significance as the only criterion of treatment effects. Critics of these procedures (e.g., Cook & Leviton, 1980; Eysenck, 1978; Slavin, 1984; Wilson & Rachman, 1983) are equally justified in objecting to a mechanistic approach to literature synthesis that sacrifices most of the information contributed in the original studies and includes studies of questionable methodological quality and questionable relevance to the issue at hand.

In an earlier article (Slavin, 1984), I evaluated the actual practice of meta-analysis in education by examining eight meta-analyses conducted by six independent sets of investigators, comparing their procedures and conclusions against the studies they included. I found that all of these meta-analyses had made errors serious enough to invalidate or call into question one or more major conclusions. In reviewing several meta-analyses published after my article went to press, I have seen misapplications of the procedure that are at least as serious (Slavin, 1985). Yet the misuses of meta-analysis in education do not in themselves justify a return to traditional review procedures.

In this paper, I propose an alternative to both meta-analytic and traditional reviews that is designed to draw on the strengths of each approach and to avoid the pitfalls characteristic of each. The main idea behind this procedure, which I call "best-evidence synthesis," is to add to the traditional scholarly literature review application of rational, systematic methods of selecting studies to be included and use of effect size (rather than statistical significance alone) as a common metric for treatment effects.

The Principle of Best Evidence

In law, there is a principle that the same evidence that would be essential in one case might be disregarded in another because in the second case there is better evidence available. For example, in a case of disputed authorship, a typed manuscript might be critical evidence if no handwritten copy is available, but if a handwritten copy exists, the typed copy would be inadmissible because it is no longer the best evidence (since the handwritten copy would be conclusive evidence of authorship).

I would propose extending the principle of best evidence to the practice of research review. For example, if a literature contains several studies high in internal and external validity, then lower quality studies might be largely excluded from the review. Let's say we have a literature with 10 randomized studies of several months' duration evaluating Treatment X. In this case, results of correlational studies, small-sample studies, and/or brief experiments might be excluded, or at most briefly mentioned. For example, Ottenbacher and Cooper (1983) located 61 randomized, double-blind studies of effects of medication on hyperactivity, and therefore decided not to include studies of lower methodological rigor. However, if a set of studies high in internal and external validity does not exist, we might cautiously examine the less well designed studies to see if there is adequate unbiased information to come to any conclusion.

The principle of best evidence works in law because there are a priori criteria for adequacy of evidence in certain types of cases. Comparable criteria could not be prescribed for all of educational research, but could be proposed for each subfield as it is reviewed. These criteria might be derived from a reading of previous narrative and meta-analytic reviews and a preliminary search of the literature.

Justification for the "Best Evidence" Principle

The recommendation that reviewers apply consistent, well justified, and clearly stated a priori inclusion criteria is at the heart of the best-evidence synthesis, and differs from the exhaustive inclusion principle suggested by Glass et al. (1981) and others, who recommend including all studies that meet broad standards in terms of independent and dependent variables, avoiding any judgments of study quality. Proponents of meta-analysis suggest that statistical tests be used to empirically test for any effects of design features on study outcomes. The rationale given for including all studies regardless of quality rather than identifying the methodologically adequate ones is primarily that the reviewer's own biases may enter into decisions about which studies are "good" and which are "bad" methodologically. Certainly, studies of interjudge consistency in evaluations of journal articles (e.g., Gottfredson, 1978; Marsh & Ball, 1981; Peters & Ceci, 1982; Scarr & Weber, 1978) show considerable variation from reviewer to reviewer, so global decisions about methodological quality are inappropriate as a priori criteria for inclusion of studies in a research synthesis. It is important to recall that much of the impetus for the development of meta-analysis came from a frequent observation that traditional narrative reviews were unsystematic in their selection of studies, and did a poor job (or no job at all) of justifying their selection of studies, arguably the most important step in the review process (see Cooper, 1984; G. Jackson, 1980; Waxman & Walberg, 1982).

However, while it is difficult to justify a return to haphazard study selection procedures characteristic of many narrative reviews, it is also difficult to accept the meta-analysts' exhaustive inclusion strategy.

The rationale for exhaustive inclusion depends entirely on the proposition that specific methodological features of studies can be statistically compared in terms of their effects on effect size. Cooper (1984) puts the issue this way:

> If it is empirically demonstrated that studies using "good" methods produce results different from "bad" studies, the results of the good studies can be believed. When no difference is found it is sensible to retain the "bad" studies because they contain other variations in methods (like different samples and locations) that, by their inclusion, will help solve many other questions surrounding the problem area. (pp. 65–66)

In practice, meta-analyses almost always test several methodological and substantive characteristics of studies for correlations with effect size, using a criterion for rejecting the null hypothesis of no differences of .05. However, in order to justify pooling across categories of studies, the meta-analyst must prove the null hypothesis that the categories do not differ. This is logically impossible, and in situations in which the numbers of studies are small and the numbers of categories are large, finding true differences between categories of studies to be statistically significant is unlikely.

One example of this is a recent meta-analysis on adaptive education by Waxman, Wang, Anderson, and Walberg (1985), which coded the critical methodological factor "control method" into eight categories: unspecified, stratification, partial correlation, beta weights in regression, raw or metric weights in regression, factorial analysis of variance, analysis of covariance, or none. In a meta-analysis of only 38 studies, the 8 × 1 ANOVA apparently used to evaluate effects of methodological quality on study outcome had highly unequal and small cell sizes and an extremely high probability of failing to detect any true differences.

The problem of the reviewer's bias entering into inclusion decisions is hardly solved by exhaustive inclusion followed by statistical tests. The reviewer's bias may just as well enter into the coding of studies for statistical analysis (Mintz, 1983; Wilson & Rachman, 1983). Worse, the reader has no easy way to find out how studies were coded. For example, most of the studies coded as "randomly assigned" in a meta-analysis on mainstreaming by Carlberg and Kavale (1980) were in fact randomly selected from nonrandomly assigned groups. To discover this, it was necessary to obtain every article cited and laboriously recode them (Slavin, 1984).

Reviews of social science literature will inevitably involve judgment. No set of procedural or statistical canons can make the review process immune to the reviewer's biases. What we can do, however, is to require that reviewers make their procedures explicit and open, and we can ask that reviewers say enough about the studies they review to give readers a clear idea of what the original evidence is. The greatest problem with exhaustive inclusion is that it often produces such a long list of studies that the reviewer cannot possibly describe each one. I would argue that all other things being equal, far more information is extracted from a large literature by clearly describing the best evidence on a topic than by using limited journal space to describe statistical analyses of the entire methodologically and substantively diverse literature.

> *. . . far more information is extracted from a large literature by clearly describing the best evidence on a topic than by using limited journal space to describe statistical analyses of the entire methodologically and substantively diverse literature.*

Criteria for Including Studies

Obviously, if a priori criteria are to be used to select studies, these criteria must be well thought out and well justified. It is not possible to specify in advance what criteria should be used, as this must depend on the purposes for which the review is intended (see Light & Pillemer, 1984, for more on this point). However, there are a few principles that probably apply generally.

First, the most important principle of inclusion must be germaneness to the issue at hand. For example, a meta-analysis focusing on school achievement as a dependent measure must explicitly describe what is meant by school achievement and must only include studies that measured what is commonly understood as school achievement on individual assessments, not swimming, tennis, block stacking, time-on-task, task completion rate, group productivity, attitudes, or other measures perhaps related to but not identical with student academic achievement (see Slavin, 1984).

Second, methodological adequacy of studies must be evaluated primarily on the basis of the extent to which the study design minimized bias. For example, it would probably be inappropriate to exclude studies because they failed to document the reliability of their measures, as unreliability of measures is unlikely in itself to bias a study's results in favor of the experimental or control group. On the other hand, great caution must be exercised in areas of research in which less-than-ideal research designs tend to produce systematic bias. For example, matched or correlational studies of such issues as special education, non-promotion, and gifted programs are likely to be systematically biased in favor of the students placed in regular classes, promoted, or placed in gifted classes, respectively (Madden & Slavin, 1983). In these areas of research, the independent variable is strongly correlated with academic ability, motivation, and many other factors that go into a decision to, for example, promote or retain a student.

Controlling for all these factors is virtually impossible in a correlational study. In research literatures of this kind, random assignment to experimental or control groups is essential. However, in other areas of research, the independent variable is less highly correlated with academic ability or other biasing factors. For example, schools that use tracking may not be systematically different from those that do not. If this is the case, then random assignment, though still desirable, may be less essential; carefully matched or statistically controlled studies may be interpretable.

Third, it is important to note that external validity should be valued at least as highly as internal validity in selecting studies for a best-evidence synthesis. For example, reviews of classroom practices should not generally include extremely brief laboratory studies or other highly artificial experiments. Often, a search for randomized studies turns up such artificial experiments. This was the case with the Glass, Cohen, Smith, and Filby (1982) class size meta-analysis, which found more positive effects of class size in "well controlled" studies than in "less well controlled" studies. Well controlled meant studies using random assignment, but this requirement caused the well controlled study category to include a number of extremely brief artificial experiments, such as a 30-minute study of class size by Moody, Bausell, and Jenkins (1973), as well as a study of effects of class size on tennis "achievement" (Verducci, 1969). Because class size is not strongly correlated with academic ability (see Coleman et al., 1966), this is actually a case in which well designed correlational studies, because of their greater external validity, might be preferred to many of the randomized experimental studies.

One category of studies that may be excluded in some literatures is studies with very small sample sizes. Small samples are generally susceptible to unstable effects. In education, experiments involving small numbers of classes are particularly susceptible to teacher and class effects (see Glass & Stanley, 1970; Page, 1975). For example, if Mr. Jones teaches Class A using Method *X* and Ms. Smith teaches Class B using Method *Y*, there is no way to rule out the possibility that any differences between the classes are due to differences in teaching style or ability between Mr. Jones and Ms. Smith (teacher effects) or to effects of students in the different classes on one another (class effects) rather than to any differences between Methods *X* and *Y*. To minimize these possibilities, a criterion of a certain number of teachers, classes, and/or students in each treatment group might be established.

In some literatures lacking a body of studies high in internal and external validity, it may be necessary to include (but not pool) germane studies using several methods, each of which has countervailing flaws. For example, if a literature on a particular topic consists largely of randomized experiments low in external validity and correlational studies high in external validity but susceptible to bias, the two types of research might be separately reviewed. If the two groups of studies yield the same result, each buttresses the other. If they yield different results, the reviewer should explain the discrepancy.

Finally, it may be important in some literatures to mention the best designed studies excluded from the review (that is, those that "just missed") to give the reader a more concrete idea of why a study was excluded and what the consequences of that exclusion are. For example, one recent meta-analysis of studies of bilingual education by Willig (1985) devoted considerable attention to describing studies excluded from the review, making the criteria for inclusion clear.

Some arbitrary limitations often placed on inclusion of studies in traditional reviews make little sense, and should be abandoned. Perhaps most common is the elimination of dissertations and unpublished reports (such as government reports or university technical reports). Often, these unpublished reports are better designed than published ones; for example, it may sometimes be easier to get a poorly designed study into a low quality journal than to get it past a dissertation committee. The most important randomized study of special education versus mainstream placement (Goldstein, Moss, & Jordan, 1966) and the Coleman Report (Coleman et al., 1966) are two examples of unpublished government reports essential to their respective literatures.

On the other hand, meta-analyses also exclude one type of study that should not be excluded: studies in which effect sizes cannot be computed. It often happens that studies fail to report standard deviations or other information sufficient to enable computation of effect sizes. While effect sizes can be computed directly from *t*-scores, *F*'s, or *p* values for two-group comparisons if *N*'s are known (see Glass et al., 1981), there are cases in which important, well designed studies present only *p* values or *F*'s for complex designs, ANCOVAs, or multiple regression analyses with too little information to allow for computation of effect sizes. Yet there is no good reason to exclude these studies from consideration solely on this basis.

Exhaustive Literature Search

Once criteria for inclusion of studies in a best-evidence synthesis have been established, it is incumbent upon the reviewer to locate every study ever conducted that meets these criteria. Books on meta-analysis (e.g., Cooper, 1984; Light & Pillemer, 1984) give useful suggestions for conducting literature searches using ERIC, Psychological Abstracts, Social Science Citation Index, and bibliographies of other reviews or meta-analyses, among other sources. In some cases, it is necessary to write to authors to request means and standard deviations or other information necessary to understand some aspect of a study. It is particularly important to locate all studies cited by previous reviewers to assure the reader that any differences in conclusions between reviewers are not simply due to differences in the pool of studies located.

Computation of Effect Sizes

In general, effect sizes should be computed as suggested by Glass et al. (1981), with a correction for sample size devised by Hedges (1981; Hedges & Olkin, 1985). The Hedges procedure produces an unbiased estimate of effect size, reducing estimates from studies with total *N*'s (experimental plus control) less than 50.

There are many statistical issues that are important in computing and understanding effect sizes, and many of these have important substantive implications. For example, there are questions of how to interpret gain scores or posttests adjusted for covariates, how to deal with unequal pretest scores in experimental and control groups, and how to deal with aggregated data (e.g., class or school means). Readers interested in statistical issues should refer to the excellent books on the conduct of quantitative syntheses (e.g., Cooper, 1984; Glass et al., 1981; Hedges & Olkin, 1985; Hunter et al., 1982; Rosenthal, 1984).

Averaging effect sizes within studies. Since many studies report a large number of effects, it may be important to compute averages of some effect sizes across particular subsets of comparisons. The amount of averaging to be done depends on the purpose and focus of the best-evidence synthesis. For example, in a general review of the effects of ability grouping on achievement, different measures of reading and language arts might be averaged. However, in a best-evidence synthesis of research on specific reading strategies, we would want to preserve information separately for reading comprehension, reading vocabulary, oral reading, language mechanics, and so on.

Similarly, in a review of effects of computer-assisted instruction we might average effects for students of different ethnicities, but in a review of compensatory education, separate effects for different ethnic groups might be preserved. However, when pooling effect sizes across studies, each study (or each experimental-control comparison) must count as one observation with effect sizes from similar measures averaged as appropriate. To count each dependent measure as a separate effect size for pooling purposes, as recommended by Glass et al. (1981), creates serious problems as it gives too much weight to studies with large numbers of measures and comparisons and violates assumptions of independence of data points in any statistical analyses (see Bangert-Drowns, 1986).

Table of Study Characteristics and Effect Sizes

No matter how extensive the literature reviewed, all studies should be listed in a table specifying major design and setting variables and effect sizes for principal studies. This table should include the names of the studies, sample size, duration, research design, subject matter, grade levels, treatments compared, and effect size(s). Other information important in a particular area of research might also be included. For example, the table might indicate which effects were statistically significant in the original research. This table is essential not only in summarizing all pertinent information, but also in making it easier to check the review's procedures and conclusions against the original research on which it was based.

In the table of study characteristics and effects sizes, results from studies for which effect sizes could not be computed may be represented as "+" (statistically significant-positive), "0" (no significant differences), or "–" (statistically significant-negative).

For examples of tables of study characteristics and effect sizes, see Willig (1985), Schlaefli, Rest and Thoma (1985), Kulik and Kulik (1984), and Slavin (1986).

Pooling of Effect Sizes

When there are many studies high in internal and external validity on a well defined topic, pooling (averaging) effect sizes across the various studies may be done. For example, let's say we located a dozen studies of Treatment *X* in which experimental and control students (or classes) were randomly assigned to treatment groups, the treatment was applied for at least 3 weeks, and fair achievement tests equally responsive to the curriculum taught in the experimental and control groups were used. In this case, we might pool the effect sizes by computing a median across the 12 studies. Medians are preferable to means because they are minimally influenced by anomalous outliers frequently seen in meta-analyses.

In pooling effect sizes, the reviewer must be careful "not to quantitatively combine studies at a broader conceptual level than the readers would find useful" (Cooper, 1984, p. 82). For example, in a quantitative synthesis by Lysakowski and Walberg (1982), it was not useful to pool across studies of cues, participation, and corrective feedback, as these topics together do not form a single well-defined category (see Slavin, 1984).

Pooled effect sizes should be reported as adjuncts to the literature review, not its primary outcome. Pooling and statistical comparisons must be guided by substantive, methodological, and theoretical considerations, not conducted wholesale and interpreted according to statistical criteria alone. For example, many meta-analyses routinely test for differences among effect sizes according to year of publication, a criterion that may be important in some literatures but is meaningless in others, while ignoring more theoretically or methodologically important comparisons (such as plausible interactions among study features).

Pooled effect sizes should never be treated as the final word on a subject. If pooled effects are markedly different from those of two or three especially well designed studies, this discrepancy should be explained. Pooling has value simply in describing the central tendency of several effects that clearly tend in the same direction. When effects are diverse, or the number of methodologically adequate, germane articles is small, pooling should not be done. Hedges and Olkin (1985) have described statistical procedures for testing sets of effect sizes for homogeneity, and these may be useful in determining whether or not pooling is indicated. However, decisions about which studies to include in a particular category should be based primarily on substantive, not statistical criteria.

Literature Review

The selection of studies, computation of effect sizes, and pooling described above are only a preliminary to the main task of a best-evidence synthesis: the literature review itself. It is in the literature review section that best-evidence synthesis least resembles meta-analysis. For example, some quantitative syntheses do use a priori selection, do present tables of study characteristics and effect size, and do follow other procedures recommended for best-evidence synthesis, but it is very unusual for a quantitative synthesis to discuss more than two or three individual studies or to examine a literature with the care typical of the best narrative reviews.

There are no formal guidelines or mechanistic procedures for conducting a literature review in a best-evidence synthesis; it is up to the reviewer to make sense out of the best available evidence.

Formats for Best-Evidence Syntheses

No rigid formula for presenting best-evidence syntheses can be prescribed, as formats must be adapted to the literature being reviewed. However, one suggestion for a general format is presented below. Also, see Slavin (1986) for an example of a best-evidence synthesis.

Introduction. The introduction to a best-evidence synthesis will closely resemble introductions to traditional narrative reviews. The area being studied is introduced, key terms and concepts are defined, and the previous literature, particularly earlier reviews and meta-analyses, is discussed.

Methods. In a best-evidence synthesis, the methods section serves primarily to describe how studies were selected for inclusion in the review. The methods section might consist of the following three subsections.

Best-Evidence Criteria describes and justifies the study selection criteria employed. Clear, quantifiable criteria must be specified, not global ratings of methodological adequacy. Stringent criteria for germaneness should be applied (e.g., studies of individualized instruction in mathematics that took place over periods of at least 8 weeks in elementary schools, using mathematics achievement measures not specifically keyed to the material being studied in the experimental classes). Among germane studies, criteria for methodological adequacy are established, focusing on avoidance of systematic bias (e.g., use of random assignment or matching with evidence of initial equality), sample size (e.g., at least four classes in experimental and control groups), and external validity (e.g., treatment duration of at least eight weeks). The literature search procedure should be described in enough detail that the reader could theoretically regenerate an identical set of articles. A section titled *Studies Selected* might describe the set of studies that will constitute the synthesis, while a section on *Studies Not Selected* characterizes studies not included in the synthesis, in particular describing excluded studies that were included in others' reviews and studies that "just missed" being included.

Literature Synthesis. The real meat of the best-evidence synthesis is in the *Literature Synthesis* section. This is where the research evidence is actually reviewed. This section would first present and discuss the table of study characteristics and effect sizes and discuss any issues related to the table and its contents. If pooling is seen as appropriate, the results of the pooling are described; otherwise, the rationale for not pooling is presented.

In a meta-analysis, the presentation of the "results" is essentially the end point of the review. In a best-evidence synthesis, the table of study characteristics and effect sizes and the results of any pooling are simply a point of departure for an intelligent, critical examination of the literature (see Light & Pillemer, 1984). In the Literature Synthesis section, critical studies should be described and important conceptual and methodological issues should be explored. A best-evidence synthesis should not read like an annotated bibliography, but should use the evidence at hand to answer important questions about effects of various treatments, possible conditioning or mediating variables, and so on. When conclusions are suggested, they must be justified in light of the available evidence, but also the *contrary* evidence should be discussed. Effect size information may be incorporated in the Literature Synthesis, as in the following example:

"Katz and Jammer (19XX) found significantly higher achievement in project classes than in control classes on mathematics computations (ES = .45) and concepts (ES = .31), but not on applications (ES = .02)."

In general, the "best-evidence" studies should be described with particular attention to studies with outstanding features, unusually high or low effect sizes, or important additional data. Studies that meet standards of germaneness and methodological adequacy but do not yield effect size data should be discussed on the same basis as those that do yield effect size data. Studies excluded from the main synthesis may be brought in to illustrate particular points or to provide additional evidence on a secondary issue. Except for the references to effect sizes, the bulk of the Literature Synthesis should look much like the main body of any narrative literature review.

One useful activity in many best-evidence syntheses is to compare review-generated and study-generated evidence (see Cooper, 1984). Review-generated evidence results from comparisons of outcomes in studies falling into different categories, while study-generated evidence relates to comparisons made within the same studies. For example, a reviewer might find an average effect size of 1.0 in methodologically adequate studies of Treatment *X*, and 0.5 in similar studies of Treatment *Y* and conclude that Treatment *X* is more effective than Treatment *Y*. However, this is not necessarily so, as other factors that are systematically different in studies of the two treatments could account for the apparent difference. This issue could be substantially informed by examination of studies that specifically compared treatments *X* and *Y*. If such studies exist and are of good quality, they would constitute the best evidence for the comparison of the treatments. Review-generated evidence can be useful in *suggesting* comparisons to be sought within studies, and may often be the only available evidence on a topic, but is rarely conclusive in itself.

Conclusions. One purpose of any literature review is to summarize the findings from large literatures to give readers some indication of where the weight of the evidence lies. A best-evidence synthesis should produce and defend conclusions based on the best available evidence; or in some cases may conclude that the evidence currently available does not allow for any conclusions.

Summary

The advent of meta-analysis has had an important positive impact on research synthesis in reopening the question of how best to summarize the results of large literatures and providing statistical procedures for computation of effect size, a common metric of treatment effects. It is difficult to justify a return to reviews with arbitrary study selection procedures and reliance on statistical significance as the only criterion for treatment effects. Yet in actual practice (at least in education), meta-analysis has produced serious errors (see Slavin, 1984).

This paper proposes one means, best-evidence synthesis, of combining the strengths of meta-analytic and traditional reviews. Best-evidence synthesis incorporates the quantification and systematic literature search methods of meta-analysis with the detailed analysis of critical issues and study characteristics of the best traditional reviews in an attempt to provide a thorough and unbiased means of synthesizing research and providing clear and useful conclusions. No review procedure can make errors impossible or eliminate any chance that reviewers' biases will affect the conclusions drawn. It may be that applications of the procedures proposed in this paper will still lead to errors as serious as those often found in meta-analytic and traditional reviews. However, applications of best-evidence synthesis should at least make review procedures clear to the reader and should provide the reader with enough information about the primary research on which the review is based to reach independent conclusions.

References

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

Carlberg, C. , & Kavale, K. (1980). The efficacy of special versus regular class placement for exceptional children: A meta-analysis. *Journal of Special Education, 14,* 295–309.

Coleman, J. S., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F., & York, R. (1966). *Equality of educational opportunity.* Washington, DC: U.S. Department of Health, Education, and Welfare.

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

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

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist, 33,* 517.

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

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

Glass, G., Cohen, L., Smith, M. L. & Filby, N. (1982). *School class size.* Beverly Hills, CA: Sage.

Glass, G. & Stanley, J.C. (1970). *Statistical methods in education and psychology.* Englewood Cliffs, NJ: Prentice-Hall.

Goldstein, H., Moss, J., & Jordan, J. (1966). *The efficacy of special class training on the development of mentally retarded children* (Cooperative Research Project no. 619). Washington, DC: U.S. Office of Education.

Gottfredson, S. (1978). Evaluating psychological research reports. *American Psychologist, 33,* 920–934.

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

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies.* Beverly Hills, CA: Sage.

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 annual convention of the American Psychological Association, Toronto.

Kulik, J. A., & Kulik, C. L. (1984). Effects of accelerated instruction on students. *Review of Educational Research, 54,* 409–425.

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

Lysakowski, R., & Walberg, H. (1982). Instructional effects of cues, participation, and corrective feedback: A quantitative synthesis. *American Educational Research Journal, 19,* 559–578.

Madden, N. A., & Slavin, R. E. (1983). Mainstreaming students with mild academic handicaps: Academic and social outcomes. *Review of Educational Research, 53,* 519–569.

Marsh, H., & Ball, S. (1981). Interjudgmental reliability of reviews for the *Journal of Educational Psychology. Journal of Educational Psychology, 73,* 872–880.

Mintz, J. (1983). Integrating research evidence: A commentary on meta-analysis. *Journal of Consulting and Clinical Psychology, 51,* 71–75.

Moody, W. B., Bausell, R. B., & Jenkins, J. R. (1973). The effect of class size on the learning of mathematics: A parametric study with fourth grade students. *Journal for Research in Mathematics Education, 4,* 170–176.

Ottenbacher, R.J., & Cooper, H.M. (1983). Drug treatment of hyperactivity in children. *Developmental Medicine and Child Neurology, 25,* 358–366.

Page, E. (1975). Statistically recapturing the richness within the classroom. *Psychology in the Schools, 12,* 339–344.

Peters, D., & Ceci, S. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. *The Behavioral and Brain Sciences, 5,* 187–255.

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

Scarr, S., & Weber, B. (1978). The reliability of reviews for the *American Psychologist. American Psychologist, 33,* 935.

Schlaefli, A., Rest, J. R., & Thoma, S. J. (1985). Does moral education improve moral judgment? A meta-analysis of intervention studies using the defining issues test. *Review of Educational Research, 55,* 319–352.

Slavin, R. E. (1984). Meta-analysis in education: How has it been used? *Educational Researcher, 13* (8), 6–15, 24–27.

Slavin R. E. (1985, March). *Quantitative review.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Slavin, R. E. (1986). *Ability grouping and student achievement in elementary schools: A best-evidence synthesis* (Tech. Rep. No. 1). Baltimore, MD: Center for Research on Elementary and Middle Schools, Johns Hopkins University.

Verducci, F. (1969). Effects of class size on the learning of a motor skill. *Research Quarterly, 40,* 391–395.

Waxman, H., & Walberg, H. (1982). The relation of teaching and learning: A review of reviews of process-product research. *Contemporary Education Review, 1,* 103–120.

Waxman, H.C., Wang, M. C., Anderson, K. A., & Walberg, H.J. (1985). Adaptive education and student outcomes: A quantitative synthesis. *Journal of Educational Research, 78,* 228–236.

Willig, A.C. (1985). A meta-analysis of selected studies on the effectiveness of bilingual education. *Review of Educational Research, 55,* 269–317.

Wilson, G.T., & Rachman, S.J. (1983). Meta-analysis and the evaluation of psychotherapy outcome: Limitations and liabilities. *Journal of Consulting and Clinical Psychology, 51,* 54–64.

Clarification

On p. 21 of the October *ER,* an alternate sigma (ς) symbol was used in the equations below. For those readers who may have been confused by this symbol, the equations now appear with the usual sigma (σ) symbol. In addition, an extraneous subscript p appeared in the first equation.

$E(e_i)=0,\ \sigma^2_{e_i}=(1-\varrho^2_i)^2/(N-1),\ \sigma^2_r=\sigma^2_\varrho+\sigma^2_e.$

"$\sigma^2_{e_c}=\sigma^2_e/r_{xx}r_{yy}$" (p. 56); "$E(d)=\delta$" (p.101);

"$\sigma^2_e=4(1+\delta^2/8)/N$" (p. 101).