ED 297 015 TM 011 982

AUTHOR	Kulik, James A.; Kulik, Chen-Lin C.
TITLE	Meta-analysis: Historical Origins and Contemporary Practice.
PUB DATE	Apr 88
NOTE	39p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150) -- Information Analyses (070)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Generalization; Literature Reviews; *Meta Analysis; *Research Methodology; *Statistical Analysis
IDENTIFIERS	*Historical Background

ABSTRACT

The early and recent history of meta-analysis is outlined. After providing a definition of meta-analysis and listing its major characteristics, developments in statistics and research are described that influenced the formulation of modern meta-analytic methods. Major meta-analytic methods currently in use are described. Statistical and other research developments contributing to meta-analysis include the introduction of combined tests, combined treatment effects, use of percentages as outcome variables, and use of correlations as outcomes. Meta-analytic approaches reviewed include Glass' methodology, Hedges' modern statistical methods, Hunter and Schmidt's validity generalization, and Rosenthal's methods. Problems affecting meta-analysis include inflated sample sizes, non-independent measures in statistical analyses, the failure to take experimental design into account when estimating effect sizes and sampling errors, and the development of inappropriate statistical methods for testing the influence of study features on study outcomes. Four tables and two graphs are included. (TJH)

Meta-analysis: Historical Origins And Contemporary Practice

James A. Kulik & Chen-Lin C. Kulik

The University of Michigan

A Paper Presentation at the 1988 Meeting

Of the American Educational Research Association

New Orleans

Meta-analysis has a long past and a short history. Its history begins in 1976 when Glass first used the term in his presidential address to the American Educational Research Association to describe the statistical analysis of findings from a large number of independent studies. But the roots of meta-analysis go back much farther. Reviewers have been using numbers to give readers a sense of review findings since the early years of this century. Sometimes crude, sometimes sophisticated, these early quantitative reviews paved the way for the development of a variety of meta-analytic methods during the past decade.

The purpose of this chapter is to describe both the past and the recent history of meta-analysis. We begin by defining meta-analysis and listing its major characteristics. We then describe developments in statistics and in research reviewing that influenced the formulation of modern meta-analytic methods. Finally, we describe and assess major meta-analytic methods in use today.

What is Meta-Analyis?

Glass (1976, p. 3) described meta-analysis in three words as the "analysis of analyses"--certainly the most succinct definition that has yet been proposed for this methodology. He went on to define meta-analysis more formally as the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings. According to Glass, the meta-analyst (a) uses objective methods to find studies for a review; (b) describes the features of the studies in quantitative or quasi-quantitative terms; (c) expresses treatment effects of all studies on a common scale of effect size; and (d) uses statistical techniques to relate study features to study outcomes.

Several aspects of Glass's characterization of meta-analysis are especially worthy of note.

1. A meta-analysis covers review results. It encompasses results found in objective searches of a research literature. Glass did not use the term to describe analysis of a planned series of investigations.

2. A meta-analysis is an application of statistical tools to summary statistics, not raw data. The meta-analyst's observations are means, standard deviations, and results from statistical tests. An analysis of raw scores is a primary analysis or secondary analysis; it is not a meta-analysis.

3. A meta-analysis covers a large number of studies. Glass's meta-analysis on effectiveness of psychotherapy covered 475 studies (Smith, Glass, & Miller, 1980). His meta-analysis on class size covered 77 reports (Glass, Cahen, & Filby, 1982). Reviews that cover only a handful of studies may be mini-analyses; they are not meta-analyses.

4. A meta-analysis focuses on size of treatment effects, not just statistical significance. Reviews that do not base their conclusions on effect sizes and relationship strengths differ in a critical way from Glass's meta-analytic reviews.

5. A meta-analysis focuses on relations between study features and outcomes. The meta-analyst's goal is not simply to summarize a whole body of literature with a single average effect size or overall significance level. A meta-analyst also tries to determine how study features influence effect sizes.

Not all users of meta-analytic methods would accept this characterization of the area. Rosenthal (1984), for example, uses the term *meta-analysis* in a much broader sense than Glass does. For Rosenthal, meta-analysis is the use of statistical techniques either to combine or compare either effect size measures or probability levels from either two studies or more than two studies. For Rosenthal, therefore, an experimenter who combines probability levels from two of her own experiments is carrying out a meta-analysis. Hedges (1986) also uses the term meta-analysis in a broad sense. For him, meta-analyses are reviews that make explicit use of quantitative methods to express the results of studies or to combine those results across studies. These newer definitions of meta-analysis have not yet caught on, however, and Glass's characterization of the area seems most consistent with common usage.

The term *meta-analysis* has been criticized as a poor name for quantitative reviewing. One objection to the term is that it is grander than it need be. To some researchers it suggests analysis not only at a different level from primary or secondary analysis but also analysis at a higher level. Researchers who carry out primary and secondary analyses naturally feel somewhat offended by this connotation of the term. Another problem with the term *meta-analysis* is that it suggests taking apart rather than putting together. Some reviewers consider *synthesis* to be a better word than *analysis* to describe a review's function. Users of Glass's methodology have suggested a variety of alternative names for his approach--*research integration*, *research synthesis*, and quantitative reviews, among others--but none of these terms has yet come into common usage.

Meta-analytic Antecedents

In the early 1930s statisticians developed elegant statistical tools for combining results from series of planned experiments. At the same time reviewers were applying seat-of-the-pants statistical methods to the often messy accumulations of research results that they found in the literature. The work carried out during this time is still exerting an influence on meta-analysis.

Statistical Developments

Statistical approaches developed during the 1930s for combining results from a series of studies were of two types. One approach required researchers to combine probability levels from the studies. The other required researchers to first determine whether experiments produced homogeneous results and then to make combined estimates of treatment effects.

Combined tests. Most methods for combining probability levels are based on a simple fact (Mosteller & Bush, 1954). If the null hypothesis is true in each study in a set, then p values from statistical tests of all studies will be uniformly distributed between zero and one. That is, the number of outcomes with p values between, say, 0.5 and 0.6 will be the same as the number between 0.1 and 0.2. This property of p values makes it possible to combine them to obtain new probabilities. One transforms probabilities to values that can be added and then transforms the combined value to a new probability.

Fisher (1932) was one of the first to devise a means for transforming and combining p values, and his approach continues to be one of the best known and most often used. Fisher's method requires the researcher to take the natural logarithm of the one-tailed p value of each study in a set and to multiply the value by -2. Each of the resulting quantities is distributed as chi square with 2 degrees of freedom. Since the sum of independent chi squares is also distributed as chi square, an overall test of significance is provided by the sum of these logs.

$$\chi^2 = -2 \sum \log_e p \qquad (1)$$

Stouffer's method (Mosteller & Bush, 1954) is also a popular approach to combined probabilities, and it is even simpler than

Fisher's to use. The method requires the analyst to add standard normal deviates, or Z values, associated with obtained p values and then divide the sum by the square root of the number (n) of studies being combined.

$$Z_C = \frac{\Sigma Z}{\sqrt{n}} \qquad (2)$$

All that one needs to apply Stouffer's method is a table of normal-curve deviates, paper and pencil, and a few minutes of time.

These methods for combining probabilities have much to offer to researchers who are combining results from several of their own investigations. Researchers can use the methods even when they no longer have access to the original data from the experiments. They can apply the tests without doing time-consuming calculations. And they can use them without worrying about restrictive assumptions, such as homogeneity of variances within studies. About the only thing that researchers have to be concerned about when using the tests is the independence of the data sets whose p levels are being combined.

Rosenthal (1984), a leader in the development of a methodology for meta-analysis, believes that combining probabilities can also be a useful methodology for research reviewers. His interest in this methodology goes back at least to 1963 when he used combined tests to show that experimenter bias can significantly influence the results of social science experiments. Rosenthal, however, does not recommend the use of combined probabilities on a stand-alone basis in meta-analytic reviews. He recommends that reviewers supplement combined probabilities with analysis of effect size measures.

Other leaders in meta-analysis, however, do not even see a limited role for combined tests in research reviews. Their reasons for disliking combined probabilities are not hard to understand. First, with hundreds of studies and thousands of subjects encompassed in a meta-analytic review, these methods will almost always produce statistically significant results. They seldom tell a reviewer anything that cannot be guessed. Second, these methods provide no information about effect size. They do not help a reviewer decide whether overall effects are large and important or small and unimportant. And third, combined probability methods provide no information about moderator variables, or study features that may be used to separate sets of studies into subsets that differ in their effects.

Combined treatment effects. Cochran's method of estimating combined treatment effects requires researchers to reconstruct the means, sample sizes, and mean squares within conditions for all studies in a set and then to combine the data into an overall analysis of variance in which studies are regarded as one factor. Like procedures for combining probabilities, Cochran's method of combining treatment effects was developed to deal with results from a planned series of studies (Cochran, 1937, 1943; Cochran & Cox, 1957). He did not develop his methods for use in research reviews.

Cochran considered a variety of situations in which data from related experiments might be combined. He noted that all the experiments might be of the same size and precision, or that the experiments might differ in size and precision. He noted that effects might be more variable in some studies than in others, and that treatments might have different magnitudes in different studies. Cochran discussed a variety of ways of testing for such complications in data from supposedly identical experiments, and he also proposed several ways of overcoming the effects of such complicating factors.

Major contributors to the meta-analytic literature have commented favorably upon Cochran's approach. Hedges and Olkin (1982), for example, have stated that the statistical ideas proposed in Cochran's earliest papers on combining estimates have stood the test of time. In a 1978 paper Rosenthal commented that the only real disadvantage of Cochran's method is that it requires a lot of work to use, especially when the number of studies grows from just a few to dozens, scores, or hundreds.

Nevertheless Cochran's approach to combining study results is seldom used in research reviews. We know of no reviewer, for instance, who has applied Cochran's methodology in a social science review. The major problem is that direct application of Cochran's methods requires all results to be reported in the same unit of measurement. Studies collected by social science reviewers usually contain results on different scales. Transformation of results to a common scale is necessary before methods like Cochran's can be applied.

Beyond that, Cochran worked out his procedures for planned series of experiments, not for independent results located in the literature, and his worked examples do not cover situations reviewers typically encounter. In Cochran's illustrations of his methods, for example, studies are never nested within levels of another factor. In meta-analytic data sets, nesting is the rule. The meta-analyst investigating the relationship between study source and treatment effects, for example, will have one set of studies nested under the category of dissertations and another set of studies nested under the category of journal articles. In addition, because Cochran's focus is on planned replications of a study in specific times and places, he is usually able to consider studies as a fixed factor in his analyses. Studies found in the literature differ from one another in innumerable ways, some of which are known and some unknown, and they must usually be regarded as a random factor in experimental designs.

In addition, Cochran analyzed sets of experiments that varied only slightly in sample size and experimental precision. Cochran did not consider cases in which the magnitude of variation in study size was large. Hedges (1984) has pointed out that studies in a meta-analytic data set may vary in size by a factor of 50:1. Under such circumstances, Hedges argues, conventional analysis of variance is impossible because of its requirement of homogeneity of error variances. Nor did Cochran consider cases in which some studies use simple two-group, post-test only designs and other studies use complex designs involving covariates and blocking.

Overall, therefore, although Cochran's goal of estimating overall treatment effects was similar to the goal of today's meta-analysts, Cochran dealt with experiments very different from those that meta-analysts typically encounter. His procedures therefore are not directly applicable in meta-analytic reviews. Cochran's work would have to be extended and perhaps revised before it could serve as a guidebook for meta-analysts.

Early Quantitative Reviews

At the same time as statisticians were working out ways for handling results from sets of studies, reviewers were independently developing ways to quantify review results. Some reviewers developed simple approaches involving little more than counting positive and negative results and reporting whether resulting box-scores were too lopsided to be attributed to chance factors. Other reviewers developed methods that were considerably more sophisticated.

Counting positive and negative results. Counting negative and positive results in an area can be done in a number of ways. Reviewers can consider results with p values below .50 to be positive and results with p values above .50 to be negative. Or reviewers can count the number of statistically significant results supporting or contradicting a hypothesis. Or they can form several categories of results: significant positive, mixed, and significant negative.

Social scientists have been using this approach in reviews since early in this century, and boxscores can be found in some of the best known reviews in education and psychology. The method was used, for example, by Paul Meehl (1954) in his influential book Statistical vs. Clinical Prediction. At the core of the book is Meehl's review of 20 studies that pitted the predictions of clinical psychologists against those of simple actuarial tables. Meehl reported that in half the studies, actuarial predictions were reliably superior to those of clinicians, and in all but one of the remaining studies, there was no difference in accuracy of the clinical and actuarial predictions. Costly, labor-intensive clinical predictions came out on top in only 1 of 20 studies. The boxscore was so lopsided that Meehl needed no statistical test to get the message across: Clinical predictions produced a very small yield for their cost.

Chu and Schramm (1968) used counts in a different way in their research review on learning from television, but their review also turned out to be influential. These authors located a total of 207 studies of effectiveness of instructional television. Learning in conventional classrooms was compared with learning from instructional television in each of the studies. Students learned more from instructional television in 15% of the cases; they learned less in 12%; and there was no difference in amount learned in 73% of the cases. Chu and Schramm also noted that the effects of instructional television varied with educational level. The vote for instructional television was better at lower educational levels, poorer at higher levels.

Most meta-analytic methodologists today look upon these counting methods with disfavor. Rosenthal (1978) rightly points out that these methods are usually low in power. A chi-square test of number of positive vs. negative results, for example, will often fail to detect a significant effect even when the effect size in the population is as large as 0.5. Hedges and Olkin (1980) have shown that with low effect sizes, the difficulty of detecting a significant effect may increase as the number of studies increases.

A further problem with vote counts is the meager information they yield. Glass et al. (1981) put the case against boxscores this way:

> A serious deficiency of the voting method of research integration is that it discards good descriptive information. To know that televised instruction beats traditional classroom instruction in 25 of 30 studies--if, in fact, it

> does--is not to know whether television wins by a nose or in a walkaway (p. 95).

Finally, reviewers using counting methods will usually find it very difficult to determine whether subgroups of studies differ in their effects.

Percentages as outcome variables. Long before the development of meta-analysis, some reviewers found themselves in situations where they could provide more than just a count of negative and positive findings. When results from all studies on a topic were reported in percentage terms, reviewers could use more powerful, parametric statistical techniques to summarize findings. They could record a percent score for each study and then treat the set of percent scores as a data set for further analysis.

Eysenck's (1952) well-known review on the effects of psychotherapy used this method of research integration. Eysenck found 19 studies on the improvement of neurotic patients after psychotherapy, and compared the consolidated study results to improvement rates for patients treated either custodially or by general practitioners. Eysenck reported these overall results:

> Patients treated by means of psychoanalysis improve to the extent of 44 per cent; patients treated eclectically improve to the extent of 64 per cent; patients treated only custodially or by general practitioners improve to the extent of 72 per cent. There thus appears to be an inverse correlation between recovery and psychotherapy; the more psychotherapy, the smaller the recovery rate (p. 322).

These results were widely noted in the professional literature and popular press at the time they appeared, and they have had a far-reaching impact on psychology in the years since. Review results could hardly have been presented more strikingly than they were in Eysenck's review.

Underwood's influential review (1957) on interference and forgetting also covered studies that reported results in percentage terms. The starting point for Underwood's review was his perplexity over the disagreement between classic and modern results in studies of retention. Early studies, like those by Ebbinghaus, often showed very high rates of forgetting; more recent studies showed much lower rates. Underwood noted that in most of the older studies, the individuals who served as subjects had learned other material in earlier stages of the experiment. In more recent experiments, naive subjects were the rule.

Underwood wondered whether a subject's experience in learning lists made the difference in study results.

Underwood was able to locate 14 studies with clear results on retention of lists of words. For each study he calculated the percent correct on the last list, and he also calculated the number of lists previously learned. What Underwood found was remarkable (Figure 1). The amount of forgetting could be predicted with great accuracy from the number of lists previously learned. The rank-order correlation between the two variables was -.91. This quantitative analysis of review results provided a classic demonstration of the power of proactive inhibition in forgetting.

With reviews such as this one we come closer to meta-analysis than we do with statistical work on combining probabilities or treatment effects. Underwood's goal was not simply to combine study results, but to show by using quantitative methods sources of regularity and variance in study results. His focus was on studies found in the literature, not on a planned series of experiments. In his hands each study became a data point and a bit of evidence to support a point of view. One senses in this work the beginning of the meta-analytic attitude: the belief that quantitative tools can be used to make sense of a body of research findings.

Correlations as outcomes. Study outcomes are sometimes reported in correlational terms in psychology and education. The reviewer who is attempting to reach conclusions in such cases has advantages over the reviewer working with average scores on psychological scales. Correlations are in themselves indices of relationship strength, and they are independent of the original units of measurement. Because of such characteristics, studies using correlation coefficients are ideally suited for use in quantitative reviews.

Erlenmeyer-Kimling and Jarvik's review (1963) of genetics and intelligence is a good example of an early review that took full advantage of the characteristics of the correlation coefficient. This review covered 99 correlation coefficients representing degree of similarity in intelligence of related individuals. The 99 coefficients came from 52 studies covering a period of 50 years. Erlenmeyer-Kimling and Jarvik classified these coefficients into ten groups on the basis of genetic and environmental similarity of those involved in the correlational pairings.

They found that the magnitude of the correlation coefficients increased regularly as degree of genetic similarity increased (Figure 2). In addition, Erlenmeyer-Kimling and Jarvik reported that for most relationship categories, the median correlation was very close to the theoretical value predicted on the basis of genetic relationship alone. Environmental similarity also contributed to correlation size, but its influence appeared to be smaller than was the influence of genetic similarity. The demonstration was so compelling that it has continued to challenge researchers, theorists, and educators for more than 25 years.

Reviews such as this one bring us to the threshold of meta-analysis. Erlenmeyer-Kimling and Jarvik's review has so much in common with later quantitative reviews that it can almost be classified as a meta-analysis. The review covers numerous studies found in a diverse literature; it measures effects or relationships in all studies on a common scale; it codes study findings according to some central feature or features; and it finally shows that the features explain a great deal of variation in study results. Most of the ingredients for meta-analysis are present in the review. All that it really lacks is a name for its methodology.

Meta-analytic Approaches

The presidential address of the American Educational Research Association provides an ideal platform for reconceptualization of issues in educational research, and in 1976 AERA president Gene Glass took full advantage of the opportunity the platform provided. His presidential address gave quantitative reviews a name and an identity. The speech changed--perhaps for all time--our conception of what social science reviews can be.

Glass's Meta-Analytic Methodology

Glass (1976) distinguished between three types of research endeavors: primary studies, secondary studies, and meta-analyses. Primary researchers carry out basic experimental and field studies; secondary analysts reanalyze data from such studies; and meta-analysts organize the results from many primary studies in order to draw general conclusions. Glass stressed the importance of meta-analytic activity. Without it, researchers could be overwhelmed by the quantity of unorganized findings in any area of science.

> Five hundred studies on class size or ability grouping can accumulate: they will defy simple summary. Their meaning can no more be grasped in our traditional narrative,

> discursive review than one can grasp the sense of 500 test scores without the aid of techniques for organizing, depicting, and interrelating data (1976, p. 4).

What made these points particularly compelling was Glass's description of his own meta-analytic work. The meta-analysis that Glass described most fully in his AERA address covered results from nearly 500 controlled evaluations of the effects of psychotherapy (Smith et al. 1980). To carry out the analysis, Glass and his colleagues first expressed results of each evaluation as a standardized mean difference in scores of experimental and control groups, and they then coded each study for its major features. From extensive multivariate analysis, they concluded that psychotherapy is effective, raising the typical client from the 50th to the 75th percentile of the untreated population. They also concluded that different types of therapy (e.g., behavioral and nonbehavioral) differed little in their overall effectiveness.

Glass's other major meta-analytic synthesis of research is equally impressive (Glass et al., 1982). It focuses on the relationship between class size and student learning. The research literature in this area was too variable to be covered by the methods that Glass used in his synthesis of psychotherapy research. Glass was able to assume a fairly uniform definition of experimental and control treatments in his meta-analysis of psychotherapy findings; he could not make a comparable assumption in his meta-analysis of class size findings. Classes varied too much in size from study to study; one study's small class could be another study's large class. Another complication was the possibility of a nonlinear relation between class size and student learning. Glass and his colleagues suspected that the effect of adding 20 students to a class of 20 would be different from the effect of adding 20 students to a class of 200. Glass devised ways of handling these complications and concluded that the relationship between class size and student learning is best described as a logarithmic relationship.

Among the many contributions that Glass made to quantitative reviewing, four seem to us to be especially important. First, Glass demonstrated that the standardized mean difference could be used as a convenient unit-free measure of effect size in reviews covering experimental research. Glass's use of this measure greatly extended the number of research topics that could be covered in quantitative reviews. Cohen and others had already demonstrated that the standardized mean difference provided a useful index of effect size in experimental work, but Glass was among the first to appreciate the contribution that this index could make to research reviews.

Second, Glass demonstrated that the number of studies available on important social science questions was much larger than many reviewers imagined it to be. Eysenck's (1952) landmark quantitative review on psychotherapy, for example, had covered only 19 studies. Glass's meta-analysis covered 475. Glass pointed out that large bodies of studies were available on other important questions: class size, computer-based instruction, learning from television, etc. Meta-analyses have appeared in each of these literatures in the years since Glass gave his address.

A third contribution was the demonstration that the influence of dozens of study features might be explored in reviews. Earlier quantitative reviewers categorized study results by one or two features. Underwood (1957), for example, classified studies by number of lists learned by subjects before the last list. Erlenmeyer-Kimling and Jarvik (1963) classified study findings by genetic and environmental similarity of paired individuals. Glass and his colleagues classified studies on more than 20 variables. The variables covered not only features of the treatment but also methodological features of studies, setting features, and characteristics of publications in which they were found.

Finally, the analytic methods that Glass used went far beyond the methods previously used in quantitative reviews. Glass, for example, developed regression equations relating size of treatment effect--the dependent variable--to such factors as therapy type, type of client, nature of outcome measure, etc. The equations gave Glass a way of determining how effective behavioral and verbal therapies would be if both were evaluated in studies of the same type. Nothing remotely like this had ever been done before in research reviews.

The importance of Glass's development of meta-analysis was widely recognized at the time of his address to the American Educational Research Association. Within a few years of the address, hundreds of meta-analyses were being carried out on the literatures of the social and health sciences (Kulik, 1984). If imitation is the surest index of admiration, Glass's admirers were legion. But Glass's methodology also had its critics. Soon after the publication of reports on Glass's first meta-analysis, criticisms of the methodology appeared in print (Eysenck, 1978; Mansfield & Busse, 1977; Presby, 1978). And the publication of Glass's work on class size stimulated a new wave of criticism (Educational Research Service, 1980; Slavin, 1984).

The major criticisms of Glass's meta-analyses are four (Glass et al., 1981, ch. 7). First, Glass's meta-analyses are said to give too much attention to low-quality studies. Second, Glass's meta-analyses have been criticized for being too dependent on published results, which may differ from results that do not get into print. Third, Glass's meta-analyses are said to mix apples and oranges. And fourth, they have been criticized for covering multiple results derived from the same studies. With multiple representation of a study in a data set, samples sizes may be inflated, thus creating a misleading impression of reliability of results.

The first two of these criticisms seem to us to fall wide of the mark. Glass's reviews have done as much as anyone's to focus attention on the influence that study quality and publication bias have on study results. Glass has taken great pains to include in his reviews studies from a variety of sources and studies with a variety of methodological features. His meta-analyses have produced challenging evidence on the relationship between strength of social science findings and both study quality (Glass et al., 1981, ch. 7) and publication bias (Glass et al., 1981, ch. 3). To criticize Glass for paying too little attention to study quality and publication bias is to miss the point of Glass's meta-analytic activities.

The third criticism of Glass's meta-analyses deserves closer examination. This is the criticism that Glass's meta-analyses mix apples and oranges. It should be pointed out, first of all, that all nontrivial reviews cover a variety of studies, and so in a sense all reviews, quantitative as well as literary ones, mix apples and oranges. In covering studies of different types of therapy in a single review, therefore, Glass did just what other good reviewers do. In reviewing studies of class sizes in different types of schools, Glass also did nothing novel. To produce meaningful conclusions, reviews have to have adequate scope. They cannot limit their focus to studies that exactly replicate one another.

But having said this, we must add that Glass may have gone farther than other reviewers in mixing results. We must recognize that the standardized mean difference is a statistical index that gives a reviewer extraordinary freedom to combine disparate studies. The meta-analyst can transform outcomes from entirely different experiments using entirely different measures into standardized mean differences and then easily overlook the fact that the two measures cover different things. Literary reviewers must think long and hard before deciding to describe in a single paragraph studies with different outcome measures; meta-analysts can put such studies into a single analysis with the greatest of ease. Some critics believe that this is exactly what Glass and his colleagues did in their meta-analyses. Freed of some of the constraints that ordinary reviewers feel, they may have mixed incompatibles.

In their study of psychotherapy, for example, Glass and his colleagues (Smith et al., 1980) mixed results not only from different types of therapy but also from different types of outcome measures. They calculated effects of psychotherapy on such different measures as palmar sweat, inkblot scores, therapist ratings of adjustment, grade-point averages, and self-ratings of improvement. No matter what the original unit of measurement, Glass and his colleagues expressed the difference between treated and control subjects in standard deviation units. They analyzed the collection of all indices of effect size in the same regression analysis and reached the following overall conclusion: "The average study showed a 0.68 standard deviation superiority of the treated group over the control group" (Smith & Glass, 1977, p. 754). The reader might well ask: A superiority of 0.68 standard deviations of what? Of palmar sweat? Self-satisfaction? Academic achievement? Job performance? The answer is that the superiority is in some unspecified combination of these measures. Whether the answer is satisfactory for researchers and practitioners remains to be seen.

The fourth criticism--that Glass's meta-analyses lump together nonindependent results--also seems to us to have some validity. Glass and his colleagues often code several effect sizes from a single study and routinely include all the effect sizes in a single regression analysis. Glass's analysis of psychotherapy effects, for example, covered 475 studies, but some of his analyses were based on nearly 1800 effect sizes. Glass's analysis of class size covered 77 studies but the data analyses covered 725 effect sizes. These numbers indicate an inflated *N*--a sample size much larger than the number of independent observations. When a study is represented two, three, four, or five times in a data set, it is difficult for an analyst to determine the amount of error in statistics describing the set, and it is virtually impossible for the analyst to estimate the actual degree of correlation among study features. The results from regression analyses on such data sets should be treated with some caution.

To keep things in perspective, however, we must say that these are small quibbles considering the overall importance of Glass's contributions. Glass not only devised a method for a specific problem but he saw clearly the wider implications in the use of his method. He worked through innumerable details in thє application of meta-analysis so that his writings continue to be the best source of meta-analytic guidelines. The value of Glass's work is beyond question, and its importance seems likely to continue to increase in the years ahead.

Hedges' Modern Statistical Methods

The statistical methods that Glass used in his meta-analyses were conventional ones, such as analysis of variance and regression analysis, but Glass applied these techniques to a novel type of data set. Instead of using these methods with raw observations, Glass applied them to summary study statistics. Hedges (1984) has recently commented on Glass's use of conventional statistics in research synthesis:

> Such use seemed at first to be an innocuous extension of statistical methods to a new situation. However, recent research has demonstrated that the use of such statistical procedures as analysis of variance and regression analysis cannot be justified for meta-analysis. Fortunately, some new statistical procedures have been designed specifically for meta-analysis (p. 25).

Hedges (1984) is one of the major architects of what he has called "modern statistical methods for meta-analysis" (p. 25).

One of Hedges' first contributions to meta-analysis was his demonstration that the effect size statistics usually calculated for meta-analyses were biased estimators of an underlying population effect (Hedges, 1982a). Hedges proposed a correction for Cohen's effect size estimator d that removed this bias:

$$d^u = \left(1 - \frac{3}{4 (n_e + n_c - 2) - 1} \right) d \qquad (3)$$

where d^u is the unbiased estimator and n_e and n_c are the sample sizes for the experimental and control groups.

Other meta-analysts soon reported that use of this correction had at most a trivial effect on their results. Bangert-Drowns, Kulik, and Kulik (1983), for example, calculated 27 effect sizes with and without Hedges' correction. They reported that uncorrected and corrected effect sizes correlated .999, and in most cases agreed to two decimal places. In view of the small difference that the correction makes, many meta-analysts today do not bother to make it.

Hedges (1982a) also showed that his unbiased estimator had a sampling distribution of a noncentral t times a constant. Furthermore, with large sample sizes, the distribution of Hedges's unbiased estimator is approximately normal with standard deviation

$$s^2(d) = \left(\frac{1}{n_e} + \frac{1}{n_c} \right) + \frac{d^2}{2 \left(n_e + n_c \right)} \qquad (2)$$

In his earlier writings, Hedges implied that this formula was the only one needed to calculate the sampling error of an effect size.

> The variance of d is completely determined by the sample sizes and the value of d. Consequently, it is possible to determine the sampling variance of d from a single observation. The ability to determine the nonsystematic variance of d (the variance of ϵ) from a single observation of d is the key to modern statistical methods for meta-analysis. This relationship allows the meta-analyst to use all the degrees of freedom among different d values for estimating systematic effects while still providing a way of estimating the unsystematic variance needed to construct statistical tests (Hedges, 1984, p. 33)

We have criticized Hedges before for this description of factors determining sampling error of effect size estimates (C. Kulik & J. Kulik, 1985; J. Kulik & C. Kulik, 1986). We pointed out that standard errors of effect sizes are not only a function of sample size and population effects but they are also influenced by experimental design. With a given population effect and sample size, for example, the error in measuring a treatment effect can be large or small, depending on whether covariates were used in the experimental design to increase the precision of measurement of the treatment effect. For example, when an effect d is measured with an analysis of covariance design, its variance is given by

$$s^2(d) = \left(1 - r^2 \right)\left(\frac{1}{n_e} + \frac{1}{n_c} \right) + \frac{d^2}{2 \left(n_e + n_c \right)} \qquad (3)$$

where r is the correlation between the dependent variable and the covariate.

Hedges has acknowledged this point in his recent writings on meta-analytic methodology (Hedges, 1986). He mentions that the formulas that he has presented as modern statistics for meta-analysis apply only to what can be called "operative" effect sizes, and these effect sizes are not usually appropriate for use in meta-analysis. Hedges has also conceded that adjustments of the sort we described must be used to make his formulas suitable for use in meta-analytic work. He has not yet given detailed guidance, however, on incorporating these adjustments. It is safe to say that reviewers should not attempt to use Hedges' methodology, however, without consulting his 1986 statement.

Hedges (1983) next recommended use of the standard error of the effect size in tests of homogeneity of experimental results. To test the influence of study features on effect sizes, for example, Hedges suggested using homogeneity tests. He recommended first testing the homogeneity of a set of effect sizes, $d_1,\ldots\ldots,d_k$, from k experiments by calculating the statistic

$$H = \sum w_i \ (d_i - d_.)^2 \qquad (4)$$

where $w_i = 1/s^2(d)$. If all k studies share a common effect size, then the statistic H has approximately a chi square distribution with $(k - 1)$ degrees of freedom. The test simply indicates whether the variation among observed effects is greater than one would predict from the reliability of measurement of the individual effect size statistics.

When homogeneity of effects cannot be assumed, Hedges uses an analogue to the analysis of variance to determine whether effects are a function of specific study features. He first divides the studies on the basis of a selected feature into two or more groups. He then determines whether between-group variance in means is greater than would be expected from within-group variation in scores. The between-group homogeneity statistic (H_B) is calculated as follows:

$$H_B = \sum w_{j.} \ (d_{j.} - d_{..})^2 \qquad (7)$$

where $d_{..}$ is the overall weighted mean across all studies ignoring groupings; $d_{j.}$ is the weighted mean of effect size estimates in the jth group; and $w_{j.}$ is the geometric mean of within-cell variances for the jth group. Hedges points out that when there are p groups and the groups share a common population effect size, the statistic H_B has approximately a chi square distribution with $(p - 1)$ degrees of freedom.

Hedges (1984) has noted that this analogue and conventional analysis of variance produce very different results for the same data sets. One set of data that he has used for this demonstration is presented in Table 1. The data come from six studies of the effects of open education on student cooperativeness. Hedges judged three of the studies to be high in treatment fidelity and three to be low. Hedges wanted to determine whether treatment fidelity significantly influenced study results.

He first used conventional analysis of variance to test for the effect of treatment fidelity (Table 2). The test did not lead to rejection of the null hypothesis, $F(1,4) = 4.12$, $p > .10$. Hedges' H_B test, however yielded a chi square of 7.32, $p < .05$. On the basis of this test, Hedges concluded that treatment fidelity has a significant effect on study results. It is interesting to note that Formula 7 can be applied without weighting study statistics by study size. The homogeneity statistic for unweighted means equals 7.75, $p < .05$.

To see why conventional analysis of variance and Hedges' homogeneity test produce different results, we must look more closely at the actual data. The data layout in Table 2 is simply an expansion of the data in Table 1. The means for the experimental and control groups in Table 3 were derived in the following way. For each study

$$\overline{ES} = \frac{\bar{x}_e - \bar{x}_c}{s_x}$$

$$= \frac{\bar{x}_e - \bar{x}_c}{s_x} - \frac{\bar{x}_c - \bar{x}_c}{s_x}$$

$$= \bar{z}_e - 0$$

$$= \bar{z}_e$$

The pooled variance for each study is equal to 1 because the within-study pooled standard deviation for each study was used in the standardization of scores. The sample variances for experimental and control groups should be approximately equal to this pooled variance.

From Table 4 we can see that the results described by Hedges may be regarded as coming from a three-factor experiment, the factors being fidelity categories (A), studies (B), and treatments (C). Studies are nested within fidelity categories but crossed with treatment groups. The linear model for this design (Winer, 1971, p. 362) is

$$z_{ijkn} = \gamma_k + \alpha\gamma_{ik} + \beta\gamma_{j(i)k} + \epsilon_{ijkn} \qquad (8)$$

Two things should be noted. First, the model does not include terms for main effects of categories and studies. These terms do not appear because the standardization of scores within studies makes it impossible for study effects to exist independently of interaction effects. Second, studies must be considered a random, sampled factor, not a fixed factor, in situations like this one (Cronbach, 1980; Hedges, 1983). That is, we are interested in knowing whether treatment fidelity generally influences effects in studies like these. We do not want to limit our generalizations to a specific set of six studies that differ from one another in innumerable known and unknown ways. The population of settings in which open education might be used encompasses much more than is covered by these six settings.

Table 4 presents results from an unweighted means analysis of variance of Hedges' data. The unweighted means analysis was used because study sizes are unlikely to reflect factors relevant to the experimental variables, and there is no compelling reason for having the frequencies influence the estimation of the population means. The test for effect of fidelity category on effect size produces $F(1,4) = 4.12$, $p > .10$). It should be noted that this F is identical to the F reported by Hedges for a conventional analysis of variance, in which study means are used as the dependent variable. This result should not come as a surprise. Data from nested designs such as this one can often be tested with a simpler analysis of variance using study means as the experimental unit (Hopkins, 1982).

It is also noteworthy that an inappropriate test of the effect of fidelity category would use the within-cells mean square as the denominator in the F ratio. Such a test produces an F ratio of 7.75, identical to the result of Hedges' homogeneity test with unweighted means. The similarity of this incorrect result to results of the homogeneity test should alert us to the possibility that the homogeneity test may be based on inappropriate variance estimators.

Hedges has argued that the conventional analysis of variance results are wrong and should not be trusted because meta-analytic data sets cannot meet the analysis of variance requirement of homogeneity of error variance. With different cell sizes, Hedges argues, error variances cannot be assumed to be equal. Our reconstruction of Hedges' data shows that heterogeneity of within-cell variances is not a problem. Because scores are standardized within studies, all within-cell variances are approximately equal to 1. There also seems to be little reason to reject the assumption of homogeneity of variance of study means within fidelity categories. Although sampling errors certainly are different for the study means, sampling is only one factor that contributes to error in measuring study effects.

The problem to us seems not to be in the analysis of variance approach to these data but in Hedges' homogeneity approach. In Hedges' homogeneity formula, each term of the form $(d_{j.} - d_{..})^2$ is actually an estimate of the variance between groups of studies. Each weight $w_{j.} = 1/s^2_{j.}$ is the geometric mean of several within-study variances. Therefore each term of the form

$$H_B = \sum w_{j.} (d_{j.} - d_{..})^2$$

is actually a ratio of a between-group variance to variance within studies. The problem is that within-study variance is not the appropriate variance to use to test the significance of a group factor when studies are a random factor nested within groups. In our view Hedges has provided an analogue to the wrong model of analysis of variance for meta-analytic data.

What can we say overall about Hedges' (1984) modern methods for statistical analysis? First, Hedges has been highly critical of the use of conventional statistics in meta-analysis. He has criticized conventional effect size estimators for bias, but the amount of bias in these indicators is so small that few investigators today correct their effect sizes using Hedges' correction. Second, Hedges has devised a formula for calculating standard errors of effect sizes. Although this formula gives an accurate estimate of the standard error of what we have called operative effect sizes, it does not always yield the right standard errors for the interpretable effect sizes used in meta-analysis. Hedges (1986) has recently conceded that corrections are needed before his formulas for effect size and standard errors of effect sizes can be used in meta-analyses. Third, Hedges has criticized the use of conventional analysis of variance in meta-analysis and recommends instead the use of a chi-square analogue to analysis of variance. Such a test seems to us to be inappropriate for use with meta-analytic data sets. We believe therefore that Hedges' suggested modern methodology for meta-analysis needs careful scrutiny.

Hunter and Schmidt's Validity Generalization

Although he developed statistical tools for summarizing results from correlational research, Glass did not use these techniques extensively in his own research. His major meta-analyses covered experimental studies, not correlational ones. He left to others the job of meta-analyzing studies in the psychometric tradition, and Hunter and Schmidt soon took the lead in this endeavor (e.g., Hunter, Schmidt, & Jackson, 1982).

Hunter and Schmidt's first quantitative reviews predated the development of meta-analysis. In a 1973 paper they investigated differential validity of job prediction tests for blacks and whites. They located 19 studies that contained a total of 410 comparisons of validity coefficients for the two groups. They calculated the average of the two validity coefficients in each comparison, and then from these average coefficients and sample sizes, they developed a expected distribution of significant and nonsignificant study results. They found that the pattern of significant and nonsignificant results in the 410 comparisons was consistent with the hypothesis of no racial difference in test validities They concluded therefore that there was one underlying population validity coefficient that applied equally to black and white populations.

Schmidt and Hunter extended this work and had soon formulated a set of general procedures for reviewing validity studies of employment tests. They referred to their methodology as validity generalization. The methodology requires a reviewer of test validities to first form a distribution of observed validity coefficients. Next, the reviewer must determine whether most of the variation in validity coefficients can be attributed to sampling error. Hunter and Schmidt have developed a cumulation formula for sampling error that helps the reviewer make this determination. To complete the job, the reviewer finally determines whether remaining variation in results can be explained by such factors as (a) study differences in reliability of independent and dependent variable measures; (b) study differences in range restriction; (c) study differences in instrument validity; and (d) computation, typographical, and transcription errors.

Hunter and Schmidt soon realized that their work on validity generalization had much in common with Glass's work on meta-analytic methodology. In a 1982 book, in fact, they proposed that the two methods could be combined into one overall approach. They called the combined approach state-of-the-art meta-analysis. Analysts using the method calculate effect sizes for all studies and correct them for any statistical and measurement artifacts that may have influenced them. The analyst then examines variation in the adjusted effect sizes to see if it can be explained, or explained away, by such factors as sampling error. If not, the analyst examines selected study features to see whether these features can explain variation in study results.

Although details of Hunter and Schmidt's methodology have changed with time, the underlying theme of their work has remained constant: Study results that appear to be different on the surface may actually be perfectly consistent. A good deal of variation in study results is attributable to sampling error. Sample sizes are too small for accurate estimation of parameters in most studies. Add to the effects of sampling error the influence of range restriction, criterion unreliability, and so on, and you have ample reason to expect variable results from studies of a phenomenon that produces consistent effects.

Hunter and Schmidt's developed methodology has much in common with Hedges' methodology. It therefore shares some of the weaknesses of Hedges' approach. For example, Hunter and Schmidt point out that Cohen's effect size estimator *d* and the correlation coefficient *r* are related by the following formula when sample sizes are equal:

$$t = \frac{\sqrt{n}}{2} d = \sqrt{n - 2}\, \frac{r}{\sqrt{1 - r^2}} \qquad (8)$$

where $n_e = n_c = n/2$ and n is the total sample size. This formula oversimplifies the relationship between test statistics, effect sizes, and correlation coefficients. It applies to results from simple two-group experiments with no covariates or blocking, but it does not apply to results from more complex designs.

Furthermore, Hunter and Schmidt, like Hedges, provide only one formula for sampling error of effect sizes:

$$s_d^2 = \frac{4}{n} \left(1 + \frac{d^2}{8} \right) \qquad (9)$$

This formula does not give an accurate indicator of the error of effect sizes when more complex designs are used to measure treatment effects. C. Kulik and J. Kulik (1985), and more recently Hedges (1986), have discussed the problem with such standard error formulas.

A unique feature in Hunter and Schmidt's meta-analytic methodology is adjustment of effect size measures for range restriction and criterion unreliability. Although range-restriction and criterion-unreliability adjustments are sometimes easily made with validity coefficients, they are usually troublesome to make with experimental studies. Reports of experimental research seldom provide the data that reviewers need to make the adjustments. Before making these adjustments in reviews, meta-analysts should also consider the degree to which the adjustments increase error in measurement of treatment effects (Hunter et al., 1982, p. 59). Finally, before making the adjustments, meta-analysts should take into account the expectations of readers of research reviews. Most research readers expect to find actual results summarized in reviews, not the results that might be obtained with theoretically perfect measures and theoretically perfect samples. For reasons such as these, most meta-analysts have been reluctant to endorse the use of the adjustments that Hunter and Schmidt espouse. Rosenthal (1984), for example, has written:

> Since correction for attenuation and for range restriction are not routinely employed by social researchers, greater comparability to typical research can be obtained by presenting the uncorrected results (p. 30).

Rosenthal's Meta-analytic Methods

Robert Rosenthal was making important contributions to quantitative reviews before Glass gave the area its current name. Rosenthal's interest in the topic can be traced back at least to the early 1960s when he began comparing and combining results of studies dealing with experimenter expectancies. In 1976, the year in which Glass's first meta-analysis appeared, Rosenthal published a landmark synthesis of findings from 311 studies of interpersonal expectancies. Among its innovations were measurement of size of study effects with d, the standardized mean difference between an experimental and a control group, and the statistical analysis of the relation between study features and d. In a 1984 book Rosenthal described the approach to quantitative research reviewing that he developed over the years, and in a 1985 book he and Mullen presented a set of 14 computer programs in Basic computer language for carrying out these analyses.

Rosenthal (1984) distinguishes between eight different types of methods available for meta-analysis and he has organized these techniques into a three-way classification. Meta-analytic methods may involve (a) combination or comparison, (b) effect sizes or probabilities, and (c) two studies or more than two studies. Rosenthal recommends using different statistical techniques for each cell in this layout. For combining probabilities from two studies, for example, he recommends using tests like Stouffer's. For comparing effect sizes from more than two studies, Rosenthal recommends what he calls focused tests. Rosenthal's focused tests are formally identical to the homogeneity tests advocated by Hedges.

Rosenthal's approach to meta-analysis is above all else eclectic and tolerant. Rosenthal has a good word to say about almost any method that has ever been used to treat statistically results from multiple experiments. Rosentnal puts side by side, for example, the method of counting positive and negative findings and Cochran's method of reconstructing analyses of variance. He shows that they produce very different conclusions when applied to the same set of data. But Rosenthal does not indicate clearly which is to be preferred. He simply mentions mildly that judging significance by counting positive and negative results may lack power and that Cochran's test may be time-consuming with large sets of data. Rosenthal leaves it up to the individual meta-analyst to choose between methods.

But Rosenthal does have some preferences and some of these are idiosyncratic. Rosenthal looks favorably upon the practice of combining probability levels from different studies located by a reviewer; most other meta-analysts do not. He applies meta-analytic methods to as few as two related studies of a topic; most other meta-analysts insist on having more than two studies available before they try to find the pattern in the set of results.

Among the most controversial aspects of Rosenthal's methodology is his retrieval of effect sizes, without apology, from the sample size and the value of a test statistic associated with a study. Other meta-analysts, including ourselves (C.-

L. Kulik & J. Kulik, 1985; J. Kulik & C.-L. Kulik, 1986), have pointed out that effect-size indices such as *d* cannot be calculated from these two factors alone. A meta-analyst needs to know in addition something about the experimental design that produced the test statistic: whether the experimental design used blocking, matching, or any other device to increase the power of the statistical test.

For example, Rosenthal converts *t*- and *F*-statistics to the effect size indicator *d* by using the following equation:

$$F = t^2 = (1/n_e + 1/n_c) d \tag{10}$$

where n_e and n_c are the sample sizes for the experimental and control groups. This formula accurately summarizes the relation between an interpretable effect-size index *d*, *F*, and *t* only when *F* or *t* comes from a posttest-only, two-independent-group experiment without covariates or blocking. When *t* and *F* statistics come from other experimental designs (and they usually do), Rosenthal's formula does not apply. When *F* comes from an a comparison of gain scores in experimental and control groups, for example, the formula relating *F*, *t*, and *d* is:

$$F = t^2 = 2(1 - r) (1/n_e + 1/n_c) d \tag{11}$$

where *r* is the correlation between pre- and post-scores.

A related problem is Rosenthal's estimation of size of treatment effects from sample sizes and the probability levels associated with the treatment effects. These two factors provide an even poorer basis for estimating size of effect than do sample size and test-statistic value. Meta-analysts who know the sample size and the probability level associated with a treatment effect also need to know what kind of statistical test produced the probability level. With a given sample size and a given probability level associated with the treatment, for example, effect sizes can vary widely depending on whether a parametric or nonparametric test was used in a study (Glass et al., 1981, p. 130-131).

Finally, Rosenthal proposes applying contrast weights to studies in what he calls focused statistical tests. Rosenthal uses these focused tests to determine whether certain studies produce stronger effects than others do. Use of contrast weights makes sense with factors with fixed levels; contrast weights are not appropriate for random, sampled factors (Hays, 1973, p. 582), and studies carried out independently by different investigators at different times in different places under a myriad of different circumstances surely represent a sampled factor rather than one with fixed levels.

Conclusions

For more than 50 years now, reviewers and statisticians have been trying to develop ways to integrate findings from independent studies of research questions. For most of those 50 years the methods in use have been simple and unsophisticated. Reviewers counted studies that supported or rejected their hypotheses, o. they combined probability levels of small numbers of studies without adequately testing for the homogeneity of results in the studies. Occasionally reviewers using such methods produced powerful and compelling reviews, but the results of use of quantitative methods in reviews were too unpredictable for the methods to catch on.

The year 1976 proved a watershed year in quantitative reviewing. In that year both Glass and Rosenthal produced quantitative reviews that made use of the standardized mean difference as an index of effect size in individual studies. Since that time developments in meta-analytic methodology have been rapid. Although some of the developments have been positive, other developments are of more questionable value. Among the developments that are most troubling to us are the use of `nflated sample sizes and nonindependent measures in statistical .lyses, the failure to take experimental design into account in .imating effect sizes and sampling errors, and the development of inappropriate statistical methods for testing the influence of study features on study outcomes.

References

Chu, G. C., & Schramm, W. (1968). Learning from television: What the research says. Washington, DC: National Association of Educational Broadcasters.

Cochran, W. G., & Cox, G. M. (1957). Experimental designs (2nd ed.). New York: Wiley.

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. Journal of the Royal Statistical Society, Supplement, 4, 102-118.

Cochran, W. G. (1943). The comparison of different scales of measurement for experimental results. Annals of Mathematical Statistics, 14, 205-216.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. Journal of Abnormal and Social Psychology, 65, 145-153.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences (Revised Edition). New York: Academic Press.

Cronbach, l. J. (1980). Toward reform of program evaluation. San Francisco: Jossey-Bass.

Educational Research Service (1980). Class size research: A critique of recent meta-analyses. Phi Delta Kappan, 62, 239-241.

Erlenmeyer-Kimling, L., & Jarvik, L. F. (1963). Genetics and intelligence: A review. Science, 142, 1477-1479.

Eysenck, H. J. (1952). The effects of psychotherapy: An evaluation. Journal of Consulting Psychology, 16, 319-324.

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

Fisher, R. A. (1932). Statistical methods for research workers (4th Ed.). London: Oliver and Boyd.

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

Glass, G. V, Cahen, L. S., Smith, M. L., & Filby, N. N. (1982). School class size: Research and policy. Beverly Hills, CA: Sage.

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

Hays, W. L. (1973). Statistics for the social sciences, 2nd Edition. New York: Holt, Rinehart, & Winston.

Hedges, L. V. (1982a). Estimation of effect sizes 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. (1963). A random effects model for effect sizes. Psychological Bulletin, 93, 388-396.

Hedges, L. V. (1984). Advances in statistical methods for meta-analysis. In W. H. Yeaton & P. M. Wortman (Eds.), Issues in data synthesis. New Directions for Program Evaluation, no. 24. San Francisco: Jossey-Bass. Pp. 25-42.

Hedges, L. V. (1986). Issues in meta-analysis. In E. Z. Rothkopf (Ed.), Review of research in education, no. 13. Washington, DC: American Educational Research Association. Pp. 353-398.

Hedges, L. V., & Olkin, I. (1980). Vote counting methods in research synthesis. Psychological Bulletin, 88, 359-369.

Hedges, L. V., & Olkin, I. (1982). Analyses, reanalyses, and meta-analysis. Contemporary Education Review, 1, 157-165.

Hopkins, K. D. (1982). The unit of analysis: Group means versus individual observations. American Educational Research Journal, 19, 5-18.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1983). Meta-analysis: Cumulating research findings across studies. Beverly Hils, CA: Sage.

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

Kulik, C.-L. C., & Kulik, J. A. (1985, July). Estimating effect sizes in quantitative research integration. Ann Arbor, MI: The University of Michigan, Center for Research on Learning and Teaching.

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.

Kulik, J. A., & Kulik, C.-L. C. (1986, April). Operative and interpretable effect sizes in meta-analysis. Paper presentation at the annual meeting of the American Educational Research Association. San Francisco. (ERIC Document Reproduction Service No. ED 275 758)

Light, R. J., & Smith, P. V. (1971). Accumulating evidence: Procedure for resolving contradictions among different research studies. Harvard Educational Review, 41, 429-471.

Mansfield, R. S., & Buse, T. V. (1977). Meta-analysis of research: A rejoinder to Glass. Educational Researcher, 6, 3.

Meehl, P. E. (1954). Clinical versus statistical prediction. Minneapolis: University of Minnesota Press.

Mosteller, F. M., & Bush, R. R. (1954). Selected quantitati e techniques. In G. Lindzey (Ed.), Handbook of social psychology: Vol. 1. Theory and method. Cambridge, MA: Addison-Wesley.

Mullen, B., & Rosenthal, R. (1985). BASIC Meta-analysis: Procedures and programs. Hillsdale, NJ: Lawrence Erlbaum Associates.

Presby, S. (1978). Overly broad categories obscure important differences between therapies. American Psychologist, 33, 514-515.

Rosenthal, R. (1984). Meta-analytic procedures for social research. Applied social science research methods series, vol. 6. Beverly Hills: Sage.

Rosenthal, R. (1963). On the social psychology of the psychological experiment: The experimenter's hypothesis as unintended determinant of experimental results. American Scientist, 51, 268-283.

Schmidt, F. L., Gast-Rosenberg, I., & Hunter, J. E. Validity generalization: Results for computer programmers. Journal of Applied Psychology, 65, 1980, 643-661.

Schmidt, F. L., Berner, J. G., & Hunter, J. E. (1973). Racial differences in validity of employment tests: Reality or illusion? Journal of Applied Psychology, 58, 5-9.

Schmidt, F. L., Gast-Rosenberg, I., & Hunter, J. E. (1980). Validity generalization: Results for computer programmers. Journal of Applied Psychology, 58, 5-9.

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

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.

Underwood, B. J. (1957). Interference and forgetting. Psychological Review, 64, 49-60.

Winer, B. J. (1971). Statistical principles in experimental design, 2nd Edition. New York: McGraw Hill.

Table 1

Effect Sizes from Six Studies of the Effects of Open Education on Cooperativeness (After Hedges, 1984, p. 28)

Study	Treatment Fidelity	n_e	n_c	ES	$s^2(ES)$
1	Low	30	30	0.181	0.0669
2	Low	30	30	-0.521	0.0689
3	Low	280	290	-0.131	0.0070
4	High	6	11	0.959	0.2819
5	High	44	40	0.097	0.0478
6	High	37	55	0.425	0.0462

Table 2

Analysis of Variance Model for Hedges' Data Using Studies as Experimental Unit

Model A: $y_{ij} = \mu + a_i + \beta_{j(i)}$

Source	*df*	*E(MS)*	Example		
			df	*MS*	*F*
Fidelity category *(I)*	$I - 1$	$\sigma_{\beta}^{2} + J\sigma_{a}^{2}$	1	0.634	4.12
Study within category (J:I)	$I(J - 1)$	σ_{β}^{2}	4	0.154	

Table 3

Reconstructed Cell Means and Variances for Six Studies of the Effects of Open Education on Cooperativeness

Treatment Fidelity Category	Study	Teaching Method	n	$\bar{z}$	$s^2(z)$
Low	1	Open	30	0.181	~1.0
		Conventional	30	0.000	~1.0
Low	2	Open	30	-0.521	~1.0
		Conventional	30	0.000	~1.0
Low	3	Open	280	-0.131	~1.0
		Conventional	290	0.000	~1.0
High	4	Open	6	0.959	~1.0
		Conventional	11	0.000	~1.0
High	5	Open	44	0.091	~1.0
		Conventional	40	0.000	~1.0
High	6	Open	37	0.425	~1.0
		Conventional	55	0.000	~1.0

Table 4

Analysis of Variance Model for Hedges' Data Using Effects on Individuals as Experimental Unit

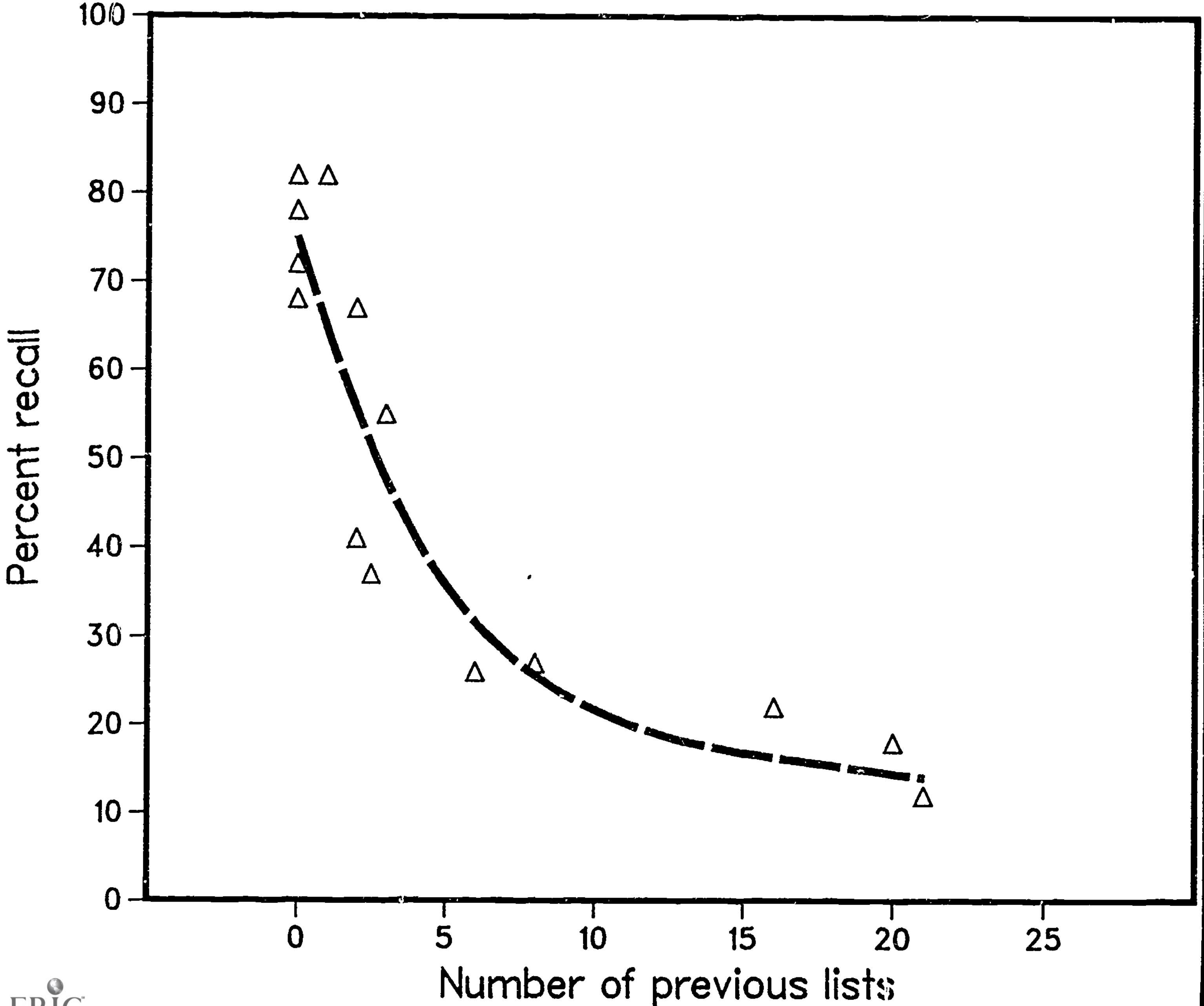
Model B: $z_{ijkn} = \gamma_k + \alpha\gamma_{ik} + \beta\gamma_{j(i)k} + \epsilon_{ijkn}$

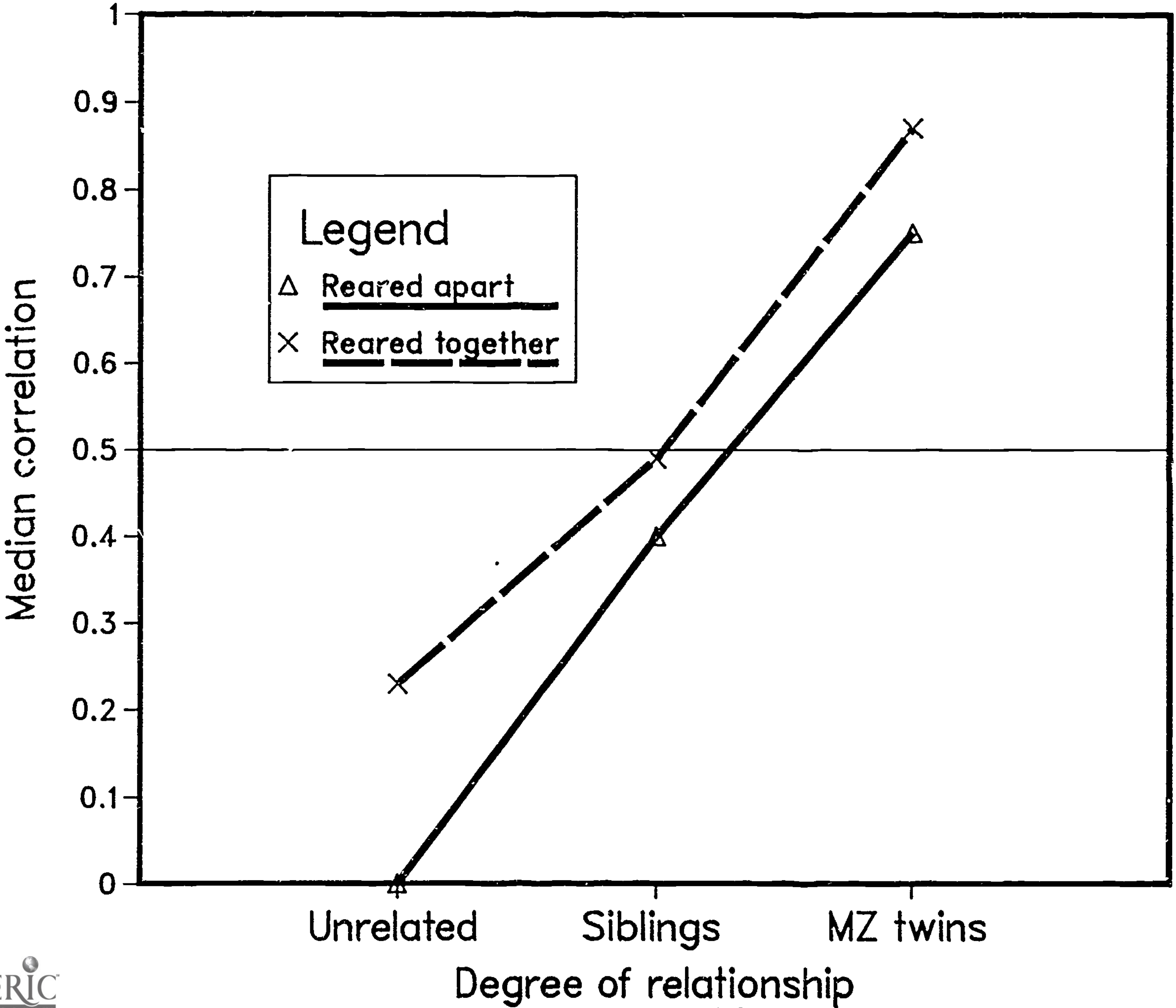
Source	*df*	*E(MS)*	Example		
			df	*MS*	*F*
Method *(K)*	*K - 1*	$\sigma_\epsilon^2 + N\sigma_{\beta\gamma}^2 + JN\sigma_{\alpha\gamma}^2 + IJN\sigma_\gamma^2$	1	2.069	0.677
Fidelity x method *(IK)*	*(I - 1)(K - 1)*	$\sigma_\epsilon^2 + N\sigma_{\beta\gamma}^2 + JN\sigma_{\alpha\gamma}^2$	1	7.75	4.12
Study within category x method *((J:I)K)*	*I(J - 1)(K - 1)*	$\sigma_\epsilon^2 + N\sigma_{\beta\gamma}^2$	4	1.88	1.88
Within cell	*IJK(N - 1)*	σ_ϵ^2	281	1.00	

Figure Captions

Figure 1. Percent recall as a function of previous lists learned based on 14 different studies. (After Underwood, 1957).

Figure 2. Median correlations for individuals with varying relatinships based on 52 studies. (After Erlenmeyer-Kimling and Jarvik, 1963.)


100
90
80
70
60
50
40
30
20
10
0
Percent recall
0
5
10
15
20
25
Number of previous lists


Legend
Δ Reared apart
× Reared together
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Median correlation
Unrelated
Siblings
MZ twins
Degree of relationship