PART I: META-ANALYTIC METHODOLOGY

CHAPTER 1

THE CONCEPT OF META-ANALYSIS

Keeping up with the literature of education becomes a more difficult task each year. The *Current Index to Journals in Education* last year listed more than 17,000 articles published in 700 journals. *Research in Education* indexed an additional 9000 documents, and *Comprehensive Dissertation Abstracts* listed more than 6000 dissertations in education. The number of research studies published next year will undoubtedly be greater, and in the year after next, an even larger number of studies is likely to be added to the literature.

Researchers have long been aware of the need for organizing this vast literature so that it will be more useful to policy makers, administrators, teachers, and other researchers. But the case for research synthesis has seldom been made as convincingly as it was in Gene Glass's 1976 presidential address to the American Educational Research Association.

> Before what has been found can be used, before it can persuade skeptics, influence policy, affect practice, it must be known. Someone must organize it, extract the message. . . We face an abundance of information. Our problem is to find the knowledge in the information. We need methods for the orderly summarization of studies so that knowledge can be extracted from the myriad individual researches (Glass, 1976, p. 4).

Glass pointed out that ordinary research reviews have not done the job. Reviewers usually select studies for review by haphazard processes. They describe study findings in vague and imprecise narrative summaries. They usually report so little about their methods that readers are unable to judge the adequacy of their conclusions.

For real progress in education, Glass pointed out, three types of analyses will be necessary: primary analyses, secondary analyses, and meta-analyses. Primary analysis is the original treatment of data in a research study, usually carried out under the direction of those who designed the study. Secondary analysis is the reanalysis of data for the purpose of answering the original research questions with better statistical techniques. Secondary analysis is carried out by individuals who have access to the original study data, but most often the secondary analyst is someone not involved in the design of the original study. Meta-analysis is the quantitative treatment of review results.

Meta-analysts carry out statistical analyses of quantitative summaries of individual experiments.

Primary analyses usually get the lion's share of attention in educational research. Foundations fund primary studies; journals vie to publish their results; and reputations rise and fall on the basis of their conclusions. Some secondary analyses have been carried out with enough flair to compete with primary analyses for attention. Thus, Elashoff and Snow's *Pygmalion Revisited* is almost as well known as Rosenthal and Jacobson's *Pygmalion in the Classroom.*

In 1976 when Glass spoke about meta-analysis in his AERA presidential address, quantitative research reviews were not a major concern in education. In 1976 no books had been written on meta-analytic methodology. No research review was widely acclaimed as a classic of meta-analytic literature. To Glass, the neglect of meta-analysis was troubling:

> The literature on dozens of topics in education is growing at an astounding rate. In five years time, researchers can produce literally hundreds of studies on IQ and creativity, or impulsive vs. reflective cognitive styles, or any other topic.
>
> In education, the findings are fragile, they vary in confusing irregularity across contexts, classes of subjects, and countless other factors. Where ten studies might suffice to resolve a matter in biology, ten studies on computer assisted instruction or reading may fail to show the same pattern of results twice (p. 3).

Just as statistical analyses were needed to make sense of hundreds of test scores gathered for a primary experiment, so too was statistical analysis needed to make sense of the hundreds of study results available on most research questions in education.

Definition of Meta-Analysis

Glass (1976, p. 3) defined meta-analysis 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. A meta-analysis by Glass and his colleagues on effectiveness of psychotherapy covered 475 studies (Smith, Glass, & Miller, 1980). Their meta-analysis on class size covered 77 reports (Glass, Cahen, Smith, & Filby, 1982; Smith & Glass, 1980). 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.

In recent years some writers have used the term *meta-analysis* in a broader sense than Glass does. Rosenthal (1984), for example, uses the term to describe almost any attempt to combine or compare statistical results from two or more studies. For Rosenthal an experimenter who combines probability levels from two of her own experiments is carrying out a meta-analysis. Such broad 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 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.

Examples of Meta-Analysis

In his 1976 address, Glass cited several examples of meta-analytic work. Included in his examples were a few selected quantitative reviews from the literature, some meta-analyses currently underway in his laboratory, and most important of all, a meta-analysis of findings on psychotherapy conducted by himself and his colleagues. The meta-analysis on psychotherapy was a tour-de-force.

For this analysis, Glass and his colleagues first expressed results of 500 controlled evaluations of psychotherapy as standardized mean differences in

scores of treated and untreated 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 and his colleagues later described the results of their analysis in detail in a book on the evaluation of psychotherapy (Smith *et al.*, 1980).

The other major meta-analytic synthesis of research by Glass and his colleagues was equally impressive (Glass *et al.*, 1982; Smith & Glass, 1980). It focused on the relationship between class size and student learning. Glass and Smith felt that the research literature in this area was too variable to be covered by the methods that they used in their synthesis of psychotherapy research. In that meta-analysis Glass and Smith were able to assume a fairly uniform definition of experimental and control treatments; they could not make a comparable assumption in their 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 Smith 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. They devised ways of handling these complications and finally concluded that the relationship between class size and student learning could be accurately described as a logarithmic relationship.

Glass and Smith's Contributions

Nothing like these reviews had ever been seen before in education. Their scope was almost unparalleled. Seldom in the history of education had reviewers assembled 400 studies on a single research topic, and no earlier reviewers had ever shown so much ingenuity in reducing the results of so many studies to common terms. Seldom had reviewers examined so carefully factors that might affect study results, and probably no one before had ever shown such statistical flair in examining relationships between study features and outcomes.

It now seems clear that Glass and Smith's meta-analytic reviews broke new ground in several directions. Besides contributing to our substantive knowledge, they contributed to the methodology of research reviewing. After 1976 research reviews would never again be the same as they had been before. Four fundamental contributions changed the way researchers thought about research reviews.

Glass and Smith's reviews demonstrated, first of all, that standardized mean differences between experimental and control groups could be used as convenient unit-free measures of effect size in reviews covering experimental research. Glass and Smith's use of such effect sizes greatly extended the

number of research topics that could be covered in quantitative reviews. Cohen (1977) and others had already demonstrated that an index of effect size could be a useful tool in designing experimental studies, but Glass and Smith were among the first to appreciate the contribution that this index could make to research reviews.

Second, Glass and Smith's reviews demonstrated that the number of studies available on important social science questions was much larger than many reviewers had imagined it to be. Eysenck's (1952) landmark quantitative review on psychotherapy, for example, had covered only 19 studies. Glass and Smith's meta-analysis on psychotherapy findings covered 475. Glass and Smith's meta-analysis on class size covered 77 studies and nearly 725 separate comparisons.

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. In their psychotherapy analysis, 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 and Smith used went far beyond the methods previously used in quantitative reviews. They developed regression equations, for example, to relate size of treatment effect—the dependent variable—to such factors as therapy type, type of client, nature of outcome measure, etc. The equations gave them 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.

Glass saw clearly in 1976 that his synthesis methodology had potential not only for illuminating findings on psychotherapy but for clarifying findings in other areas as well. In his 1976 address Glass discussed briefly the application of his methodology to such topics as socioeconomic status and achievement, and he also pointed out other areas of educational research where large literatures existed waiting to be meta-analyzed. Glass gave as examples the literatures on reading research, class size, programmed instruction, instructional television, school integration, computer-assisted instruction, and modern math curricula. He believed that meta-analytic methods could be profitably applied to each area.

The years following Glass's address have proved him to be a good predictor of future educational research developments. Within a few years he and his colleagues had published meta-analyses on class size, programmed instruction, and computer-assisted instruction. Other meta-analyses were soon published on such topics as reading research, school integration, and modern math curricula. Within five years of Glass's address, a bibliography appeared with more than 250 entries on meta-analysis (Lamb & Whitla, 1981). The meta-analytic results came from education, psychology, other social sciences, and the

health sciences. If imitation is the surest index of admiration, then Glass's admirers were legion.

Criticisms

But meta-analytic methodology also attracted critics. The first criticisms of the method appeared in print soon after Glass and his colleagues reported their results on the effectiveness of psychotherapy (Eysenck, 1978; Mansfield & Busse, 1977; Presby, 1978). The publication of Glass and Smith's work on class size stimulated a new wave of criticism (Educational Research Service, 1980; Slavin, 1984).

Glass and his colleagues have described four major criticisms of their meta-analytic reviews (Glass *et al.*, 1981, chap. 7). First, their meta-analyses are said to give too much attention to low-quality studies. Second, their meta-analyses have been criticized for being too dependent on published results, which may differ from results that do not get into print. Third, the 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 and Smith'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 and Smith have taken great pains to include in their reviews studies from a variety of sources with a variety of methodological features. Their meta-analyses have produced challenging evidence on the relationship between strength of social science findings and both study quality (Glass *et al.*, 1981, chap. 7) and publication bias (Glass *et al.*, 1981, chap. 3). To criticize Glass for paying too little attention to study quality and publication bias is to miss the point of his meta-analytic activities.

The third criticism of Glass and Smith's meta-analyses deserves closer examination. This is the criticism that meta-analysts 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. 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 and Smith'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. Their analysis of psychotherapy effects, for example, covered 475 studies, but some of their analyses were based on nearly 1800 effect sizes. Glass and Smith'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 sampling units in the analysis. 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 the 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.

Meta-Analysis of Findings in Education

And what have we learned from the meta-analyses of educational findings that have appeared in the literature since Glass's address? Do they lead to important conclusions? Although individual results have been carefully scrutinized, criticized, and defended, the cumulative findings from meta-analyses of educational research have not been examined. How much has been done? Are the findings dependable? Are they important?

Our purpose in this issue is to answer these questions. Our major goals are two. First, we wish to present enough of the background and methodology of meta-analysis so that readers will be in a position to judge meta-analytic contributions to education on their own. Second, we wish to lay out meta-analytic findings in major areas of educational research so that readers will see just what these findings are.

CHAPTER 2

ANTECEDENTS OF META-ANALYSIS

Quantitative reviewing has a long past. Reviewers have been using numbers to give readers a sense of review findings for at least 50 years. Since the early 1930s they have used special statistical tools for combining results from series of planned experiments. And for just as long a time they have been using simple counting and averaging techniques to summarize the haphazard accumulations of research results found in the literature. The work carried out before Glass's formulation of meta-analytic methodology in 1976 is still exerting an influence on research reviews.

The purpose of this chapter is to examine early efforts to deal with what are now recognized as meta-analytic problems. We first describe the developments in statistical methodology that are relevant for research reviewers. We then look at early applications of quantitative methodology to review literatures. The main points that we make in this chapter are that the need for quantitative methods in research reviews has been recognized for a long time and that pre-Glassian attempts to develop a meta-analytic methodology were far from complete.

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:

$$X^2 = -2 \sum \log_e p \qquad [2.1]$$

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:

$$Z_c = \frac{\Sigma z}{\sqrt{n}} \qquad [2.2]$$

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. 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 or hundreds.

Nevertheless, Cochran's approach to combining study results is seldom used in research reviews. We have not heard of any reviewer, for instance, who has applied Cochran's methodology in a review of educational research findings. The major problem seems to be that direct application of Cochran's methods requires all results to be reported in the same unit of measurement. Studies collected by reviewers in education 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. They would have to be extended and perhaps revised before they could be used in meta-analytic work.

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 findings 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 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 Outcomes

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 first found 19 studies on the improvement of neurotic patients after psychotherapy. He then determined the consolidated improvement rate and compared it to improvement rates for patients treated custodially and by general practitioners. Eysenck reported:

> 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 (1957) influential review 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 2.1). 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 the 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. One senses in his 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.

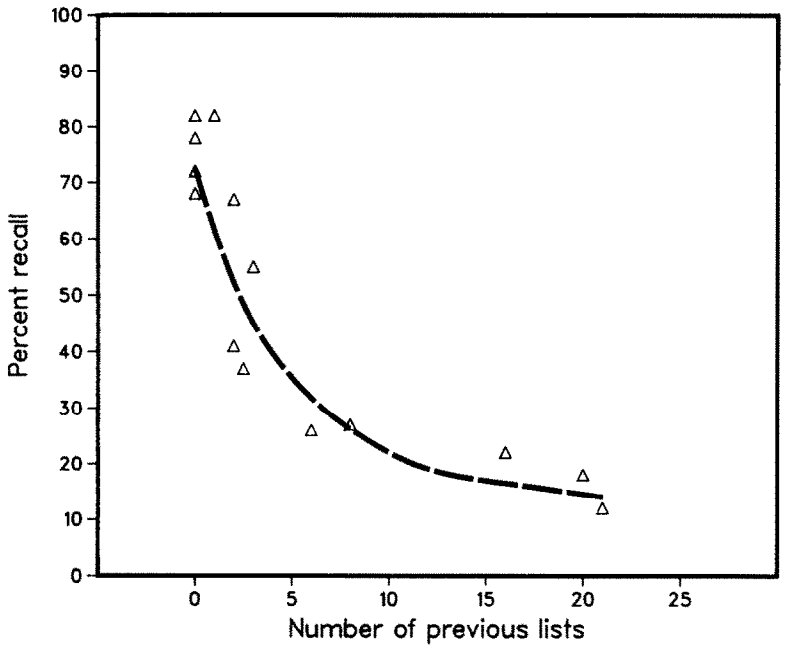

Figure 2.1
Percent Recall as a Function of Previous Lists Learned, Based on 14 Studies Analyzed by Underwood (1957)

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 2.2). In addition,

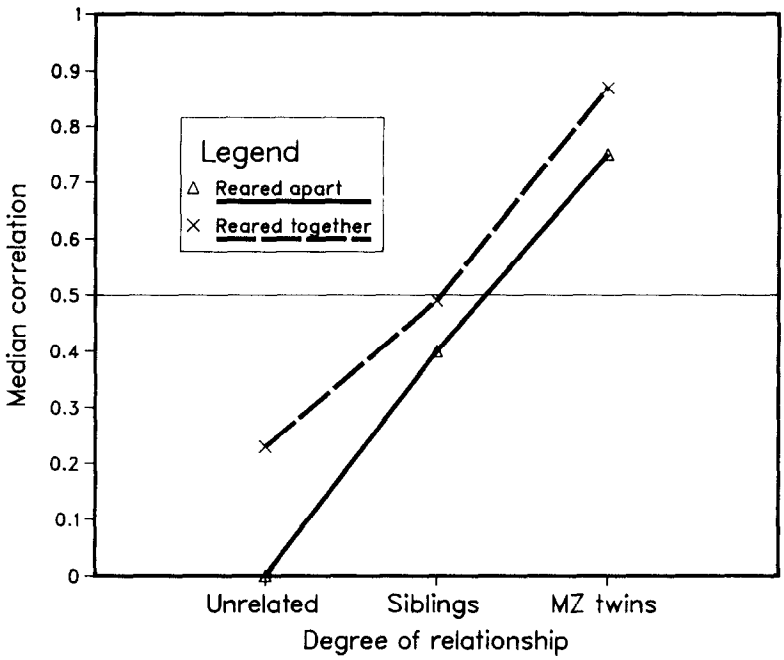

Figure 2.2
Median Correlations for Individuals with Varying Relationships, Based on 52 Studies Analyzed by Erlenmeyer-Kimling and Jarvik (1963)

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 that of genetic similarity. The demonstration was so compelling that it has continued to challenge researchers, theorists, and educators for more than 25 years.

Conclusion

Reviews such as those by Underwood and by Erlenmeyer-Kimling and Jarvik bring us to the threshold of meta-analysis. The reviews have so much in common with later quantitative reviews that they can almost be classified as meta-analyses. They cover numerous studies found in a diverse literature; they report effects or relationships in all studies on a common scale; they arrange study results according to some central study feature or features; and they finally show that the feature explains some of the variation in study results. Most of the ingredients for meta-analysis are present in these reviews. All that they really lack is a special name for their methodology.

CHAPTER 3

RECENT APPROACHES

The presidential address of the American Educational Research Association provides an ideal platform for conceptualization of issues in educational research, and in 1976 AERA president Gene Glass took full advantage of the opportunity that 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.

What became apparent during the decade following Glass's presentation, however, was that his conceptions did not emerge from a vacuum. Other methodologists were working on similar problems, and they began to point out relationships between their work and Glass's. They also began to build bridges between different conceptions of quantitative reviews.

Five methodologists have been especially influential as friendly critics of classic meta-analysis. They are Hedges, who has developed what have been called *modern statistical methods for meta-analysis* (Hedges, 1984; Hedges & Olkin, 1985); Hunter and Schmidt, who have contributed *state-of-the-art meta-analysis* (Hunter, Schmidt, & Jackson, 1982); Rosenthal (1984), who has formulated a taxonomy of meta-analytic methods; and Slavin (1986), who has developed the method of *best-evidence research synthesis*. This chapter presents a description of these four perspectives on meta-analytic methodology.

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 these new statistical procedures for meta-analysis.

One of Hedges' first contributions to meta-analysis was his demonstration that the effect sizes usually calculated by meta-analysts were biased estimators of an underlying population effect (Hedges, 1982a). For this demonstration, Hedges focused on Cohen's (1977) effect-size estimator d, an index that is very similar to Glass's index of effect size. Cohen's d is calculated by subtracting the average score of the untreated group from the average score of the treated group and then divding the remainder by the pooled standard deviation for the two groups. Hedges proposed a correction for d that removed its bias,

$$d^u = \left(1 - \frac{3}{4 (n_e + n_c - 2) - 1} \right) d \ , \qquad [3.1]$$

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. For our meta-analysis with Bangert-Drowns on effects of coaching on test performance (Bangert-Drowns, Kulik, & Kulik, 1983b), for example, we calculated 27 effect sizes with and without Hedges' correction. We found 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 use this correction.

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

$$s_d^2 = \left(\frac{1}{n_e} + \frac{1}{n_c} \right) + \frac{d^2}{2 \left(n_e + n_c \right)} . \qquad [3.2]$$

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)

In earlier papers we have criticized this description of factors determining sampling error of effect size estimates (C. Kulik & Kulik, 1985; J. Kulik & Kulik, 1986). We pointed out that standard errors of effect sizes are a function not only of sample sizes and population effects but also of experimental designs. 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

$$s_d^2 = (1 - r^2)\left(\frac{1}{n_e} + \frac{1}{n_c}\right) + \frac{d^2}{2\left(n_e + n_c\right)}, \qquad [3.3]$$

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*. It should be noted that such operative effect sizes are usually inappropriate 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, $d_1, \ldots d_j$, from j experiments by calculating the statistic

$$H = \sum w_j (d_j - d_{.})^2, \qquad [3.4]$$

where $w_j = 1/s_d^2$. If all J studies share a common effect size, then the statistic H has approximately a chi square distribution with $(J - 1)$ 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:

$$H_B = \sum w_{i.} (d_{i.} - d_{..})^2 \quad , \qquad [3.5]$$

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

Table 3.1
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	M_{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

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 3.1. 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 3.2). 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. When Formula 3.5 is applied without weighting study statistics by study size, the chi-square for testing homogeneity 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 3.3 is simply an expansion of the data in Table 3.1. 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 3.4 we can see that the results described by Hedges may be regarded as coming from a three-factor experiment, the factors being fidelity

Table 3.2
Analysis of Variance Model for Hedges' Data Using Studies as the Experimental Unit

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

Source	*df*	*E(MS)*	*df*	Example *MS*	*F*
Fidelity category *(I)*	$I - 1$	$\sigma^2_\beta + J\sigma^2_\alpha$	1	0.634	4.12
Study within category *(J:I)*	$I(J - 1)$	σ^2_β	4	0.154	

Table 3.3
Reconstructed Cell Means and Variances for Six Studies of the Effects of Open Education on Cooperativeness

Treatment Fidelity Category	Study	Teaching Method	n	M_z	s_z^2
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

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

$$z_{ijkn} = \gamma_k + \alpha\gamma_{ik} + \beta\gamma_{j(i)k} + \epsilon_{ijkn} \ . \qquad [3.6]$$

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 and category effects to exist independently of interaction effects. Second, studies must be considered a random, sampled factor, not a fixed factor, in this situation (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 specific settings.

Table 3.4 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

Table 3.4
Analysis of Variance Model for Hedges' Data Using Effect Sizes on Individuals as the Experimental Unit

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

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

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 should not be trusted because meta-analytic data sets do not meet the analysis of variance requirement of homogeneity of error variance. With different cell sizes, Hedges asserts, error variances cannot be assumed to be equal. Our reconstruction of cell means and variances for Hedges' data set (Table 3.3) 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.

To us the problem 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 $(d_{i\,.} - d_{..})^2$ is actually an estimate of the variance between groups of studies. Each weight $w_{i\,.} = 1/s^2_{i\,.}$ is the geometric mean of several within-study variances. Therefore each term of the form

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

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' 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 correlational studies, 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 (Schmidt, Berner, & Hunter, 1973). 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 an 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 then extended this work and formulated a set of general procedures for reviewing validity studies of employment tests (Schmidt, Gast-Rosenberg, & Hunter, 1980). 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 their 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*. An analyst using the method calculates effect sizes for all studies and corrects 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:

$$t = \frac{\sqrt{n}}{2} d = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}\,, \qquad [3.7]$$

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:

$$s_d^2 = \frac{4}{n}\left(1 + \frac{d^2}{8}\right) \qquad [3.8]$$

This formula does not give an accurate indicator of the error of effect sizes when more complex designs are used to measure treatment effects. We (C. Kulik & 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 (Rosenthal, 1963). In 1976, the year in which Glass and Smith's first meta-analysis appeared, Rosenthal published a landmark synthesis of findings from 311 studies of interpersonal expectancies (Rosenthal, 1976). 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 (Mullen & Rosenthal, 1985).

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 such as 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 most statistical methods that have been used to treat results from multiple experiments. He puts side by side the method of counting positive and negative findings, for example, 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 does not indicate which is to be preferred. He instead points out 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. Kulik & Kulik, 1985; J. Kulik & Kulik, 1986), have pointed out that effect-size indices such as d cannot be calculated from these two factors alone. A meta-analyst also needs to know something about the experimental design that produced the test statistic. Did it involve blocking or matching, for example, 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:

$$F = t^2 = \left(1/n_e + 1/n_c \right) d \ , \qquad [3.9]$$

where n_e and n_c are the sample sizes for the experimental and control groups. This formula accurately summarizes the relation between an effect size d and test-statistics F and t only when F and 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 a comparison of gain scores in experimental and control groups, for example, the formula relating F, t, and d is:

$$F = t^2 = 2\left(1 - r \right) \left(1/n_e + 1/n_c \right) d \ , \qquad [3.10]$$

where r is the correlation between pre- and postscores.

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 focused statistical tests. He 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.

Slavin's Best-Evidence Synthesis

Slavin (1986) has advocated a type of meta-analytic review, called *best-evidence synthesis*, in which a reviewer bases conclusions on statistical treatment and logical analysis of a small number of studies the reviewer considers to be most relevant to a topic and most methodologically sound. Slavin describes the approach as an application of the principle of best evidence in law, which specifies that not all evidence on a case carries the same weight. According to Slavin, best-evidence syntheses combine the best features of traditional reviews and meta-analyses. Like traditional reviews, best-evidence syntheses contain judgments about the credibility of different research results. Like meta-analyses, they express experimental effects in quantitative form.

Although attractive in theory, the approach has several practical limitations. First, best-evidence syntheses usually cover relatively few studies. Slavin's synthesis (1987a) on ability grouping, for example, covered 43 studies, whereas our synthesis on ability grouping covered 109 studies (J. Kulik & Kulik, 1987). Slavin's synthesis (1987b) on mastery learning covered 18 studies, whereas our synthesis on the topic covered 106 studies (J. Kulik, Kulik, & Bangert-Drowns, 1988). Becker's synthesis (1988) on microcomputer-based instruction covered 17 studies, whereas our synthesis (C. Kulik & Kulik, 1988b) on computer-based instruction covered 34 microcomputer-based studies and 220 studies carried out on terminals and mainframes. With reduced pools of studies, Slavin and his colleagues have been able to carry out only the most rudimentary statistical analyses. They have not been able to draw statistically defensible conclusions about the relationships between study features and outcomes, and their reviews have ended up being highly speculative.

Second, analyst biases can play too large a role in the conduct of best-evidence syntheses. We have evaluated in detail two major best-evidence reviews (C. Kulik & Kulik, 1988a; J. Kulik & Kulik, 1987). We concluded that study-inclusion criteria used in the reviews were arbitrary and overly restrictive. Use of such criteria unnecessarily reduced the number of studies available for analysis. We also found that the criteria set up for study inclusion were not uniformly applied. Some studies that did not meet the criteria were included in the best-evidence syntheses; some studies that fit the criteria were excluded. Overall, therefore, we concluded that the best-evidence approach does not provide sufficient safeguards against personal biases of the analyst.

Conclusion

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, or 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 standardized mean differences as an index of effect size. 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 (a) the increasing use of formulas that ignore experimental design factors in the calculation of effect sizes and sampling error, and (b) the advocacy and use of inappropriate statistical methods for testing the influence of study features on study outcomes.

CHAPTER 4

GUIDELINES FOR META-ANALYSIS

Not all meta-analytic evidence is equal. Anyone who has read meta-analytic reports knows that meta-analyses vary in quality. Some can stand up to the most careful scrutiny. Others are so flawed that few conclusions can be drawn from them.

Reports on good and poor meta-analyses are sometimes found side-by-side in the educational literature. Readers who wish to distinguish between the two must pay attention to a variety of factors. How were studies located for the meta-analyses? Are all selected studies relevant to the topic? How were study features coded? How were effect sizes calculated? Were appropriate statistical methods used? Were sample sizes inflated?

The purpose of this chapter is to present guidelines for evaluating the meta-analytic literature. The chapter is meant, first of all, to document the considerations that guided the review of the meta-analytic literature that follows. Second, it is designed to serve as a set of guidelines for review of other meta-analytic literature. Third, it is meant to help readers who wish to conduct meta-analyses themselves. It presents many of the main issues that will confront prospective meta-analysts as they read meta-analytic conclusions, study meta-analytic reports, and begin carrying out meta-analyses of their own.

Finding Studies

Researchers who write conventional narrative reviews of research have sometimes been criticized for the way they go about finding studies. Their search procedures often seem casual and unsystematic. They often locate studies by happenstance rather than design, and their own studies and those of their students usually seem to get a disproportionate amount of attention. Not many studies can be described in a typical conventional review, and so authors of such reviews seldom try to locate all the studies that have been conducted on a research question.

Meta-analysts can make these same mistakes. They too can use unsystematic search procedures, and they too can stop searching for studies before they have located an adequate number. Meta-analyses based on such sloppy search procedures, however, will be seriously flawed. No amount of quantitative ingenuity can make up for poor search and selection procedures.

Defining the Review Area

Meta-analytic reviews often cover broad, loosely defined topics: open learning, computer-based instruction, mastery learning, psychotherapy, etc. One problem with such loosely defined topics is that they suggest different things to different people. Psychotherapy, for example, may be anything from a few sessions with a college counselor to a process of personality restructuring that takes years to complete. Mastery learning may be anything from a slogan to a precise set of rules covering every aspect of student-teacher interaction.

Meta-analysts who choose to analyze the literature on such topics as psychotherapy or mastery learning must first decide therefore about the type of psychotherapy or mastery learning they wish to study. Do they wish to include in their analyses studies of everything ever referred to as psychotherapy or mastery learning? Or do they wish to use more restrictive definitions? The meta-analyst must set up explicit criteria for including and excluding studies.

These criteria should not be too broad. The criteria should show that the meta-analyst has a reasonable respect for use of terms in the literature. If primary researchers usually distinguish between counseling and psychotherapy, then defining psychotherapy so broadly that it includes counseling studies may be a mistake. Conclusions about psychotherapy that are based in part on studies of counseling may be misleading. If most researchers think of controlled studies of tutoring instruction as being different from studies of class size, conclusions based on an amalgam of the two types of studies may also be misleading.

The criteria that the meta-analyst sets up should not, on the other hand, be too restrictive. When inclusion criteria are too restrictive, the meta-analyst may not find enough studies to carry out an adequate analysis. The resulting conclusion may be a weak assertion that "more research is needed." Even if the investigator finds an adequate number of studies, the investigator may be unable to find any interesting or illuminating relationships between study features and outcomes when studies are highly uniform. An investigator who looks at only one type of implementation of a treatment will almost certainly find out nothing interesting about the variations in a treatment that make it more or less effective.

Screening Studies

In addition to defining the variations in treatment that are acceptable for study in a meta-analysis, the analyst must also consider the variations in study

quality that are typical in the literature. Should the analyst include all studies of an issue in a meta-analytic review? Or should the analyst include only the best studies? There are two distinct issues involved. The first is setting up methodological criteria for including the studies. The second is adhering to such criteria.

Glass and his colleagues (Glass *et al.*, 1981) believe that meta-analysts should be tolerant of possible methodological flaws when assembling studies for a meta-analysis. Their meta-analyses therefore cover studies that vary both in quality and source. They cover both true experiments and quasi-experiments. Authors of the studies covered in their meta-analyses include distinguished researchers and students. Glass and his colleagues believe that researchers know too little about the study features that affect study outcomes, and they think that meta-analyses can help us find out about such relationships. But meta-analyses will provide good answers only if meta-analysts examine studies that vary in their features.

Slavin (1986) holds the opposing opinion. He believes that only the best evidence should be used in forming a judgment about effects in an area and that evidence from low-quality studies should be given little or no weight. Slavin compares the research reviewer to a judge in a court of law. The judge would use the best possible evidence in reaching a decision. According to Slavin, the research reviewer should do the same.

The best-evidence approach has proved to be more attractive in theory, however, than in practice. As we pointed out in the preceding chapter, best-evidence reviews conducted so far have been based on very small samples of studies. The number of studies is usually so small, in fact, that best-evidence reviewers are unable to analyze their results statistically. Best-evidence reviewers therefore usually end up substituting speculation for statistical demonstration.

Describing Study Features

The reviewer who has located a pool of studies for a meta-analysis faces a second major task: to describe the various characteristics of each study as objectively as possible. Most meta-analysts use a coding sheet in this task. Use of a coding sheet helps ensure that each study is examined in the same way.

One reason for coding study features is obvious. The coding of study features provides the basis for statistical analysis of relationships between study features and outcomes. In order to examine such relationships, the analyst needs to include in an analysis study features on which there is adequate variation. It is impossible to establish statistically a relationship between a study feature and effect sizes when all or almost all the studies in a set have the same feature. If all or almost all studies of computer-based instruction have been carried out in math classes, for example, it is impossible to show that the effectiveness of computer-based instruction is a function of the

subject taught with computer assistance. Only variables with adequate variation need be examined when the analyst's purpose is to demonstrate a relationship between study features and effect sizes.

But there is a second reason for describing study features precisely and quantitatively in meta-analysis. Meta-analysts want to present an accurate picture of the literature on a research question, and they use descriptive statistics for this purpose. They present a statistical portrait of study features to show which settings have been overstudied, which have been understudied, and which have not been studied at all. Meta-analysts therefore often code studies for features on which there is little or no variation. Such coding information can be useful for researchers who are planning to do further studies in the area.

How should meta-analysts choose the study features to describe? First, they should use the existing meta-analytic literature to see which study features have been coded successfully in the past. The prospective meta-analyst should also use the meta-analytic literature to see how such features were studied. Coding study features that have been used in earlier meta-analyses helps ensure cumulativity in the meta-analytic literature and in educational research.

Meta-analysts should also make use of a second important source when deciding about study features to code. The analyst should read a sample of studies located for the meta-analysis. This preliminary reading of studies is necessary because the appropriate study features for an analysis vary from meta-analytic area to area. The analyst who fails to become intimately acquainted with the literature in an area is likely to miss some of the important features of studies in the area.

Several different kinds of study features can be coded. First, the analyst can code features of the experimental treatment. Features of this sort are ordinarily unique to an area of research. In a study of computer-based instruction, for example, an analyst might code studies for the way in which the computer was used (e.g., drill, tutorial, management, simulations) and the type of computer used in the study (e.g., microcomputer or mainframe with terminals). Such coding categories would be irrelevant for most other areas of educational research. Examples of coding categories developed especially for meta-analyses in specific areas are presented in Table 4.1.

In addition to describing experimental treatments, meta-analysts usually wish to describe the experimental methodology used in a study. Features that are descriptive of both internal and external validity can be coded. Table 4.2 contains a list of study features that can be coded reliably in many areas of educational research. Coding studies on these categories does not require a high degree of inference.

Finally, most meta-analysts code studies for features of their settings and for publication source. Setting features that are usually of interest include the type of school in which the study was conducted, the ability level of the subjects treated, etc. An especially important publication feature is the type of

Table 4.1
Categories for Unique Features of Specific Research Areas

Computer-Based Instruction	Mastery Learning	Ability Grouping
Type of application	Pacing	By specific ability
Computer-assisted	Self-paced	No
Computer-managed	Group-paced	Yes
Computer-enriched		
	Mastery level required	Material adjusted
Type of interaction	70% or less	No
Off-line	71 – 80%	Yes
Terminal with mainframe	81 – 90%	
Microcomputer	91 – 100%	Flexible grouping
		No
Relation to group work	Tutorial assistance available	Yes
Supplementary	No	
Substitutive	Yes	Target group
		Disadvantaged
	Formal demonstration needed	All students
	No	Talented
	Yes	

Table 4.2
Categories for Coding Design Features

Subject assignment (random vs. nonrandom)
Control for instructor effects (same teacher for experimental and control groups vs. different teachers)
Control for historical effects (concurrent experimental and control classes vs. classes taught in different semesters or years)
Control for test author bias (standardized test vs. teacher-developed test)
Control for test scoring bias (objective test vs. essay test)
Statistical control for group equivalence (Use of covariates, blocking, or gain scores vs. no control)

publication in which the study was reported. Table 4.3 contains some features of settings and publications that have been coded in meta-analyses.

Most meta-analysts have given too little attention to the reliability of their coding of study features. Glass and his colleagues, however, have reported some data on this issue (Glass *et al.*, 1981, chap. 4). They found adequate reliability for the type of coding that meta-analysts typically do. They reported, for example, 75% agreement in coding of features of studies of drugs vs. psychotherapy. The meta-analytic literature does not contain many other reports on percentage agreement in using coding categories in meta-analysis, and it contains even fewer inter-coder reliability coefficients. Meta-analysts need to pay more attention to such matters.

Table 4.3
Categories for Coding Study Settings and Publication Features

Course content (mathematics, science, social science, language and reading, combined subjects, others)
Grade level (elementary, junior high, senior high, college, adult education)
Ability level of population (low, average, high)
Year of report
Source of report (unpublished, dissertation, published)

There is an alternative available for meta-analysts who do not report reliability of their coding. That is to include in their reports the main features of all the studies included in an analysis. Including such detail in a meta-analytic report lets other researchers determine for themselves whether individual studies have been assigned to correct coding categories. It makes it possible for researchers to redo meta-analyses with revised and expanded

coding of study features. The meta-analytic literature will not be a truly cumulative literature until meta-analytic reports include such detail routinely.

Describing Outcomes

An effect size is a general measure of the magnitude of a treatment effect on a dependent variable, expressed in such a way that the treatments in many different studies can be directly compared. Effect sizes have been measured in several different ways by meta-analysts. In a meta-analysis on Keller's Personalized System of Instruction, for example, we used as our measure of effect size the difference in percentage of right answers on final examinations of experimental and control students (J. Kulik, Kulik, & Cohen, 1979a). In a meta-analysis on advanced organizers, Hamaker used proportion of correct answers on experimental and control tests (Hamaker, 1986). Walberg and his colleagues have used cost per student hour as a measure of treatment effect in meta-analyses focusing on cost-effectiveness (Niemiec, Sikorski, & Walberg, 1987).

The most familiar measure of effect size, however, is the standardized mean difference between outcome scores of experimental and control groups. Cohen (1977), who popularized the idea of effect sizes in education and psychology, used the term *effect size* to describe two different kinds of quantities. We have called these quantities *operative* and *interpretable* effect sizes. The distinction between the two types of effect size is a critical one to grasp, both for those calculating effect sizes and for those who wish to read the meta-analytic literature critically.

Interpretable effect sizes are calculated by dividing a treatment effect expressed in raw units (y-units) by the standard deviation of y. Cohen used the symbol d to stand for the interpretable effect size calculated for a posttest-only, independent-group design. He added primes and subscripts to the symbol d to denote interpretable effect sizes calculated for other experimental designs. For example, Cohen used the symbol d_4' for the interpretable effect size calculated for one sample of n differences between paired observations (1977, p. 49):

$$d_4' = \frac{M_y^e - M_x^e}{s_y} , \qquad [4.1]$$

where M_x^e is the mean of the experimental group on a pretest and M_y^e is its mean on the posttest. Cohen (1977, p. 46) used the symbol d_3' for the interpretable effect size calculated for a study in which the mean of one experimental group is compared to a theoretical population mean. He pointed out, however, that all such interpretable effect sizes are conceptually equivalent and can be interpreted on a common scale. This is because the standardizing unit for interpretable effect sizes is always the standard deviation of y (or, s_y).

Operative effect sizes are a different matter. Operative effect sizes are calculated by dividing a treatment effect expressed in y-units by either the standard deviation of y or by a standard deviation from which sources of variation have been removed by one or another adjusting mechanism designed to increase power, e.g., covariance, regression, or blocking. Operative effect sizes are identical to interpretable effect sizes only in experiments that do not remove irrelevant variation from the dependent variable, e.g., unblocked, posttest-only experiments. For other experimental designs, operative effect sizes are calculated with special formulas. The operative effect size d for paired observations for one sample would be estimated from (Cohen, 1977, p. 63):

$$d = \frac{M_y^e - M_x^e}{s_y \sqrt{1 - r_{xy}}} . \qquad [4.2]$$

Cohen used the symbol d without subscripts or primes to represent operative effect sizes calculated for a variety of experimental designs. Although denoted by a common symbol, operative effect sizes calculated for different experimental designs are not conceptually equivalent because different standardizing units are used in calculating them. Operative effect sizes cannot therefore be interpreted in a single way. Operative effect sizes are useful, however, because they can be employed directly to find values of power in power tables.

A critical point to grasp is this: Meta-analysis must be based on interpretable, not operative, effect sizes if it is to produce interpretable results. Operative effect sizes have an undesirable property that makes them inappropriate to use in meta-analysis. With a given raw-score unit, they vary not only as a function of the size of raw treatment effect but also as a function of the experimental design used to investigate this effect. Two investigators studying the same phenomenon who find identical treatment effects (when effects are expressed in raw-score units) would report different operative effect sizes if they used research designs that controlled different amount of irrelevant variance in a posttest. The two investigators would report the same interpretable effect sizes, however, for identical raw treatment effects.

Meta-analytic methodologists have not stressed the distinction between operative and interpretable effect sizes in their writings, and meta-analytic practitioners have sometimes calculated operative effect sizes instead of interpretable ones for their analyses. Analyses based on operative effect sizes are flawed. The effect size values used in such analyses are almost invariably too large.

Rosenthal and Rubin (1978), for example, reported extraordinarily large effect sizes for a study by Keshock (1971) in which teacher expectancies were manipulated:

> Gains in performance were substantially greater for the children whose teachers had been led to expect greater gains in performance. The sizes of the effects varied across the four grades from nearly half a standard deviation to nearly four standard deviations. For all subjects combined, the mean effect size was 2.04 (p. 383).

The spectacularly large effect sizes reported for this study are surprising since the author of the study reported no significant effect of the treatment. Our reexamination of Rosenthal and Rubin's calculations showed, however, that the treatment effects of Keshock's study were standardized not in terms of variation in achievement but rather in terms of variation in achievement gains (J. Kulik & Kulik, 1986). In other words, Rosenthal and Rubin used operative effect sizes rather than interpretable ones in their meta-analysis. We calculated interpretable effect sizes for this study and determined that the average effect, rather than being 2.04, was 0.44.

Hedges's colleague Becker (1983) also reported operative effect sizes for studies in her analysis of gender differences in susceptibility to influence. Slavin and Karweit (1984) reported operative effect sizes in their analysis of effects of intra-class grouping. Because many meta-analysts do not report effect sizes by study in their reports, it is difficult to determine how many other meta-analytic reports in the literature are based on inflated operative effect sizes.

Other concerns about calculation of effect sizes seem less important to us than this one. For example, Glass (1981) has argued at some length that the proper unit for standardizing treatment effects is the control group standard deviation. He believes that experimental treatments may affect variation in performance as well as average performance, and he therefore recommends calculation of effect sizes from a standard deviation that cannot be affected by the experimental treatment: the control group standard deviation.

Some methodologists, however, have ignored Glass's advice and calculated effect sizes using pooled standard deviations. Their reasons for doing so are both practical and theoretical. Hunter and Schmidt, for example, recommend use of pooled standard deviations on theoretical grounds (Hunter *et al.*, 1982, chap. 4). Pooled standard deviations, they remind us, have smaller standard errors than those based on single groups. Other meta-analysts use pooled rather than control group standard deviations in calculating effect sizes as a matter of necessity. Many primary researchers do not report separate standard deviations for experimental and control groups when standard deviations are similar for the two groups. If effect sizes are to be calculated for such studies, they must be calculated using pooled standard deviations.

We think that control group standard deviations are the ideal standardizing unit when control groups are very large. When the control group is small, standardizing treatment effects on pooled standard deviations is probably a good idea. When control group standard deviations are not separately reported, we recommend calculating effect sizes using pooled standard deviations. We believe that meta-analysts will inevitably have to use some judgment in this matter. It is too much to expect that all analysts will find identical effect sizes for every study they examine. It is not too much to expect, however, that the

use of judgment will lead experienced analysts to calculate very similar effect sizes for the same studies.

Hedges (1982a) has argued that effect sizes calculated by Glass are not unbiased estimators of an underlying population parameter, and he has proposed a slight adjustment in effect sizes to make them unbiased estimators. As we have already pointed out, the correction that Hedges recommends has very little effect in actual data sets. In one of our analyses, we found that corrected and uncorrected effect sizes correlated .999 and that corrected and uncorrected effect sizes usually agree to two decimal places (Bangert-Drowns, Kulik, & Kulik, 1983). Although analysts are free to make Hedges' adjustments, we do not believe that making this adjustment has any important effect on analyses and we treat with equal respect analyses that make and do not make this adjustment.

Analyzing Results

The major problem in statistical analyses of effect size data, in our opinion, is inflation of sample sizes. Few meta-analysts are content to have their sample sizes be equal to the number of studies that they have located. Instead, they use study findings rather than the study as the unit in their analyses. In doing this, they seem to be following Glass's lead. Smith and Glass (1977) located 475 studies for their analysis of effects of psychotherapy, but they coded 1766 effect sizes. Glass and his colleagues located 77 studies for their analysis of effects of class size, but they coded 725 effect sizes (Glass, Cahen, Smith, & Filby, 1982).

Glass was aware of the problem of inflated sample sizes, however. He referred to it as the problem of lumpy data, and he considered it to be a serious issue. He was unwilling, however, to simply code one effect size for each separate study, and he used Tukey's *jackknife method* as a way of getting around the problem (Mosteller & Tukey, 1977). Meta-analysts who share Glass's reluctance to use the number of studies as a sample size have for the most part ignored the special techniques Glass advocated as a safeguard against inflated sample sizes.

Other analysts have come up with different solutions to the problem of exaggerated sample sizes. Gilbert, McPeek, and Mosteller (1977), for example, carried out an early study on the effectiveness of surgical treatments. They used survival percentage as the dependent variable and surgical intervention as the independent variable in the analysis. Because some studies located contained more than one finding on effectiveness of surgical intervention, they restricted the number of outcomes coded from any one study to two. When two comparisons were used, they gave each a weight of one-half. Their purpose, in their own words, was to prevent any one paper from having an undue effect on the total picture.

Our own solution is even more conservative. We seldom code more than one effect size from one study for any given analysis, but we often carry out

more than one separate analysis in a report. For example, in our meta-analysis on elementary computer-based education, we first carried out an overall meta-analysis on effects on student learning. Achievement effects for this analysis were averaged over boys and girls, grade levels, and subject matters. We next coded separate effect sizes for these various subgroups and carried out a separate set of analyses that examined effect sizes by sex, grade-level, and subject-matter. We also carried out separate analyses on other outcomes of computer-based instruction: attitudinal effects, time-on-task effects, and retention effects. In each analysis, however, the number of effects was equal to the number of studies with relevant data.

The reason for taking such pains in the analytic process is simple. We do not want to find spurious relationships that will not be replicated by other analysts. We want our findings to endure beyond the next meta-analysis. The best way for meta-analysts to ensure the longevity of their findings, in our opinion, is to carry out statistical analyses that show a respect for the way in which the sample of studies was drawn. When the sampling unit is the study, the number of studies ordinarily sets the upper limit for sample size.

The other feature that leads to unwarranted conclusions about relationships between study features and outcomes is the use of homogeneity testing procedures. We have shown in Chapter 3 that these procedures produce chi-square values that are spuriously high. Hedges (1983) has pointed out that these tests are analogous to analysis of variance tests, but he has failed to note that they are analogous to analysis of variance models that are inappropriate for meta-analytic data sets. Homogeneity tests are appropriate to use only if the analyst can make the assumption (a) that effects of a treatment do not vary from study to study, or (b) that the sample of settings included in a meta-analysis is exhaustive and no other settings exist to which findings may be applied.

Hunter and his colleagues have pointed out another problem in the statistical methods most often used in meta-analysis (Hunter *et al.*, 1982). Meta-analysts usually examine numerous relationships between study features and outcomes, but usually give little attention to the fact that they are carrying out multiple tests. Hunter, Schmidt, and Jackson point out:

> When effect size estimates are regressed on multiple study characteristics, capitalization on chance operates to increase the apparent number of significant associations. Since the sample size is the *number of studies* and many study properties may be coded, this problem is potentially severe. There is no purely statistical solution to this problem (p. 142).

Glass and his colleagues were aware of this problem from the start. In their meta-analyses on psychotherapy, they tried to overcome the problem by applying a clustering approach to their data (Smith & Glass, 1977). They cluster analyzed their pool of studies and found that the studies fell naturally into two basic types. One type consisted of behavioral therapies; the other type consisted of nonbehavioral therapies. Smith and Glass found that the outcomes of the two basic types of studies were similar.

Another reasonable approach to the problem of multiple measures is to rely on past findings and theory in examining relationships between study features and outcomes. After hundreds of meta-analyses, meta-analysts may be able to give up their unfettered empiricism. Rather than throwing scores of variables into a regression equation, they can select variables of theoretical interest for their analyses, or they can include in their analyses only those variables that have been shown to be important in past meta-analyses. Such careful and thoughtful pruning may cut down on the number of "significant" findings in meta-analyses, but it may also lead to a more cumulative science in the long run.

Conclusion

Meta-analysts can fail in many different ways. They can define the areas that they wish to analyze too broadly or too narrowly. They can be too lenient or too strict in rejecting studies with methodological flaws. They can try to investigate study features that cannot be coded reliably and objectively. In calculating effect sizes, they can use inappropriately restricted standard deviations, and in analyzing results, they can greatly inflate their true sample sizes. Making the wrong choices in a meta-analysis inevitably leads to results that cannot be replicated and conclusions that cannot be trusted.

It is possible, on the other hand, for meta-analysts to avoid many of these pitfalls. Their analyses can focus on a variety of credible studies of a well-defined experimental treatment. Coding of features and effects can be reliable and valid, and analyses do not have to be flawed statistically. If the meta-analyses of the future are carried out with greater attention to potential pitfalls, a clearer picture should emerge of the overall findings of educational research.

PART II: META-ANALYTIC FINDINGS

CHAPTER 5

DESIGN AND PUBLICATION FEATURES

Before Glass's development of formal methods for meta-analysis, reviewers had often speculated about the relationship between features of studies and their outcomes. Were results different in quasi-experiments and true experiments? Did results differ in long- and short-term studies of a phenomenon? Did study quality have an important influence on study outcome?

The very first meta-analysis conducted by Glass and his colleagues provided some evidence on the nature and size of such relationships. Smith and Glass's (1977) meta-analysis on psychotherapy effects showed that published studies indeed reported more positive effects than did unpublished ones. Glass also found a strong relationship between type of outcome measure and size of effects. Studies in which outcomes were measured as change in therapist ratings produced stronger results than did studies in which outcomes were measured as change in physiological responses, for example. With the publication of numerous other meta-analytic reports during recent years, it has become possible to say even more about features of studies that influence study outcomes. It is now fairly certain that studies of different types produce predictably different findings.

The purpose of this chapter is to explore meta-analytic findings on relationships between study features and outcomes. What relationships have been investigated? What has been found? How strong are the relationships? What might cause the relationships? How does the fact of such relationships affect our reading of the research literature? This chapter focuses on meta-analytic studies that are of a sufficient scale and methodological quality to provide dependable results. It examines the analyses for consistency of findings and for size of relationships, and it suggests several plausible interpretations of the observed relationships.

Data Sources

To summarize the findings on study features and outcomes in all meta-analyses would be a demanding task. Different meta-analysts have coded

different sets of study features and used different statistical methods in their analyses. Some have used sample sizes too small to produce dependable results, and some have used inflated sample sizes in their statistical tests, giving a false appearance of dependability to their results. Systematic analysis of such a mixed bag of findings might not repay the effort that went into it.

Our effort here is more limited. Instead of surveying a large and varied group of analyses, we survey a small group. All of the analyses we examine were conducted during recent years, and they have several features in common. First, each of the analyses covers approximately 100 separate studies. Second, each examines a common group of study features coded in a similar way in all studies. Third, each is based on analyses of interpretable not operative effect sizes. Fourth, in each meta-analysis, inferential statistics are based on sample sizes that are not inflated.

The first meta-analysis covered 99 studies of computer-based instruction at the elementary and secondary levels, and the second covered 155 studies of computer-based instruction at the postsecondary level (C. Kulik & Kulik, 1986; C. Kulik, Kulik, & Shwalb, 1986; J. Kulik, Kulik, & Bangert-Drowns, 1985; Bangert-Drowns, Kulik, & Kulik, 1985; C. Kulik & Kulik, 1988b). Included in these two meta-analyses are studies of computer-assisted, computer-managed, and computer-enriched instruction. The third meta-analysis covered 109 studies of ability grouping (J. Kulik & Kulik, 1987). Studies included in this meta-analysis focus on comprehensive between- and within-class grouping programs, special grouping of gifted youngsters, and Joplin plan approaches. The fourth meta-analysis covered 106 studies of mastery learning systems (J. Kulik, Kulik, & Bangert-Drowns, 1988). Included in the analysis are studies of both Keller's Personalized System of Instruction and Bloom's Learning for Mastery.

Findings

A few study features appear to be related dependably to study effects (Table 5.1). First of all, study effects are somewhat higher in journal articles than in dissertations. Findings in journal articles, however, do not differ systematically from findings in unpublished reports. Second, effects reported in short studies are somewhat larger than effects reported in long studies. Third, findings in studies with controls for instructor effects appear to be smaller than are findings in studies without such controls.

It is important to note that some study features often thought to relate strongly to effect size do not show a systematic relation in our analyses. True experiments with random assignment of subjects to groups produce about the same effects as do quasi-experiments. Experiments that use local tests produce about the same results as do experiments with commercial standardized tests. The older literature in a field produces about the same picture as does the recent literature.

Table 5.1
Average Effect Sizes By Study Feature in Four Meta-Analyses

Study feature	Precollege CBI		Postsecondary CBI		Mastery learning		Ability grouping	
	N	M_{ES}	*N*	M_{ES}	*N*	M_{ES}	*N*	M_{ES}
Random assignment								
Yes	27	0.26	62	0.33	28	0.52	24	0.09
No	67	0.32	86	0.29	73	0.52	91	0.21
Control for instructor effects								
Yes	34	0.23	90	0.26	45	0.52	24	0.08
No	57	0.35	49	0.42	53	0.52	91	0.21
Control for test author bias								
Local test	24	0.26	115	0.30	81	0.56	6	−0.13
Local & commercial	12	0.28	9	0.33	10	0.42	4	0.04
Commercial	58	0.33	24	0.33	10	0.30	105	0.21
Duration of treatment								
1 – 4 weeks	15	0.47	53	0.41	5	0.62	1	−0.35
5 weeks or more	75	0.27	95	0.24	96	0.52	114	0.19
Source of publication								
Unpublished	28	0.34	27	0.19	21	0.62	10	0.37
Dissertation	49	0.23	62	0.22	15	0.39	48	0.12
Published	17	0.46	59	0.44	65	0.52	57	0.20
Publication year								
Before 1960	0	— —	0	— —	0	— —	29	0.26
1960–1969	9	0.36	10	0.12	1	0.81	67	0.20
1970–1979	53	0.32	99	0.31	78	0.49	17	−0.03
1980-present	32	0.25	39	0.34	22	0.62	2	0.19

Reviewers sometimes disregard results from quasi-experiments on the grounds that such studies produce undependable results. They sometimes throw away results from studies using locally devised tests on the grounds that locally developed tests are generally biased. And they sometimes disregard results from older studies when results from newer studies are available. The meta-analyses that we have examined here suggest that none of these practices is defensible. Reviewers who limit their study pools on the basis of such methodological factors may be limiting the scope of the conclusions that they can reach.

It is also important to note that none of the relationships between study features and effect size can be characterized as strong. Effects usually vary a good deal when studies are characterized by a single feature. Distributions of effects usually overlap when studies are of different types. There is a good deal of overlap, for example, in the distribution of effects in dissertations and journal articles.

Certain study features are related to effect size in some areas, but not in others. The type of test used in evaluating experimental outcomes seems to be related to effect size in studies of mastery learning, for example, but it seems unrelated to effect size in studies of computer-based instruction and in studies of grouping. Conversely, controlling for instructor effects seems to make a difference in studies of computer-based teaching and in studies of grouping; it does not seem to have much of an effect in studies of mastery learning. The possibility of study features having different effects on findings in different areas considerably complicates the interpretational task for meta-analysts.

Interpretations

Meta-analysts can establish a correlation between effect sizes and study features fairly easily. Showing what causes the correlation is a more difficult task. Significant correlations sometimes arise from obscure and complicated causes.

Dissertations vs. Journal Articles

The relationship between publication source and study findings is a good case in point. Publication source of a study is often significantly related to study outcome. Results found in journal articles are usually more positive than are results from dissertations. In addition to the syntheses we have reviewed here, other authors have reported a difference between journal and dissertation results (Bangert-Drowns, Kulik, & Kulik, 1984; Glass *et al.*, 1981, pp. 64–68). The relationship between publication source and effect size is, in fact, one of the best documented findings in the meta-analytic literature.

But the explanation of the relationship is still controversial. A number of writers have attributed the difference in journal and dissertation findings to publication bias (e.g., Clark, 1985). This is the purported tendency of

researchers, reviewers, and editors to screen reports for publication on the basis of size and statistical significance of effects rather than on the basis of study quality. If it exists, such publication bias would make journals an unreliable source for information about the effectiveness of experimental treatments. Researchers would do well to avoid biased journals and base their judgments about research topics instead on unpublished findings in doctoral dissertations.

We have noted in our earlier articles, however, that journal studies and other studies are carried out by different persons working under different conditions (e.g., J. Kulik, Kulik, & Bangert-Drowns, 1985b). The typical author of a journal article differs from the typical dissertation writer in research experience, resources, professional status, and in many other respects. If the weakness of dissertation results is attributable to the inexperience of dissertation writers, then dissertations would be a poor source for information on the effectiveness of treatments.

Long vs. Short Studies

It is also clear that strong effects are more often reported in short than in long studies. That is, teaching interventions like computer-assisted instruction and mastery teaching look better when evaluated in short studies than in long ones. In studies lasting four weeks or less, the average effect of computer-based instruction, for example, is to raise test scores by 0.42 standard deviations. In studies where computer-based instruction is provided for several months, a semester, or a whole year, its effects are less dramatic. The average effect of computer-based instruction in such studies is to raise examination performance by 0.26 standard deviations.

Some have argued that this relationship is the opposite of what it would be if the treatment were truly effective. With brief programs, the argument goes, one should expect only small effects. With prolonged exposure to an effective treatment, however, effects should be larger. This argument would be correct if both long and short studies used the same outcome measures of achievement—say, an overall achievement test such as the Stanford Achievement Test. With a month of improved instruction, students in the experimental group might be able to surpass control group students in educational age by a few days, or a week at most. With a year-long improvement in instruction, gains of a few months would be possible, and such gains would translate into larger effect sizes. But short-term studies do not use the same tests as long term studies do. They use tests that cover small content areas in great detail. Longer studies use overall tests. With tests tailored to the amount covered in an instructional period, there is no reason to expect effect sizes, which are measured in standard deviation units, to be larger with long studies.

It is not clear, however, why effects should be significantly larger in short studies. Novelty, of course, could be a factor. A novelty, or Hawthorne, effect occurs when learners are stimulated to greater efforts simply because of the

novelty of the treatment. When a treatment grows familiar, it loses its potency. But it is also possible that shorter experiments produce stronger results because short experiments are more carefully controlled. In short experiments, it is usually possible to use precisely focused criterion tests, to keep control group subjects from being exposed to the experimental treatment, and so on. There is little empirical evidence available that allows us to choose between these explanations for the difference in findings of long and short experiments.

Control for Instructor Effects

The four meta-analyses also produced evidence that controlling for instructor effects by having a single instructor teach experimental and control groups influences the outcomes of evaluations. Effects were larger when different instructors taught experimental and control classes; effects were smaller when a single instructor taught both classes. This result showed up in studies of computer-based classes at the precollege level, in studies of computer-based classes at the postsecondary level, and in studies of ability grouping. Although the result did not show up in our latest meta-analysis of mastery-based teaching, it did show up in an earlier analyses of studies of Keller's Personalized System of Instruction (J. Kulik, Kulik, & Cohen, 1979a).

This effect can be produced by selective assignment of teachers to experimental and control groups in an experiment. If stronger teachers are usually assigned to experimental classes and weaker teachers to conventional classes, two-teacher experiments would be expected to produce stronger effects than one-teacher experiments. True effects attributable to the experimental treatment would be magnified by teacher differences in this case, and the more trustworthy studies would be those that control for instructor effects. Another possible explanation for the effect, however, is treatment contamination. If teaching an experimental class has generally beneficial effects on a teacher's performance, two-teacher experiments would also be expected to produce stronger effects than one-teacher experiments do because the control groups in one-teacher experiments would get some of the benefits of the treatment. The important point, however, is that in this case, the more trustworthy experiments would be those without a control for instructor effects because such experiments would be freer than other experiments from the effects of treatment contamination.

The implications of these two hypotheses are different. If one- and two-teacher experiments produce different results because of selective assignment of teachers in two-teacher studies, then reviewers should discount findings from two-teacher experiments. If one- and two-teacher experiments produce different results because of treatment contamination in one-teacher studies, then reviewers should discount findings from one-teacher experiments. At this point not enough evidence is available to choose between these two alternative explanations, and no one can say for certain which type of study produced the more accurate evidence about treatment effects.

Conclusion

The primary lesson that meta-analysis has taught us is to beware of strong claims about relations between study features and outcomes. Study features are not often significantly related to study outcomes in carefully done meta-analyses. And in the analyses where relations between study features and outcomes are significant, the relations are seldom strong. The variation in study outcomes that characterizes experimentation in the social sciences and perhaps in the natural sciences as well cannot be explained so simply as some reviewers hope it can. Glass has expressed the situation very well:

> . . . the findings of contemporary research fit together poorly. Variance of study *findings* is only modestly predictable from study characteristics; in nearly all instances, less than 25% of the variance in study results can be accounted for by the best combination of study features . . . The condition of most social and behavioral research appears to be that there is little predictability at either the individual or study level (Glass *et al.*, 1981, p. 230).

CHAPTER 6

INSTRUCTIONAL SYSTEMS

Skinner's development of programmed textbooks and teaching machines in the late 1950s ushered in a new era in the history of educational innovation. The era was distinguished both by its commitment to instructional research and by its reliance on laws of learning as a guide in instructional development. Among the instructional systems that became well-known during this time were programmed instruction, modular instruction, and mastery learning systems. In addition to these behaviorally based systems, other approaches became popular that placed more emphasis on social factors and individual choice in learning. Included among these systems were peer and cross-age tutoring and open education.

These instructional systems have proved to be of great interest to meta-analysts. Their reviews of the literature on instructional systems provide a basis for estimating the size of effects ordinarily achieved in real-life settings with the introduction of innovative systems of instruction.

Computer-Based Instruction

CBI programs have been developed to supplement or replace elements of a variety of conventional programs for a variety of learners. These programs typically follow a format originally popularized by Skinner and other proponents of programmed instruction during the 1950s and early 1960s. The format calls for careful attention to instructional objectives, development of a careful sequence of small steps that learners can follow to master these objectives, provision of opportunities for learners to respond frequently during instruction, and provision of immediate feedback to learners on the correctness of their responses. CBI programs developed in this format may provide students with drill and practice on material originally presented in a more conventional format; they may provide tutorials in which students are taught new facts and concepts and are quizzed on their understanding of these concepts; they may provide computer evaluation of student performance on

formative quizzes, along with careful record keeping on student performance; or they may provide a combination of some or all of these services.

Performance on Examinations. Meta-analyses show that CBI on the average increases examination scores by about 0.35 standard deviations in studies carried out in precollege settings. In meta-analyses conducted with Bangert-Drowns, we found an average effect size of 0.42 in 32 studies at the elementary school level (J. Kulik, Kulik, & Bangert-Drowns, 1985) and an average gain of 0.26 in 42 studies carried out in secondary schools (Bangert-Drowns, Kulik, & Kulik, 1985). In a recent update of these analyses, we reported an average gain of 0.31 standard deviations in a total of 99 studies in elementary and high schools.

Findings of other meta-analysts on elementary and secondary school CBI are similar. Niemiec and Walberg reported a gain of 0.37 standard deviations in 48 elementary school studies (Niemiec, 1985; Niemiec & Walberg, 1985). Hartley (1977) reported an average gain of 0.41 standard deviations in 33 studies in elementary and high school mathematics. Willett, Yamashita, and Anderson (1983) reported an average gain of 0.22 standard deviations in 11 studies of precollege science teaching. Schmidt, Weinstein, Niemiec, and Walberg (1985) reported an average CBI effect of 0.57 standard deviations in 18 studies in special education. Burns (1981) found an average CBI effect of 0.36 standard deviations in 44 studies in which computers provided drill and practice or tutorial instruction in elementary and high school mathematics.

CBI has a similar record in evaluations carried out with older learners. Our meta-analysis with Shwalb found an average effect size of 0.42 standard deviations in 23 studies involving technical training or basic skills instruction for adult students (C. Kulik, Kulik, & Shwalb, 1986). Our meta-analysis of college findings on CBI covered 99 studies and yielded an average effect of 0.26 standard deviations (C. Kulik & Kulik, 1986). In our updated analysis, we found an average effect of 0.30 standard deviations in 149 postsecondary studies (C. Kulik & Kulik, 1988b).

Attitudes. Our meta-analyses have examined attitudinal effects of CBI separately from other effects. These meta-analyses show that CBI promotes positive attitudes toward computers. The average effect of CBI in a total of 6 precollege studies was to raise ratings of computers by 0.51 standard deviations; the effect was 0.27 standard deviations in 13 studies at the college level (C. Kulik & Kulik, 1988b).

This meta-analysis also showed that attitudes toward courses improve when CBI is part of the teaching program. Average effect size on attitudes toward courses was 0.28 standard deviations in 22 studies; the effects were 0.39 in 2 precollege studies and 0.27 in 20 postsecondary studies.

Our meta-analysis also suggested that CBI does not have much of an effect on student attitudes toward the subject being taught. Average effect size was 0.05 in 34 studies. It was 0.06 in 16 studies at the precollege level and 0.04 in 18 studies at the postsecondary level.

Instructional time. Several meta-analyses have reported that CBI can be used to reduce instructional time. Orlansky and String (1979) found that CBI reduced instructional time by 30% in 30 reports of CBI in military settings. In a meta-analysis with Shwalb, we found a 29% reduction in instructional time associated with use of CBI in 13 studies of adult education (C. Kulik *et al.*, 1986). We also found that CBI reduced instructional time by 38% in 11 studies carried out in college courses (C. Kulik & Kulik, 1986). Finally, in an updated analysis, we found a 30% reduction in instructional time in 32 studies of postsecondary education (C. Kulik & Kulik, 1988b).

In summary, evaluations typically show that CBI programs increase instructional effectiveness. Students in programs with CBI components typically (a) outperform control students on course examinations and standardized achievement measures, (b) develop more positive attitudes toward computers, and (c) hold more positive attitudes about their courses than control students do. In addition, programs with CBI components often reduce the time needed for instructing students.

Keller's Personalized System of Instruction

Keller's PSI is a self-paced, mastery-oriented teaching method used primarily in college courses. Keller (1968) described the five central features of the method in a widely cited paper: PSI courses are (a) mastery oriented, (b) student proctored, and (c) self-paced courses that use (d) printed study guides to direct student learning and (e) occasional lectures to stimulate and motivate the students. Keller and his colleagues have also written a number of guides for use of PSI. The model of teaching has been so explicitly described and widely discussed that teachers throughout the world have been able to use it in their courses.

Performance on examinations. In a recent meta-analysis, we synthesized findings on achievement examinations in 72 separate evaluations of the use of PSI in college courses (J. Kulik, Kulik, & Bangert-Drowns, 1988). In the typical study, students taught by PSI scored 0.49 standard deviations higher than control students did on examinations. The effects of PSI were equally clear on objective and short-essay exams and on teacher-made and national tests. We also reported that the typical improvement on a 100-point final examination was 7.7 percentage points (J. Kulik *et al.*, 1979a). In an analysis of results from approximately 50 PSI and PSI-like courses, Robin (1976) reported similar but slightly higher improvement scores. In a quantitative synthesis of findings from studies in the natural sciences, Aiello (1981) reported a slightly smaller PSI effect of 0.36 standard deviations. Willett *et al.* (1983) reported a gain of 0.49 standard deviations for PSI learners in 7 studies conducted in precollege science courses.

Follow-up Performance. Nine of the studies that we reviewed reported results from retention measures administered weeks or even months after the conclusion of a course (J. Kulik *et al.*, 1988). The average effect of PSI in these studies was to raise student performance by approximately 0.83 standard deviations. Thus, the effects of PSI were clearer on follow-up than on final examinations.

Attitudes. Our meta-analysis with Bangert-Drowns also provided evidence for positive affective results of PSI (J. Kulik *et al.*, 1988). Sixteen of the studies in the meta-analysis contained results on overall satisfaction with instruction; 11 of these studies contained results on overall satisfaction with course quality. The average effect of PSI was to raise ratings of satisfaction with course quality by 0.41 standard deviations. Thus, students liked courses better when they were taught with PSI. Ten studies in this meta-analysis contained findings on attitudes toward subject matter. The average effect of PSI was to raise scores on these attitude measures by 0.38 standard deviations. Students therefore liked their subjects more when these subjects were taught with PSI.

Instructional Time. Our meta-analysis with Bangert-Drowns covered 5 studies that examined student use of time in both PSI and control classes (J. Kulik *et al.*, 1988). Overall, PSI students spent about 10% more time on courses. We concluded that PSI and control classes made roughly equal demands on student time, but the two kinds of courses demanded different kinds of time commitments. PSI classes required students to spend more time on individual work; conventional classes required students to spend more time in lecture attendance, class discussions, etc.

In summary, evaluation studies have typically shown that students in PSI courses: (a) outperform control students on course examinations and on standardized tests; (b) outperform control students on follow-up examinations administered weeks or months after completion of a course even when these follow-up courses are taught by conventional means; and (c) express more positive attitudes towards their courses and the subject matter taught than do other students. In addition, PSI and conventional classes make roughly equal demands on students in terms of amount of time but quite different demands in terms of the way students use this time.

Bloom's Learning for Mastery

Bloom's Learning for Mastery is a group-based, mastery-oriented teaching approach. Bloom's model requires teachers to administer formative quizzes to students to diagnose learning weaknesses and then to provide special remedial activities for those students who fail to perform at a predetermined level on the tests. Evaluations suggest that precollege and college teachers who follow Bloom's LFM model will improve their instructional effectiveness.

Bloom's LFM model was described at length in a paper that suggested that all students could achieve at the same high level in courses if they were given the time and type of instruction that they individually needed (Bloom, 1968). Individual pacing and a high mastery standard, stressed in Bloom's original writings on learning for mastery, received less emphasis in later practitioner-oriented writings. In recent years LFM has become a group-based approach. In actual applications, mastery requirements seem to have been relaxed, and in some cases, remediation is required only of those who score below 80% on formative quizzes. In addition, mastery is sometimes assumed for students who simply attend a remedial session (sometimes a group discussion). Actual LFM courses can therefore be quite diverse in their features. The common denominator of LFM courses, however, is frequent formative quizzing and remediation sessions following the quizzes.

Performance on examinations. A number of researchers have reported syntheses of findings on Bloom's LFM. The researchers have examined different parts of the LFM literature, and they have reached different conclusions about LFM's effectiveness. At one extreme, Bloom's synthesis suggested that LFM might raise student achievement by one to two standard deviations (Bloom, 1984). An analysis by Guskey and Gates (1985), covering studies of both college and precollege implementations, suggests strong positive effects of LFM, but no effects as strong as those reported by Bloom. Average effect in the 35 studies reviewed by Guskey and Gates was an increase in exam performance of 0.78 standard deviations. At the other extreme, Slavin (1987) concluded from his synthesis of results from 17 studies that there was no evidence that LFM raised student achievement at all. Although the average effect of LFM in the 17 studies examined by Slavin was an increase in student performance of 0.26 standard deviations, the average effect size was virtually zero in the 7 studies that Slavin considered to be adequately controlled.

None of these syntheses, however, seems completely adequate. Bloom's (1984) review is too selective to warrant serious attention. It focuses on a few dissertations carried out by Bloom's students at the University of Chicago, none of which provides an adequate comparison of LFM versus conventional instruction. The average effect size reported by Guskey and Gates (1985) is inflated by methodological flaws in their synthesis. Some of the effects reported by Guskey and Gates were standardized on between-class rather than within-class standard deviations. Other effects were calculated from formative quizzes rather than summative tests. When questionable effect sizes are eliminated from Guskey and Gates' synthesis, the average effect size drops to 0.47 standard deviations (C. Kulik & Kulik, 1986–87).

Slavin's (1987) synthesis was meant to cover precollege studies that compared results of group-based mastery and conventional teaching procedures. Only 7 of the 18 studies located by Slavin actually meet this criterion, however (C. Kulik & Kulik, 1988a). Some of the studies compare individualized mastery learning programs to individualized programs without mastery procedures. In other studies, the experimental treatment does not

involve a requirement of mastery. In addition, Slavin's review does not cover some sound comparisons of LFM and conventional instruction. Like Bloom's review, Slavin's seems idiosyncratic and selective.

In our recent comprehensive meta-analysis with Bangert-Drowns, we found that the average effect from LFM procedures was an increase of 0.48 standard deviations in examinations in 15 precollege studies and an increase of 0.68 standard deviations in 19 studies at the college level (J. Kulik *et al.*, 1988). Willett *et al.* (1983) reported a gain of 0.50 in 8 mastery learning studies conducted in precollege level science courses—a result that is very similar to ours. Overall, therefore, Bloom's approach seems to be at least as effective as Keller's.

Instructional time. Slavin (1987) has suggested that the gains achieved in LFM classes might come at an instructional cost. The LFM approach requires teachers to add formative tests and remediation sessions to instruction without making other changes in teaching. The hope of the originators of LFM was that increased student time requirements in the early part of a course, however, would be offset by quicker learning and therefore reduced time requirements in the later parts of a course. So far, there is no good evidence available that LFM greatly increases or decreases instructional time requirements. In our meta-analysis, we found that in three implementations of LFM, instructional time for LFM students was approximately 10% greater than that for conventional students (J. Kulik *et al.*, 1988).

In summary, Bloom's LFM procedures seem to produce notable positive effects on the learning of both school children and college-age learners. Although attitudinal effects of the method have not been studied often, what little evidence exists suggests that such effects are also positive. Although Bloom's method does not reduce instructional time, increased time demands of LFM procedures are probably not prohibitive for actual classrooms.

Individual Learning Packages

This approach was popularized by Glaser in his Individually Prescribed Instruction, but individual learning packages have been used in many other teaching systems (Glaser & Rosner, 1975). In this approach, students are provided with packages (or modules) on which they can work individually at their own pace. The learning packages often contain objectives and self-quizzes, and students are often free to choose among several alternative ways of mastering the objectives.

Systems of individualized instruction that use learning packages usually follow a basic diagnostic-prescriptive teaching cycle. The teacher first finds out what the pupil knows and then provides learning materials that are appropriate for the pupil. After the pupil works individually with the materials, the teacher assesses progress and finally requires additional work

and additional evaluation if the pupil has not reached the mastery level. Systems like IPI were derived around the country and given names like Unipac, Omapac, etc. Such systems are obviously related to Keller's PSI and Bloom's LFM, but there is a difference in emphasis. Keller's PSI emphasizes repeated testing and individual remediation until a student reaches a mastery standard; Bloom's method emphasizes frequent quizzes and remediation sessions; in Glaser's IPI and related systems, the emphasis is on specially constructed modular materials, student self-pacing, and multiple entry and exit points.

Performance on examinations. Hartley (1977) synthesized findings from 51 studies of ILP in elementary and secondary mathematics classes. She found that ILP raised student achievement by only 0.16 standard deviations. She pointed out that while results were disappointing for ILPs in general, they were especially poor for IPI. Our meta-analysis with Bangert synthesized findings from 49 studies of ILPs in Grades 6 through 12 (Bangert, Kulik, & Kulik, 1983). We found that ILPs raised student achievement by only 0.10 standard deviations, and we were unable to identify study features that were related to size of effects. Willett *et al.* (1983) found an average effect of 0.12 in 102 studies of use of ILPs in precollege science teaching.

Attitudes. Our meta-analysis with Bangert also looked at outcomes from ILPs in other areas (Bangert *et al.*, 1983). We reported that courses that used ILPs extensively had about the same effects on student attitudes as did other courses. For example, we found that ILPs raised student attitudes toward subject matter by only 0.14 standard deviations, a small effect. Willett *et al.* (1983) also examined effects of ILPs on student attitudes toward science. They found an average effect of 0.16 standard deviations in 10 studies, a finding very similar to our own.

In summary, the record of effectiveness of this system of instruction is very modest. Effects of ILPs on student examination scores are positive but quite small, averaging about 0.15 standard deviations. Average improvement in student attitudes toward subject matter are similar, also averaging about 0.15 standard deviations.

Programmed Instruction

Programmed instruction was described by Skinner (1954) in a now-classic article "The Science of Learning and The Art of Teaching." The principles Skinner hoped that programmed teaching machines would embody were instruction in small steps, learner response at each step, and immediate feedback on the adequacy of the learner's response. Originally presented on teaching machines, PI was later presented in textbook form. The short fill-in-

the-blank frames of programmed instruction are now commonplace in instructional manuals and school workbooks.

Hartley (1977) synthesized findings from 40 studies of PI in precollege mathematics teaching. She reported that PI's average effect was to raise student achievement by 0.11 standard deviations. Our meta-analysis with Shwalb found a similar average effect size of 0.08 standard deviations in 47 studies at the precollege level (C. Kulik, Kulik, & Shwalb, 1982). Our meta-analysis with Cohen and Ebeling reported a slightly larger average effect (0.24) in 56 studies of the use of PI in college courses (Kulik, Cohen, & Ebeling, 1980). Aiello (1981) restricted her analysis to studies carried out in natural science courses. She found an average effect size of 0.24 in 23 studies at the precollege level and 0.29 in 22 studies at the college level. Willett *et al.*'s analysis (1983) covered effects of PI in precollege science courses. These investigators found an average effect size of 0.17 standard deviations in 51 studies.

Three of these meta-analyses reported a significant positive relationship between size of study effect and year of study. PI has apparently produced stronger effects in studies carried out in more recent years. In our meta-analysis with Shwalb, studies carried out before 1965 yielded an average effect size of -0.03 standard deviations; studies carried out after 1965 yielded an average effect size of 0.18. Hartley's (1977) meta-analysis and our meta-analysis with Cohen and Ebeling (J. Kulik *et al.*, 1980) reported similar relationships between study finding and study year. If the relationship is, in fact, a real one, it can have several explanations. One possible explanation is that PI has improved with time. The PI of today may be better designed than the PI of yesterday was. Another explanation is that null results on PI became old hat during the 1960s; nowadays only PI success stories get written up.

It is notable that PI and CBI are blood relatives. The relationship between them is the relationship of parent to child, and a family resemblance is clear. Both PI and CBI employ short instructional frames; learners provide a response; they receive immediate feedback on their responses. In spite of such similarities, the two approaches differ in potency, especially for younger learners. CBI's effects are unmistakable; PI's are very small.

Why is PI's record so unimpressive? One suggestion is that the textbooks do not really provide a good medium for programmed teaching. Although learners are required to make a response on each frame, learners can easily get around this requirement. They can simply copy their "feedback" into answer spaces, subverting the whole idea of PI. In CAI, learners are unable to subvert the intentions of instructional designers so easily. Another suggestion is that programmed instruction grows tiresome when the learner has control over every aspect of the timing. The relatively poor ratings of courses taught with PI show that students are not very enthusiastic when PI is presented through print media. They are enthusiastic, however, when computers are used to present carefully sequenced instruction.

In summary, use of PI leads to only small improvements in effectiveness of precollege and postsecondary courses. Students taught by PI score only

slightly higher on examinations than do students taught by more conventional methods. Recent studies, however, found results that are more favorable for PI than those in early studies.

Postlethwait's Audio-Tutorial Approach

The AT approach dates back to 1961 when biologist Samuel Postlethwait began developing audiotapes and other visual and manipulative materials for remedial instruction in his introductory botany course at Purdue University (Postlethwait, Novak, & Murray, 1972). When Postlethwait's initial efforts proved successful, he decided to convert his entire course to an audio-tutorial approach. The revised course had three major components: independent study sessions, in which students learned from audiotapes and other media in self-instructional carrels; general assembly sessions, held each week, which were used for guest lectures, long films, and major examinations; and weekly integrated quiz sessions, which were held for groups consisting of between six and ten students and an instructor. Like Keller's PSI, Postlethwait's methods have been used primarily in college courses.

Our meta-analysis with Cohen covered results from 48 studies and showed that AT instruction has a significant but small overall effect on student achievement in college courses (J. Kulik, Kulik, & Cohen, 1979b). Students who received AT instruction scored 0.2 standard deviations higher on examinations than did control students. Aiello's synthesis (1981) covered 26 studies in the natural sciences at the college level and reported a similar average effect: an improvement of 0.24 standard deviations on examination scores in AT courses. Willett *et al.* (1983) reported a somewhat smaller gain (0.09 standard deviations) in 5 studies of AT in precollege science courses. Our meta-analysis with Cohen in 1979 also reported that AT instruction has no significant effect on student course evaluations or on course completions.

In summary, evaluations show that AT has a modest record in improving instructional effectiveness. Its effects on student learning, as measured by examination performance, are positive but small. Its effect on student attitudes toward instruction are equally small.

Media-Based Instruction

A number of factors have contributed to the growing interest in instructional media during the last few decades. Among these were an actual and anticipated shortage of teachers, the increasing sophistication of visual technology, and research findings that seemed to validate the use of media as a legitimate channel for educational content. Meta-analysts have examined these research findings in the past few years in order to reach overall conclusions about size of media effects.

Performance on examinations. Our meta-analysis with Cohen and Ebeling (Cohen, Ebeling, & Kulik, 1981) synthesized findings on 64 studies on media-based instruction (MBI) in college courses. Media used in the 64 studies included still projection, film, closed circuit television, and educational television. We found that on the average the use of media-based instruction raised student achievement by 0.15 standard deviations, a small positive effect. We also found that use of a control for instructor effects significantly influenced study results. Studies with different instructors for experimental and control groups reported higher effect sizes than did studies with the same teachers for both comparison groups. We also found that achievement effects differed as a function of publication year. Studies conducted in recent years reported higher effect sizes than did studies of earlier years.

Willett *et al.* (1983) synthesized findings from 75 studies of MBI in elementary and high schools. They found that the effect of MBI was to lower student achievement by 0.03 standard deviations, a trivial effect.

Attitudes. Our meta-analysis with Cohen and Ebeling in 1981 also investigated effects of MBI in other outcome areas. We found that MBI had slightly negative effects on student ratings of both instruction and subject matter. The average effect was −0.06 for 16 studies on attitudes toward instruction and −0.18 for 10 studies on attitudes toward subject matter. Willett *et al.* (1983) also examined the effects of MBI on students' attitudes toward science. They found a slightly negative effect (−0.10 standard deviations) in 16 precollege studies.

In summary, therefore, effects of MBI on students appear to be trivially small in both the cognitive and affective areas. Courses in elementary schools, high schools, and colleges produce roughly the same results when given with and without media aid.

Tutoring Programs

Peer tutoring and cross-age tutoring programs have been used widely in elementary and secondary schools during recent decades. In peer programs, pupils who have a better grasp of material in a specific area or on a specific topic tutor their same-age classmates. In cross-age programs, older pupils tutor their younger schoolmates. Tutors in highly structured and cognitively oriented programs follow a script devised by the tutoring supervisor. In less structured programs, tutors and tutees play a larger role in determining the agenda of tutorial sessions.

Performance on examinations. Hartley's (1977) meta-analysis covered 29 studies of tutoring programs in mathematics offered in elementary and secondary schools. Average effect size in the 29 studies was 0.60. Our meta-analysis with Cohen covered 52 reports on effects of tutoring on tutored

students (Cohen, Kulik, & Kulik, 1982). Most of the programs covered in the 52 reports provided tutoring in mathematics and reading, and all were offered in elementary and secondary schools. Average effect size for the mathematics tutoring programs was 0.62; average for tutoring in reading was 0.29; overall average effect size was 0.40.

We also reported finding six significant relationships between study features and effect sizes. Tutoring effects were larger in programs of shorter duration, in published rather than dissertation studies, and on locally developed rather than nationally standardized tests. Tutoring effects were also larger in more structured programs, when lower level skills were taught and tested on examinations, and when mathematics rather than reading was the subject of tutoring. These results were consistent with findings of Hartley's meta-analysis and with findings of other reviewers of the literature of tutoring.

Attitudes and self-concept. We also reported that in 8 studies, student attitudes were more positive in classrooms with tutoring programs, with the average effect size being 0.29. Finally, we reported that 9 studies contained results on effects of tutoring programs on tutee self-concept. In 7 of these, self-concepts were more favorable for students in classroom with tutoring programs. The average effect size in the 9 studies was 0.42.

Our overall conclusion therefore is that evaluation results for tutoring programs are usually positive: (a) students who receive tutoring in such programs outperform students in comparison groups on final examinations; (b) attitudes of tutored students become more positive toward the subject matter in which they are tutored; and (c) tutored students develop more positive self-concepts.

Open Education

Open education is "a style of teaching involving flexibility of space, student choice of activity, richness of learning materials, integration of curriculum areas, and more individual or small-group than large group instruction" (Horwitz, 1979, pp. 72–73). Goals of open education include the development of student responsibility for learning and honesty and respect in interpersonal relationships. Horwitz (1979) identified about 200 empirical studies that evaluated open education programs. Peterson (1979) conducted a meta-analysis of 45 of these studies. Later Hedges, Giaconia, and Gage (in Giaconia & Hedges, 1982) used most of Horwitz's studies in a meta-analysis covering 158 reports.

Performance on examinations. Peterson (1979) found an average effect size of −0.12 for student content learning. Hedges and his colleagues (1981) found an average effect size of −0.07 (Giaconia & Hedges, 1982). They also carried out extensive analyses to find factors that would explain variation in effect sizes,

Table 6.1
Estimated Average Effects of Instructional Systems in Three Major Outcome Areas

System	Examinations	Attitude toward instruction	Attitude toward subject matter	Instructional time
		Precollege effects		
CBI	0.35	0.40	0.05	—
PSI	0.50	—	—	—
LFM	0.50	—	—	—
ILP	0.15	—	0.15	—
PI	0.15	—	−0.15	—
MBI	−0.05	—	−0.10	—
Tutoring	0.40	—	0.30	—
Open education	−0.10	—	—	—
		Postsecondary effects		
CBI	0.35	0.25	0.05	−30%
PSI	0.50	0.40	0.40	+10%
LFM	0.70	—	—	—
PI	0.25	—	—	—
AT	0.20	0.10	—	—
MBI	0.15	−0.05	−0.20	—

but they were unsuccessful in this search. The average effect sizes of well-designed and poorly designed studies, for example, did not differ consistently.

Creativity. Results on creativity measures were only slightly stronger. Peterson's finding was an average effect size of 0.18 in 11 studies, and Hedges and his colleagues found an average effect size of 0.29 in 21 studies.

In summary, the record of effectiveness of programs of open education does not inspire confidence about this approach. The examination performance of pupils in such programs is slightly lower than the examination performance of pupils in conventional programs. Performance on creativity measures in open education programs is only slightly higher than the performance on such measures of those in conventional programs.

Conclusion

Our summary of the results of these meta-analytic reviews is presented in Table 6.1. The effects presented there are rounded average effects from the most dependable large-scale reviews. Results in the table cover major outcomes only. Outcomes examined in only one or two meta-analyses and in very few primary studies are not tabulated.

Several conclusions seem warranted. First, most of the well-known systems devised for improving instruction have acceptable records in evaluation studies. Second, the evaluation records are not all equally impressive. When student learning is taken as the criterion of instructional effectiveness, the records of the mastery-based systems of teaching—Bloom's LFM and Keller's PSI—are most impressive, but CBI and tutoring programs also make notable contributions to student learning. Open education programs and programs involving visual media seem not to have positive effects. Third, only CBI has proved effective so far in reducing the amount of time needed for instruction.

CHAPTER 7

INSTRUCTIONAL DESIGN

Since the 1960s educational researchers have carried out hundreds of studies of the effects of variations in design of instructional materials. Some researchers have modified instructional texts by adding learning objectives, questions, or introductory material to text passages in an attempt to make the passages easier to learn. Others have modified tests and testing procedures. Among the test features of interest have been test frequency, amount of feedback on test items, feedback timing, and the use of a mastery requirement.

Meta-analysts have recently begun to integrate results from these research studies. The findings from these meta-analyses are worth comparing to findings of meta-analyses on results from applied studies. Do the short, controlled laboratory and classroom studies of design features produce stronger effects than field evaluations do? Is there less variability in results? This chapter addresses these questions.

Instructional Text

Much of the teaching that takes place both in school and out of school is done via instructional texts. Improving the quality of instructional texts will therefore directly add to the quality of education. In addition, improving the quality of texts may make indirect contributions to classroom instruction. It seems possible that factors that improve textbook presentations will also prove to be important in other teaching situations. Learning how to produce quality text may help us to become better teachers.

Instructional designers have tried to improve texts by presenting text passages along with (a) advance organizers, (b) adjunct questions, and (c) behavioral objectives.

Advance Organizers

Advance organizers, as defined by Ausubel (1960), are introductory learning materials that give a learner a bridge between what is already known

and new material to be learned. Advance organizers play an important role in Ausubel's theory of meaningful learning. According to the theory, subsuming concepts can serve as anchors for specific pieces of new information, helping learners to organize the new information around common themes. When learners lack appropriate subsuming concepts for a learning task, advance organizers can be constructed that provide the necessary concepts.

Luiten, Ames, and Ackerson (1980) located 134 studies of the effects of advance organizers. Most of these studies were brief ones, concluding within one or two class periods. The 134 studies contained 100 effect sizes calculated from scores on tests given immediately after completion of the learning task. The average effect size was 0.21. The studies also contained 50 effect sizes calculated from results on retention tests given more than 24 hours after the completion of the learning task. Average retention effect size was 0.25.

Stone (1982) reviewed results from 29 reports on effects of advance organizers. These reports yielded 112 effect sizes with a median effect of 0.48. Stone reported that the median provided a better measure of typical effect than the average did. Like Luiten *et al.* she found that effects were retained in studies of longer duration.

Moore and Readence (1984) examined effects of graphic organizers, a special type of advance organizer, on learning from text. Graphic organizers portray the relationships among key terms used in a learning task and are usually presented in the form of tree diagrams. Moore and Readence's meta-analysis covered results from 23 studies, which yielded 161 effect sizes. They found that the average effect of graphic organizers was to increase student performance on criterion tests by 0.22 standard deviation. Their average effect size must be treated with some caution, however, because it is not consistent with effect sizes listed in their table of graphic organizer effect sizes by study variables. A reporting error seems a real possibility.

Adjunct Questions

Adjunct questions are questions added to instructional text to influence learning from the text. Rothkopf (1966), who carried out pioneering studies on adjunct questions, referred to such questions as "test-like events."

Hamaker (1986) has pointed out that experimental designs used in studies of adjunct questions vary in several important ways. For example, adjunct questions may precede text passages (prequestions) or they may follow the text (postquestions). In addition, the adjunct questions may be identical to questions on the criterion examination (repeated questions); they may be different from but related to the criterion questions (related questions); or they may be unrelated (unrelated questions).

Hamaker's review (1986) covered 61 experiments, which yielded more than 300 effect sizes. Hamaker's results for effects of factual prequestions and postquestions on repeated, related, and unrelated criterion items appear in Table 7.1. It seems clear that factual prequestions and postquestions facilitate the learning of material covered either directly (through repeated test

Table 7.1
Average Effect Sizes for Factual Adjunct Questions by Position and Criterion Test

Question position	Criterion test type		
	Repeated	Related	Unrelated
Prequestions	0.94	0.56	−0.31
Postquestions	1.00	0.44	0.06

questions) or indirectly (through related test questions) and that factual prequestions and postquestions have little or no effect on unrelated criterion questions. Because of the lack of independence among effect size measures in Hamaker's analysis, it is hard to determine how much confidence to put into his analyses of relationships between study features and outcomes. The evidence suggests, however, that higher-order questions may produce even stronger results than factual questions do.

Results of Lyday's (1983) meta-analysis of 65 studies are consistent with these results. Lyday found an overall average effect size of 0.57 standard deviations associated with the use of adjunct questions. Like Hamaker (1986), she found that the effects of adjunct questions are greater for intentional learning (repeated and related questions) and less for incidental learning (unrelated questions). She also found some evidence to suggest that effects are greater for conceptual questions than for factual ones.

Behavioral Objectives

Behavioral objectives specify in clear terms the behaviors that learners should be able to perform after completing some learning task. Mager's book *Preparing Instructional Objectives* (1962) popularized the use of such objectives in instruction. Stating objectives in behavioral style helps instructors to design appropriate learning tasks and develop appropriate tests of learner performance. And most important, behavioral objectives help learners in their study efforts. It is this last function of behavioral objectives that has been the focus of meta-analytic study.

Asencio (1984) reviewed 111 studies reporting on the effects of behavioral objectives. She found a small positive effect on student achievement in general. In 97 studies with overall achievement measures, providing students with behavioral objectives raised their achievement scores by an average of 0.12 standard deviations. Outcomes were more positive in the 10 studies that measured relevant learning (0.25), and they were slightly negative (−0.06) in the 10 studies that measured incidental learning.

Klauer's meta-analysis (1984) covered 23 reports with 52 comparisons involving experimental groups that received objectives expressed either in behavioral terms or as learning directions or questions. Klauer found that providing such objectives before an instructional text led to some improvement

in the learning of relevant material; the average improvement in criterion scores was 0.40 standard deviations in 21 studies. He also found, however, some decline in the learning of irrelevant material; the average decline in criterion scores on irrelevant items was 0.20 in 20 studies. Finally, Klauer found that instructional objectives diminished in effectiveness as length of text increased.

Test Features

Many instructional designers believe that the effectiveness of newer instructional systems is due in large part to the way in which the systems handle evaluation of student performance and feedback to students. In computer-based teaching and mastery-learning systems, for example, students are encouraged to respond frequently while learning; their responses are evaluated very often; they get quick feedback on the results of these evaluations; and they are encouraged to continue working on learning tasks until their responses show mastery of the skill to be learned.

Researchers and evaluators have carried out many studies to determine whether this emphasis on student response pays off in better student learning. Narrative reviewers have pointed to the complexity of the results. Meta-analysts have recently begun to integrate the findings. Recent meta-analyses have examined (a) frequently of evaluation of student performance, (b) amount of feedback on performance, (c) timing of feedback, and (d) use of a mastery criterion for completion of work.

Frequency of Testing

Our meta-analysis with Bangert-Drowns showed that testing students on their progress a few times a semester increases the amount that they eventually learn and that further increases in testing produce diminishing returns (Bangert-Drowns, Kulik, & Kulik, 1988). Decreasing the number of tests below two or three produces a noticeable drop in student learning. Increasing the number of conventional tests in a course beyond two or three, however, produces only very small gains in student achievement.

The meta-analysis covered 31 studies carried out in actual classrooms. It showed that students who took at least one test during a 15-week term scored more than 0.40 standard deviations higher on criterion examinations than did students who took no tests. Effects were much smaller, however, when frequently tested students were compared to other students who were also tested but less often.

This meta-analysis also examined the effects of test frequency on student attitudes. In each of four studies of this topic, frequently tested students gave their classes higher ratings than did students who were tested less often. The average effect size was 0.59 standard deviations, a large effect.

Our conclusion from this analysis was that increasing the amount of classroom testing above current levels would not have a major impact on student learning, although it might have positive effects on student attitudes. To increase instructional effectiveness, something more is needed than a simple increase in the number of tests given to students.

Informative Test Feedback

Our meta-analysis with Bangert-Drowns on effects of test feedback suggested that increasing the amount of testing will be a more productive strategy when students are given informative feedback on their responses to test items (Bangert-Drowns, Kulik, & Kulik, 1987). When testing is coupled with informative feedback to students, testing makes a clearer contribution to student learning.

This meta-analysis covered 22 studies in which students responded to questions with and without informative feedback. An initial analysis showed that feedback had virtually no effect on student learning. The average effect attributable to feedback in all 22 studies of programmed or computer-based instruction was 0.03 standard deviations. More than a third of the 22 effect sizes were negative. That is, in 8 of the 22 studies of feedback, the group that received informative feedback actually did worse on a criterion test than did the group that received no feedback. That feedback should have either no effect or negative effects on learning seemed counterintuitive to us.

Another result that was at first difficult to understand was the discrepant findings in studies using programmed and computer-based materials. Feedback effects in studies involving computer-based instruction were positive, with informative feedback raising student performance by 0.30 standard deviations. Feedback effects in studies using programmed instruction were slightly negative, with informative feedback lowering student performance by 0.05 standard deviations.

To explain these counterintuitive results, we relied on Kulhavy's (1977) analysis of research and theory on the efficacy of feedback in written instruction. Kulhavy concluded that when learners can get feedback before they compose their own answers to questions, they often learn less efficiently. Instead of thinking through their answers, they simply "peek" at the feedback. They do not bother to attend to the instruction provided. Kulhavy called this access to answers prior to the generation of responses "presearch availability."

Our meta-analysis showed that presearch availability is a strong moderator of the effects of feedback. Overall, we found that evaluations that controlled for presearch availability had larger effect sizes than did those without such a control. When students could not peek at what was intended to be feedback, the effect size was 0.38. When students could peek at feedback frames before composing their own answers, feedback effects dropped to -0.13 standard deviations. Furthermore, almost all of the computer-based presentations but only a few programmed presentations controlled for presearch availability, producing a difference in results of the two types of studies.

Delay of Feedback

Our meta-analysis on delay of feedback showed that providing learners with feedback on the correctness of their responses works best when that feedback is provided immediately (J. Kulik & Kulik, 1988). Delayed feedback appears to be more effective than immediate feedback only in special, atypical experimental situations.

The meta-analysis covered 53 separate studies. These studies could be categorized into three types: (a) applied studies with classroom quizzes and programmed materials, (b) experiments on acquisition of test content, and (c) experiments on list learning. The studies reported a wide variety of results on feedback timing and learning

The applied studies using actual classroom quizzes and real learning materials usually reported immediate feedback to be more effective than delayed. The average effect in the 11 studies of this type was 0.28 standard deviations.

Experimental studies of acquisition of test content usually produced the opposite result, with immediate feedback virtually always proving to be inferior to delayed feedback. The average effect size in 14 studies that examined immediate learning effects was −0.36. Students who received immediate feedback performed less well than did those who received delayed feedback. Results in follow-up tests were consistent with results on the original measures of learning. The average effect size in the 8 studies that used a follow-up measure was −0.44.

The average effect size in 27 laboratory studies involving list learning was 0.34. On the average, therefore, learners who received immediate feedback learned more quickly and efficiently than did learners who received delayed feedback. The variation in results in list-learning studies was very great, however. In some studies, delayed feedback seemed more effective than immediate; in other studies, the opposite was true. The studies in which delayed feedback improved student performance were studies in which the delayed feedback was presented after the whole test and the feedback consisted of both the correct responses and the stimulus material used in the original learning situation. In such studies the delayed-feedback group received what was in effect an additional complete learning trial, whereas the immediate-feedback group did not. The same design was used in experimental studies of test content acquisition. Studies of this type provide a poor test of the effects of simple feedback delay.

The overall conclusion from this analysis was that feedback is most effective when it is presented immediately after the learner makes a response. Presentation of feedback in a separate learning trial along with a second presentation of the original stimulus material also augments learning, but this fact should not be used to suggest that delayed feedback is the most efficient type of feedback for promoting human learning.

Mastery Testing

Many reviews have reported positive results from instructional programs that include mastery testing as one feature. The reviews have not focused on mastery testing *per se*, but rather on the effects of total instructional systems, one component of which is mastery testing. Our meta-analysis showed that mastery testing may indeed be the key component in these instructional programs (C. Kulik & Kulik, 1986–87). The meta-analysis also confirmed the importance of informative feedback on tests.

The meta-analysis covered 49 comparative studies of Keller- and Bloom-type programs. The analysis showed that dropping mastery testing from these programs caused instructional effectiveness to drop substantially. The average effect size in the 49 studies was 0.54 standard deviations. That is, when mastery testing was dropped from Keller- and Bloom-type courses, performance on criterion examinations dropped by 0.54 standard deviations.

The size of the effect also depended on other factors, however. First, effects were stronger in studies with a stringent mastery criterion. Effects were especially high in studies that used a mastery criterion of 95 to 100 per cent. Second, size of effect was a function of amount of additional feedback received by the experimental group. In 28 studies in which the experimental group received more quiz feedback than did the control group, the average difference in criterion scores of experimental and control students was 0.67 standard deviations. In 21 studies in which experimental and control students received exactly the same number of quizzes and the same amount of quiz feedback, average difference in criterion scores was only 0.36 standard deviations.

This meta-analysis thus confirmed the results of our other testing analyses. It showed that frequent testing can help students learn, but that frequent testing is not sufficient in itself. For maximum effectiveness, tests must be given on which students receive feedback for the purpose of remediation of errors. Mastery testing is needed, not just frequent testing.

Table 7.2
Estimated Average Effect Sizes for Instructional Design Features on Achievement Outcomes

Instructional feature	Average *ES*
Text feature	
Advance organizers	0.30
Adjunct questions	0.55
Behavioral objectives	0.30
Test feature	
Many vs. few tests	0.15
Informative vs. no feedback	0.10
Immediate vs. delayed feedback	0.30
Mastery requirement	0.35

Conclusion

Typical effect sizes found in studies of instructional design appear in Table 7.2. The effect sizes are all positive, and almost all are at least moderate in size. A very general conclusion suggested by the meta-analytic evidence therefore is that experiments on instructional design yield evidence that may be useful in the redesign of instructional materials.

CHAPTER 8

CURRICULAR INNOVATION

In the late 1950s science and mathematics educators began a major effort to develop new curricula in science that would be more responsive to the nation's scientific and technological needs. The National Science Foundation played a major role in supporting the effort, and new curricula were soon introduced in schools in a number of scientific fields. Although the curricula varied in details, they had certain elements in common. They tended to emphasize the nature, structure, and processes of science; they tended to include laboratory activity as an integral part of the class routine; and they tended to emphasize higher cognitive skills and appreciation of science (Shymansky, Kyle, & Alport, 1983).

Educators began debating the merits of these curricula soon after they were introduced into schools during the 1960s. During the 1970s evaluators and researchers carried out individual studies to determine whether the new curricula had the desired effects on student learning. Then in the 1980s several meta-analytic reports were written to consolidate the findings from individual studies. Each of the reports had its own distinct focus, and together they give a full picture of the evaluation results.

New Math

Athappilly's meta-analysis (Athappilly, Smidchens, & Kofel, 1983) of 134 studies of *new mathematics*, or *modern mathematics*, was designed to determine whether these new curricula led to improved academic achievement and attitudes. From the 134 studies, Athappilly coded 810 effect sizes. He reported that the average effect of the curricula was to raise student achievement by 0.24 standard deviations and attitude scores by 0.12 standard deviations. It is difficult to judge the reliability of differences between subsets of effect sizes in Athappilly's study because of the inflated sample size. Nevertheless, some subset averages differ enough to be notable. Effect sizes in journal articles (average = 0.42) were higher than those in dissertations (average = 0.16). Effect sizes were also higher in locally developed tests (average = 0.42) than in commercial tests (average = 0.15).

Inquiry Curricula in Science

Bredderman's analysis (1985) covered 57 evaluation reports on three major curriculum programs in science developed for elementary schools: Elementary Science Study (ESS); Science—A Process Approach (SAPA); and the Science Curriculum Improvement Study (SCIS). Each of these programs makes extensive use of laboratory activities and gives at least as much attention to the methods of science as to its content. In his meta-analysis, Bredderman attempted to avoid problems of interdependency of multiple effect sizes by using the average effect size for each study as a dependent measure. He also carried out separate analyses for different types of outcomes: science process, science content, attitudes, creativity, and so on.

Bredderman found that the average effect size was 0.52 in 28 studies that investigated knowledge of science processes. In 14 studies that examined science content, average effect size was 0.16. In 12 studies that investigated attitudinal outcomes, average effect size was 0.27. Bredderman's analysis suggested that three study features were related to study outcomes. First, effects seemed to reflect the congruence of tests with the objectives and emphases in the classes being compared. Where tests seemed to favor the control class, effect sizes tended to be low (average = 0.13); where tests seemed to be unbiased, effect sizes were higher (average = 0.24); and where tests seemed to favor experimental groups, effect sizes were even higher (average = 0.53). Second, effects were higher in studies of disadvantaged students (average effect size = 0.65) than in studies of average or advantaged students (average effect sizes were 0.30 and 0.22, respectively). And third, effects were stronger in published studies (average effect size = 0.48) than in unpublished ones (average effect size = 0.25).

Weinstein, Boulanger, and Walberg (1982) meta-analyzed the findings from studies of innovative science curricula in high schools. They located 33 studies that evaluated a total of 13 different curricular programs. Among those most frequently evaluated were the Biological Sciences Curriculum Study project (BSCS), the Physical Science Study Committee (PSSC), the Chemical Education Materials Study (CHEMS), and the Chemical Bond Approach Project (CBA).

Weinstein *et al.* calculated a total of 151 effect sizes from the 33 studies. They reported an average effect size of 0.47 on measures of conceptual learning. Conceptual learning measures were standardized achievement tests and a few local tests specifically designed to measure comprehension, analysis, and application levels of Bloom's Taxonomy. Average effect size was 0.25 on measures of inquiry learning. Among such measures were tests of controlling variables, formulating hypotheses, critical thinking, and logical operations. Average effect size on attitudinal development was 0.26. Included in measures of attitudinal development were any measure of attitude, interest, or opinion toward science or science-related concerns.

Weinstein *et al.* did not find significant relationships between outcomes of studies and the study features that they coded. They did carry out additional analyses, however, to determine whether the positive results of their analyses

might be explained by bias in the tests used in the evaluations. They concluded that although performance does reflect test content to some extent, student achievement was superior under the innovative curricula regardless of test bias and other study variables. This conclusion is similar to the one reached by Bredderman.

Shymansky, Kyle, and Alport (1983) carried out a meta-analysis of findings from innovative science programs offered in elementary schools and high schools. Their analysis was based on 105 studies of 27 distinct programs, which yielded a total of 345 effect sizes. Included among these were 130 effect sizes that reflected the effect of the new science curriculum on student achievement. The average of these 130 effect sizes was 0.37. Other averages were 0.50 for 25 effect sizes reflecting attitudes toward science, 0.41 for 10 effect sizes reflecting attitude toward instruction, 0.17 for 28 effect sizes reflecting development of process skills, 0.25 for 35 effect sizes reflecting development of analytic skills of critical thinking and problem solving, and 0.61 for 28 effect sizes reflecting development of experimental techniques.

El-Nemr (1980) carried out a meta-analysis of findings on inquiry teaching in high school and college biology courses. He defined inquiry teaching as an approach in which students see problems posed, experiments performed, and data found and interpreted. El-Nemr calculated 113 effect sizes from 59 studies that compared inquiry and conventional curricula. Thirty-nine of these effect sizes were calculated from content exams, 30 from measures of process skills, 19 from measures of critical thinking, 7 from measures of laboratory skills, and 18 from measures of attitudes and interests. He reported that the average effect size was 0.20 for student achievement, 0.50 for process skills, 0.18 for critical thinking, 0.87 for laboratory skills, and 0.19 for attitudes and interests.

Conclusion

It is clear that evaluators have reported positive effects from efforts to revise curricular in science. Our estimates of the size of effect of these curricular revisions are as follows:

Outcome	Effect size
Content learning	0.30
Learning of science processes (e.g., process skills, science methods)	0.40
Development of analytic skills (e.g., critical thinking, problem solving)	0.25
Development of lab skills and techniques	0.75
Development of positive attitudes toward subject, science, and methods	0.25

These revised curricular have therefore produced moderate effects on student performance on content examinations, on tests of higher-order thinking skills, and on measures of knowledge of science processes, but the curricula have had strong effects on the development of laboratory skills. Finally, these new curricula have had beneficial effects on students attitudes toward science.

Meta-analysts who have examined these results, however, have raised a question about the fairness of some of the criterion measures when applied to conventional curricula. Most conventional curricula in science, for example, are not geared toward development of laboratory skills. It is not surprising therefore that conventional curricula appear to do a poor job when judged by their contributions to laboratory skills. Other criterion measures used in curricular evaluations probably provide a fairer test of the effectiveness of the curricula.

CHAPTER 9

TEACHER EDUCATION AND EVALUATION

Helping teachers to develop their talents to the fullest has long been an important concern in both precollege and postsecondary education. Elementary and high school teachers usually engage in both preservice and inservice activities in their efforts to develop, maintain, and improve teaching skills. College teachers often turn to faculty development centers for help in improving their teaching. An important service offered by such centers is a formal program of teaching evaluation. The results of such evaluations are used by teachers as well as by administrators in improvement efforts.

Research findings on the effectiveness of teacher education and evaluation programs have been examined by meta-analysts in recent years. Their concerns have been several. Do teacher training programs actually help teachers do a better job? Do evaluation programs select those teachers who get across the most to students? Can student ratings be used by teachers to improve their teaching?

Teacher Education

Teacher education is a broad term that refers to preservice activities that individuals engage in before they become teachers and inservice activities carried out by teachers during their careers for the sake of professional improvement. Teacher education is thus a massive enterprise, offered by thousands of education professors, supervisors, and consultants each year for the sake of tens of thousands of individuals.

Wade (1984) carried out a large-scale meta-analysis on effects of inservice teacher training programs. The analysis covered four types of programs outcomes: (a) teacher reactions or attitudes toward the program; (b) teacher learning of what was taught in the program; (c) teacher behavior change as a result of the program; and (d) results of teacher changes as indicated by student learning. Included in the analysis were 715 effect sizes reported in 91 studies. Wade found an average effect size of 0.42 in 233 comparisons of teacher reactions or attitudes, an average effect size of 0.90 in 52 comparisons

of teacher learning, and an average effect size of 0.60 in 298 comparisons of teacher behavior. Wade also reported an increase in student test performance of 0.37 standard deviations in 132 evaluations of the effects of inservice teacher training. Wade reported that a variety of study features affected outcomes, but the reliability of the reported relationships is hard to assess because of the inflated sample sizes used in her analyses of features and outcomes.

Enz, Horak, and Blecha (1982), Sweitzer and Anderson (1983), and Yeany and Porter (1982) carried out independent meta-analyses of findings from inservice training programs for science teachers. Each of the research teams produced findings that were consistent with Wade's. Each supported the conclusion that inservice training programs had notable effects on the attitudes and behaviors of science teachers.

Sweitzer and Anderson (1983) located 68 studies in which science teachers were trained to use inquiry methods in teaching. The studies yielded 177 effect sizes. For teacher reactions, or attitude change, the average of 31 effect sizes was 0.47. For teacher learning, the average of 55 effect sizes was 0.80. For teacher behavior, the average of 60 effect sizes was 0.82. Sweitzer and Anderson also found an average of 0.67 in 19 effect sizes measuring student outcomes. It is hard to judge the reliability of reported relationships between features and outcomes in this meta-analysis because of the inflated sample sizes. It is notable, however, that Sweitzer and Anderson reported a clear difference in average effects for journal articles and dissertations. Average effect size for journals was 1.01; average effect calculated from dissertations was 0.59.

Yeany and Porter (1982) examined results in studies that were similar to those reviewed by Sweitzer and Anderson. In a typical study, teachers learned to describe and analyze teaching processes. They learned the fundamentals of specific systems for coding teaching processes, and they also learned to apply these coding systems to their own classes, to model classes, etc. Yeany and Porter do not specify the number of studies included in their meta-analysis; they list only 12 studies, however, in their bibliography. From their pool of studies, Yeany and Porter found 188 effect sizes. The average effect size on student behavior was 1.31.

Enz, Horak, and Blecha (1982) reviewed results from 16 inservice programs for science teachers. Too few studies contained findings on student attitudes and knowledge to yield reliable conclusions; only two studies measured each of these outcomes. Eleven studies, however, examined effects on teacher knowledge, and these studies yielded an average effect size of approximately 0.80.

The meta-analyses by Malone (1984) and Redfield and Rousseau (1981) examined results of training programs for teachers in a variety of areas. Redfield and Rousseau conducted a meta-analysis covering 14 studies, 6 of which involved training teachers to use higher-order questions in teaching. The training program had dramatic results on student achievement. Students of teachers who were taught to use higher-order questions performed at a level

that was 0.83 standard deviations higher than the level of students in control classes. Malone (1984) examined effects of preservice educational programs of a variety of types. Malone located 45 studies that yielded 145 effect sizes. The average of 30 effect sizes on teacher reactions, or attitudes, was 0.13; average of 12 effect sizes on teacher knowledge was 0.03; and the average of 10 effect sizes for teaching behavior was 0.11.

In summary, the difference in outcomes of preservice and inservice programs is notable. The preservice programs examined by meta-analysis to date have not had an important effect on teacher knowledge, attitudes, or behavior. The inservice programs have had clear effects on these same outcome measures. In addition, their effect on student learning has been clearly demonstrated—perhaps the ultimate testimony on the value of an educational program.

Evaluation

Like precollege teachers, college teachers engage in many activities that contribute to their professional development. They attend workshops, seminars, and conventions, and they pursue independent courses of study. Researchers have not often examined the impact of such professional development activities, however, because faculty members usually carry them out independently and informally. Education programs for college faculty would be difficult to study experimentally or meta-analytically.

Researchers have focused their attention instead on a different aspect of faculty development programs: teacher evaluation. Systematic examination of faculty evaluation programs is possible because such programs are usually formal and highly organized. Because of the importance of teacher evaluation, both correlational and experimental studies on the topic are frequently found in the literature.

Student Ratings and Student Learning

The most common type of teacher evaluation program in higher education is based on student ratings. Teachers, administrators, and researchers still argue about the validity of such ratings. Do they actually measure teaching effectiveness, which is usually thought of as the degree to which instructors facilitate student learning? Or do they measure something else, such as personality or entertainment value? Meta-analysts have synthesized the findings on teacher evaluation programs in recent years and provided some answers to this question.

Cohen (1981) carried out an analysis of results from 41 independent validity studies reporting on 68 investigations. Each of the 68 investigations took place in a different multi-section college course, and each investigation focused on the relationship between mean class rating and mean class

achievement. That is, the class was the unit of analysis in each of the investigations reviewed by Cohen.

Cohen found that the average correlation between overall instructor rating and student achievement was .43; the average correlation between overall course rating and student achievement was .47. Cohen noted that the correlations between ratings and achievement scores were even higher when students provided ratings after they received their course grades, when they were rating full-time faculty members rather than teaching assistants, and when examinations for all sections of a course were scored uniformly by someone other than the section teachers.

Overall, the validity coefficients that Cohen found for student ratings were of a magnitude that is usually considered to be high for applied studies in education and psychology. Cohen's analysis can therefore be taken to provide strong support for the validity of student ratings of measures of teaching effectiveness.

Student Ratings and Teacher Improvement

Student ratings of instruction are often collected in colleges and other educational institutions with the expectation that the ratings will provide helpful feedback for teachers to use in improving their teaching. Whether or not student ratings actually are helpful in improving teaching has been a subject of controversy among teachers, administrators, and students. Evaluation studies show, however, that feedback from student ratings leads to improved teacher performance.

A meta-analysis by Cohen (1980) covered 22 studies of the effectiveness of feedback from student ratings in college courses. End-of-course ratings for teachers who received midsemester feedback were 0.33 standard deviations higher than end-of-course ratings of control teachers. The effects of midsemester feedback were stronger when instructors discussed the feedback with consultants or received other help in interpreting and using it. Other study features sometimes thought to be important determinants of feedback effectiveness, such as the length of time available to implement changes in a course and the use of normative data in interpreting feedback, were unrelated to effect sizes in the studies.

Conclusion

Our estimates of the average effect of inservice teacher training are given in Table 9.1. The effects seem to us to be remarkably strong. They show that teachers learn what is taught in inservice programs; they change in attitudes; and they change their classroom behavior as a result of the programs. As a further consequence of the programs, students of teachers who have received inservice training learn more in the classroom. These results stand in striking contrast to meta-analytic findings on programs of preservice training. Teacher

participation in preservice programs seems to have little effect on either the teachers who receive the training or the students assigned to classrooms of such teachers.

Table 9.1
Estimated Average Effect Sizes for Inservice Education Programs

Outcome	Average *ES*
Teacher learning	0.60
Teacher behavior	0.50
Teacher attitudes	0.35
Student learning	0.55

Meta-analysts have also suggested that student evaluation of teachers can be useful in improving college teaching. Their results show that student evaluations can be used to pick out teachers who are effective in raising student performance. They also show that teachers often take student rating advice to heart and change their classroom behavior as a result of the student ratings.

CHAPTER 10

CLASS AND SCHOOL ORGANIZATION

Schools have been organized in the same way for many decades. Schools are divided into grades, and grades are usually divided into classrooms. Pupils are almost always assigned to grades on the basis of age. Classrooms are usually heterogeneous with respect to student ability in elementary schools, but they are usually fairly homogeneous in high schools, where students move in defined tracks (e.g., academic, general, vocational).

Although the patterns of school organization have changed little over the years, they have never ceased to be controversial, and educators have long searched for alternative ways to organize schools and classes. Meta-analysts have recently begun to review studies on school and classroom organization in order to help educators in their search.

Class Size

Research on class size has taken several different shapes over the years (Glass & Smith, 1979). Before 1920, most studies of class size were pre-experimental in design. Between 1920 and 1940, true but fairly primitive experimental studies were conducted. Studies of class size were dominated by a concern with large-group technology in the 1950s and 1960s and by a concern with individualized instruction after 1970. By the late 1970s, however, most educational researchers had concluded that the years of study of class size had produced only inconclusive results. By the end of the decade, class size had ceased to be a vital research area in education.

Meta-analytic Findings

Glass and Smith's (1979) meta-analysis of findings on class size rekindled interest in the area. Where others had seen only confusing irregularity, Glass and Smith saw clear relationships. They argued that research results on class size had been misconstrued by earlier reviewers.

Examination performance. Glass and Smith reported that the average difference in student achievement in small vs. large classes amounted to 0.09 standard deviations; results favored smaller classes by less than a tenth of a standard deviation. By most criteria, this effect would be considered trivial in size. Many meta-analysts would conclude from it that class size does not play an important direct role in determining student achievement.

Glass and Smith argued, however, that this overall average does not give a good picture of the importance of class size because it disregards the exact sizes of the small and large classes being compared. What is counted as a small class in some studies is considered a large class in others. Glass and Smith further argued that the relationship between class size and student achievement must be logarithmic. Achievement should increase as class sizes increase, but the size of the increments should diminish. The increase should be large when a class of 20 students is increased by 20; it should be much smaller when a class of 200 is increased by 20.

Glass and Smith also thought that the lack of quality of many studies of class size also led to underestimation of effects. Effects seemed to be stronger in studies of high quality and smaller in low quality studies. Effects were also higher in studies that appeared in books and journal articles than in unpublished reports and dissertations. When Glass and Smith fitted a logarithmic curve to the data from what appeared to be the best studies of class size, they found that results were clear and compelling. If students achieved at the 50th percentile in a class of 30, they would achieve at the 54th percentile in a class of 20; at the 57th percentile in a class of 15; and at the 61st percentile in a class of 10.

Attitudes. Smith and Glass (1980) used 59 studies from the same pool to determine how class size influenced student attitudes and feelings. They found that the effect of class size on attitudes was much larger in the typical study. Attitudes were more favorable in small classes than in large classes, and the superiority in attitude scores amounted to nearly one-half standard deviation.

Smith and Glass once again reported a good fit of a logarithmic curve to class size data. Attitudes were most favorable in the smallest classes. Increasing class sizes by a constant amount (e.g., 5 students) made more of a difference in small classes than it did in large ones. In one respect, this attitudinal analysis produced strikingly different results from the achievement analysis. In the achievement analysis, effects were clearer in better designed studies; in the attitudinal analysis, effects were less clear in such studies. Glass and Smith used the whole pool of 59 studies, however, to draw conclusions about attitudinal effects.

Response

Hedges and Stock (1983) reanalyzed Glass and Smith's class size data, using Hedges's adjusted effect sizes and a different method of estimating regression weights. Their overall results were very similar to those reported

by Glass and Smith. Hedges and Stock concluded that their reanalyses did not suggest substantial changes were needed in the conclusions originally drawn by Glass and Smith. Smaller classes still led to higher achievement than did large classes, and Hedges and Stock's statistical analysis indicated that class size accounted for a substantial amount, but not all, of the achievement variation among classes.

Educational Research Services (1980) and Slavin (1984), however, raised questions about Glass and Smith's analysis and interpretation of data. Among their criticisms were the following:

1. Glass and Smith used inconsistent standards in their analyses, and the inconsistencies suggest analyst bias. In their achievement analyses, for example, Glass and Smith based their conclusions primarily on well-controlled studies. Discounting results from less-controlled studies made effects seem larger because such studies reported smaller effects. In their attitudinal analyses, however, Glass and Smith based their conclusions on all studies. Discounting results from less-controlled studies would have made effects seem smaller because studies with few controls yielded the larger effects in the attitudinal area.
2. Glass and Smith's positive findings were influenced heavily by some studies of marginal relevance to the issue of class size and achievement. Slavin pointed out, for example, that Glass and Smith's logarithmic curve was influenced very heavily by one study in which the criterion of achievement was accuracy of throwing a tennis ball against a wall. Without this questionable study the fit of a logarithmic curve to class size data would be far poorer.
3. Glass and Smith's demonstration of the importance of class size is irrelevant to the variations in class size that are the concern of most teachers and administrators—variations from, say, 20 to 40. No evidence has been presented to show that variations in this range have important effects on student achievement.

Glass has responded to such criticisms but the response has not convinced his critics. It seems unlikely that Glass and Smith's conclusions about the importance of class size as a variable in schooling will be accepted until their findings are confirmed in other independent analyses carried out by other analysts using their own criteria for their literature searches, their own analytic strategies, etc.

Comprehensive Grouping Programs

The practice of grouping students for instruction by their academic aptitude remains as controversial today as it was when grouping was first introduced into American education at the turn of the century. Although many educators favor the practice of ability grouping, others condemn it as undemocratic. They

say that it conveys unwarranted status and prestige on pupils in in-groups while unfairly stigmatizing members of out-groups.

Educational researchers have long known that *grouping* is a loose term that can be applied to a great many educational practices. Reviewers of research therefore usually distinguish among several different categories of grouping programs: between-class and within-class programs, comprehensive programs and programs for special populations, single-class and all-day programs, and so on.

This section focuses on one major category of grouping, the comprehensive program. Within this category, we distinguish among three different approaches: between-class, within-class, and Joplin plan programs.

Between-Class Programs

These are comprehensive programs in which learners of a full range of ability are separated into classes on the basis of measured or judged aptitude and then are taught in separate classrooms by different teachers using the same or similar curricular materials. Evaluations show that on the average such programs have trivially small effects on the achievement of students and that the effects are approximately equal for students in high, middle, and low aptitude groups. Evaluation studies also show that effects of grouping on self-esteem vary according to aptitude group into which learners are placed. Learners placed into the low aptitude classes tend to show modest increases in self-esteem; learners placed into middle and high aptitude classes tend to show small decreases in self-esteem.

Examination scores. Our meta-analysis on between-class grouping at the secondary school level examined results from 33 studies of inter-class grouping of secondary students (C. Kulik & Kulik, 1982). We reported that the average effect of grouping on examination scores was trivially small—an increase in examination scores of 0.02 standard deviations—and that effects were similar at each ability level. Our later meta-analysis (C. Kulik & Kulik, 1984) of 19 studies at the elementary school level produced similar results: an overall average effect size of .07 standard deviations. We recently updated our reviews of elementary and secondary school studies and reported combined results (J. Kulik & Kulik, 1987). The average effect of grouping in 49 studies was to raise student performance by 0.06 standard deviations. Effects were 0.12 for students in high aptitude classes, 0.04 for students in middle level classes, and 0.00 for students in low ability classes.

Slavin's (1986) best-evidence synthesis reported results that were for the most part in substantial agreement with our findings. Slavin reported finding little evidence in his analysis of results to support the adoption of comprehensive, full-day grouping of pupils into different classes on the basis of ability. Average effect size in 14 studies of such programs was 0.00. But Slavin also reported that he found evidence of positive benefits from programs in which grouping was restricted to a single subject. Average effect size in 7

studies of such programs was 0.18. With larger samples of grouping studies, however, we were not able to find a reliable difference in results from programs of multiple- and single-subject grouping (J. Kulik & Kulik, 1987). We found the average effect size in 20 studies of inter-class grouping in a single subject to be 0.09, not significantly different from the average effect size of 0.03 in 29 studies of comprehensive, full-day programs.

Self-esteem. We reviewed 15 studies of the effects of grouping on student self-esteem (C. Kulik, 1985). We reported that the average effect was near-zero. That is, the average study reported that students from grouped and ungrouped classes had virtually identical self-esteem scores. We did not give much weight to this overall average, however, because we also found evidence that self-esteem effects differed according to group to which students were assigned. We reported a tendency for effects to be positive on low ability groups and negative on middle and high ability groups.

Our 1982 meta-analysis on secondary school findings on grouping also examined a small number of studies investigating effects of between-class grouping on attitudes toward school and attitudes toward school subjects. We found an average effect size of 0.37 in 8 studies of attitudes toward subject matter and an average effect size of 0.09 in 11 studies of attitudes toward school.

Within-Class Grouping

Comprehensive programs of within-class grouping, or intra-class grouping, are programs in which students of a full range of ability are separated into groups within a classroom and are then taught the same or similar curricular materials but in separate groups. Evaluations show that, like inter-class grouping, within-class grouping has trivial to small effects on student learning.

Slavin and Karweit (1984) carried out a small quantitative synthesis of findings on within-class grouping programs. On the basis of the results in eight studies, Slavin and Karweit concluded that within-class ability grouping is clearly beneficial for learners. They reported that it raised examination scores by 0.55 standard deviations on the average. We have recently criticized Slavin and Karweit's calculation of effects, however (J. Kulik & Kulik, 1987). We pointed out that some of the effect sizes reported by Slavin and Karweit (1984) were calculated by dividing treatment effects by residualized standard deviations rather than raw standard deviations. This calculating procedure produces inflated effect sizes. Slavin's average effect size of 0.55 therefore almost certainly exaggerates the size and importance of achievement effects of programs of within-class grouping.

Slavin apparently recognized the problem too. His recent review of studies of within-class grouping (Slavin, 1987) covers results from eight studies in elementary schools. In this report effect size calculations are accurate, and the average effect size is 0.34. But this figure too may overestimate the size and importance of effects from programs of within-class grouping. In our most

recent meta-analysis on grouping, we also reviewed findings from studies of within-class programs at the elementary and secondary school level (J. Kulik & Kulik, 1987). Our pool of 15 studies was twice as large as Slavin's pool. We found an increase in examination scores of only 0.17 standard deviations due to within-class grouping.

Joplin Plan

In Joplin plan programs learners are placed by aptitude rather than age into appropriate grades for instruction in a single subject, usually reading. Like programs of between-class and within-class grouping, these Joplin programs have trivial or small effects on student learning.

Slavin's best-evidence synthesis (1987) reviewed findings from 14 studies of the Joplin plan. Slavin reported that the average effect of the plan was to increase student achievement by 0.45 standard deviations. This average, however, may exaggerate the size and importance of effects that can be obtained from the Joplin grouping plan. Among the 14 studies that Slavin reviewed were a few in which the treatment does not appear to be the Joplin plan. Slavin's review also failed to include a few studies in which the Joplin plan produced strong negative effects. Our most recent meta-analysis on grouping covered 14 studies of the Joplin plan and reported an average effect size of 0.23 standard deviations (J. Kulik & Kulik, 1987).

Special Grouping of Talented Learners

Another major type of grouping program is designed to meet the special needs of children at only one part of the ability range. Grouping talented learners together for special instruction has been especially popular, and the effectiveness of this type of grouping program has been studied very often. Programs designed especially for talented and gifted students clearly differ from other grouping programs in the amount of curricular adjustment that they entail. Sometimes programs for the gifted and talented are labeled *accelerated programs*, but even when they are not, expectations are usually clear that the gifted pupils in the programs will cover course content in much greater depth if not at a faster pace than will pupils in regular classes.

Acceleration

Programs of accelerated instruction are programs in which students judged to be of high aptitude, gifted, or talented move through a curriculum at a more rapid pace than do other students, so that accelerated students are able to complete the curriculum at an early age or to take advanced courses not ordinarily taken by students of their age. Programs of accelerated instruction usually have very strong positive effects on the achievement of high-aptitude

students. Evaluation studies have not disclosed consistent nonintellective effects of acceleration.

We recently reviewed 21 reports evaluating the effectiveness of programs of accelerated instruction (J. Kulik & Kulik, 1984). The accelerated programs covered in our review were of three major kinds: grade skipping; compressing a curriculum for talented students (e.g., 4 years in 3); and extending the calendar to speed up the progress of talented students (e.g., completing the work of 4 years in 3 school years with five summer sessions). We concluded on the basis of 26 studies described in the papers that programs of accelerated instruction typically have strong positive effects on student achievement scores. Our analysis showed that examination performance of accelerates surpassed by nearly one grade level the performance of nonaccelerates of equivalent age and intelligence. Average effect size for programs of acceleration was 0.88. In addition, examination scores of accelerates were equivalent to those of same-grade but older, talented nonaccelerates. Nonintellective outcomes were not often investigated in the 26 studies, and the few results available on nonintellective outcomes were not consistent from study to study.

Between-Class Grouping

Gifted and talented students are sometimes grouped together in separate classes where they are taught with adjusted but not accelerated curricular materials. These programs have moderate positive effects on student achievement and have little effect on student self-esteem.

Our meta-analysis of ability grouping at the secondary level examined results from 14 studies of programs designed especially to enrich the education of talented and gifted secondary school students (C. Kulik & Kulik, 1982). We found that such students learned more in these special honors programs than did control students who were instructed in mixed-ability classes. The average effect of such special classes was to raise student achievement by 0.33 standard deviations. Our later meta-analysis of 9 studies at the elementary school level produced similar results: an average effect size of 0.49 standard deviations (C. Kulik & Kulik, 1984). We recently updated our reviews of elementary and secondary school studies and reported results from a combined analysis of 25 studies in elementary and secondary schools (J. Kulik & Kulik, 1987). The average effect of separate classes in the 25 studies was an increase in examination scores of high aptitude learners of 0.33 standard deviations.

We also examined the effects of special classes on the self-esteem of high aptitude learners (C. Kulik, 1985). We concluded on the basis of results in six studies that grouping programs designed especially for talented students had a trivial effect on the self-esteem of these students. Average effect in the six studies that we located was an increase in self-esteem scores of only 0.02 standard deviations.

Within-Class Grouping

Gifted and talented students are sometimes separated into small groups in otherwise heterogenous classrooms. They are then taught with adjusted but not accelerated curricular materials. These programs have moderate positive effects on student achievement.

For our most recent meta-analysis on grouping effects, we were able to locate only four studies in which talented students were taught as a separate group in the same classrooms as other students (J. Kulik & Kulik, 1987). In each of the four studies, the talented students taught in a separate group received the higher examination scores, and in three of the four studies, the difference in examination scores of students taught as a separate group and control students was statistically significant. The average effect size in the four studies was 0.62. The four reports thus seemed to provide fairly conclusive evidence on the effectiveness of within-class grouping programs designed especially for high aptitude learners.

Conclusion

Glass and Smith's meta-analyses have demonstrated that class size matters, although not an enormous amount and probably not in the way that many people think. Student learning is fairly similar in large classes of 40 or more and in small classes of 15 to 20 students. Although differences in learning may be greater with more dramatic decreases in class size (say, to 5 students), the evidence is still not conclusive on learning in extremely small classes. The evidence on the relationship between class size and student attitudes seems clear, however. Attitudes are much better in smaller classes.

Evidence on effects of grouping programs is summarized in Table 10.1. Several conclusions can be drawn. First, the strongest and clearest effects of grouping are in programs designed especially for talented students. Talented students gain more from such programs than they do from heterogeneous classes. Programs of acceleration have the strongest positive effects on students because curricula in such programs differ most from those followed by nonaccelerated students. Special within-class and between-class grouping programs designed especially for the benefit of talented students also produce positive effects on these youngsters. Separating talented students into homogeneous groups apparently enables teachers to provide learning opportunities that would not be available with heterogeneous grouping.

Programs that are designed for all students in a grade—not solely for the benefit of talented students—have significantly lower effects. Comprehensive between-class grouping raises overall achievement levels by approximately 0.05 standard deviations, a very small amount. Comprehensive within-class programs raise overall achievement by 0.15 standard deviations, and Joplin plans raise achievement by about 0.25 standard deviations. The gains from the three types of programs are not significantly different from one another,

Table 10.1
Estimated Average Effects of Grouping Programs on Three Major Outcome Areas

Program	Examination	Self-esteem	Attitude toward instruction	Attitude toward subject
		Comprehensive Grouping		
Between-class	0.05	0.00	0.10	0.35
High aptitude	0.10	−0.15		
Middle aptitude	0.05	−0.15		
Low aptitude	0.00	0.15		
Within-class	0.15			
Joplin plan	0.25			
		Special Grouping of Talented Learners		
Between-class	0.35	0.00		
Within-class	0.60			
Acceleration	0.90			

but they are large enough to be considered statistically different from a zero gain.

Critics have charged that grouping programs can have devastating effects on the self-esteem of slower students. The meta-analytic findings do not support this charge. If anything, grouping has slight positive effects on the self-esteem of slow students and slightly negative effects on the self-esteem of bright students. Talented students grow less smug when they are taught with their intellectual peers; slower students gain in self-confidence when they are taught with other slow learners.

CHAPTER 11

EQUITY

It is not necessary to document here that gender and race inequities exist. Government statistics on employment and income are widely reported in the popular press, and the message from such statistics is impossible to ignore. The disparities in opportunities between sexes and among races are great, and progress in closing these gaps has been painfully slow.

But what of the schools? Do gender and race inequities exist there? Or do children of both sexes and all races and ethnic backgrounds have the same opportunities to learn? What can schools do to promote equity? Meta-analysts have helped answer these questions in at least two different ways. First, they have integrated findings from primary studies that report on the relationship between school achievement and such factors as gender and race. Second, they have examined the effectiveness of efforts to increase equity in education.

Gender, Race, and School Achievement

Many of the topics that meta-analysts examine are broad and loosely defined. Meta-analysts therefore often struggle to find a pool of studies to analyze. There are usually many decisions to make. Is computer-based instruction in this study the same thing as computer-based instruction in that study? Should this be considered a study of ability grouping, or is it something else? The areas of gender and race do not present such definitional problems. We might therefore expect more uniformity in results in these areas.

Gender effects

Fleming and Malone's (1983) meta-analysis of findings on gender and science achievement and attitudes covered 122 dissertations, 41 journal articles, 5 fugitive documents, and results from the 1978 National Assessment of Educational Progress. These sources yielded 141 comparisons of boys and girls on science knowledge and understanding and 31 comparisons on science attitudes. Boys scored slightly higher on the average than did girls on science

knowledge and understanding (average difference = 0.14 standard deviations) and on science attitudes (average difference = 0.08 standard deviations).

Although Fleming and Malone did not use inferential statistics in their analysis, they noted that gender differences in science knowledge and understanding were smaller at the elementary level (average difference = 0.05 standard deviations) than at the high school level (average difference = 0.16 standard deviations). Gender differences in science attitudes also seemed to be a function of grade level, but differences in attitudes were more pronounced at the elementary level (average difference = 0.18 standard deviations) than at the high school level (average difference = 0.05 standard deviations).

Steinkamp and Maehr's (1983) meta-analysis also examined the relationships among gender, science attitudes, and science achievement and understanding. Their analysis was based on 255 correlations reported in 66 articles and reports. Steinkamp and Maehr reported average values for subsets of the 255 correlation coefficients. These average correlations can be easily converted to standardized mean differences between sexes. For science knowledge and understanding, Steinkamp and Maehr found 86 correlations with an average of .12, equivalent to a standardized mean difference of 0.24. They found no evidence to suggest that the gender difference decreased or increased as a function of age. The average difference in 35 elementary school comparisons was 0.30; the average difference in 38 junior and senior high school comparisons was 0.25 standard deviations.

Steinkamp and Maehr found that boys and girls differed only slightly in science attitudes. Average overall difference in attitude scores was only 0.06 standard deviations. These analysts also found no relation between grade level and size of gender difference in attitudes. The average difference between boys and girls in favorability of attitudes toward science was 0.10 standard deviations in 6 comparisons in elementary schools; the average difference was 0.06 in high schools.

Freeman's meta-analysis (1984) focused on gender differences in mathematics achievement. She coded one effect size from each of 35 studies so that inflation of sample sizes was not a problem in her statistical analysis. She found an average effect size of 0.09 in the 35 studies, indicating that achievement of males was approximately one-tenth standard deviation higher than that of females. Grade level was the only study feature related to effect size in Freeman's analysis. Average effect size was −0.19 in 7 elementary studies, but it was 0.16 in 22 high school studies and 0.19 in 6 college studies.

Race differences

Fleming and Malone also examined differences in science achievement by race and ethnic group. Their comparisons involved three groups: Anglos, Blacks, and Hispanics. Fleming and Malone reported that 49 comparisons of Anglo and Black achievement in science in elementary and high schools yielded an average difference of 0.42, favoring Anglo students. The 46 comparisons of

Anglo and Hispanic achievement yielded an average effect size of 0.31, again favoring the Anglo students.

Fleming and Malone reported that these racial and ethnic differences changed very little with grade level. Differences between Anglos and Blacks on measures of knowing and understanding science approximated 0.40 standard deviations in elementary schools, middle schools, and high schools. Differences between Anglos and Hispanics were also fairly constant through the elementary, middle, and high school years.

Finally, Fleming and Malone found only small differences among racial and ethnic groups in attitudes toward science. Anglos scored approximately 0.10 standard deviations higher than did Blacks in 11 comparisons on favorability of attitudes toward science. Anglos were also slightly more favorable toward science than Hispanics were. In 11 comparisons, scores of Anglos were 0.05 standard deviations higher than the scores of Hispanics.

Increasing Opportunity and Achievement

Efforts to increase educational opportunities and to decrease inequities are not new in education, nor is there anything new about evaluation of such efforts. For many years evaluators have been trying to determine the success of school efforts to decrease the gap between students from minority and other backgrounds. Their evaluations range from single studies of opportunity programs at specific colleges to evaluations of desegregation efforts in school systems. What is new in education is the meta-analytic attempt to quantify the evaluation results.

College Programs for High-Risk Students

Our meta-analysis with Shwalb (C. Kulik, Kulik, & Shwalb, 1983) examined effects of four types of programs designed to increase achievement of high-risk students in colleges: programs of instruction in academic skills; guidance sessions; comprehensive support programs; and programs of remedial or developmental studies. The analysis covered findings from 60 separate evaluation studies and focused on two program outcomes: achievement and persistence in college. Achievement was measured by grade-point average in regular college courses. Grades obtained in special courses designed for high-risk students were not used in calculating grade-point averages. Persistence was measured by the proportion of students initially admitted who remained enrolled in college during the period studied in an evaluation. In a few studies this evaluation period was the same as the treatment period—usually one or two semesters—but in most studies the evaluation period extended well beyond the period of student enrollment in the program.

The analysis showed that programs for high-risk students had positive effects on students. This generalization held true when program effects were measured by increases in grade-point averages and when they were measured

by increased persistence in college. The generalization also held true for different types of programs for high-risk college students: reading and study skills courses, guidance sessions, and comprehensive support services. Such programs raised student grade-point averages by 0.25 to 0.40 standard deviations (average effect size = 0.27). The special programs also raised persistence rates from 52 per cent remaining in college to 60 percent remaining (average effect size = 0.27).

We also found significant relationships between several study features and study effects. The most notable of these relationships, found in both the analysis of grade-point average and the analysis of persistence effects, was between type of college and effects. Effects on both grade and persistence rates were lower at community colleges. It may be that grade and persistence rates provide inadequate indices of program effects at community colleges, but it may also be that the programs developed for high-risk student are insufficiently challenging to prepare them well for regular college courses.

Bilingual education

Willig (1985) carried out a meta-analysis of findings from 16 studies of bilingual education. The studies yielded a total of 513 effect sizes coded from 75 independent samples. A total of 466 of these effect sizes came from comparisons of bilingual experimental groups and traditional all-English or submersion groups.

Willig found an unadjusted average effect size of 0.09 standard deviations for these 466 comparisons. An effect of this magnitude would ordinarily be taken to indicate little if any systematic effect from bilingual education. This conclusion would be similar to that reached in major narrative reviews on effectiveness of bilingual education. Willig concluded, however, that when statistical controls for methodological inadequacies of studies are applied, the positive effects of bilingual education are clear. She reported an adjusted mean effect size of 0.63 for the 466 comparisons.

Willig's adjustment was made by removing the effect of six other variables from the average effect size of 0.09. The six variables were academic domain of criterion (e.g., total language, oral production, reading, math, etc.); language of the criterion examination (English or non-English); assignment to programs; effect size formula used; type of score on the criterion examination (e.g., raw score, grade equivalent, percentile, etc.); and a variable representing the interaction effects of the language and domain of the criterion test. Willig reported that these six variables together accounted for 63% of the total variance in effect sizes.

We do not believe that Willig's adjusted means give a fair picture of the effects of bilingual education. First, Willig based her adjustment on a correlation matrix formed from 466 observations on at least 183 variables. The number of independent observations in Willig's data set is 16, the number of different bilingual programs studied and the number of reports available to Willig. Willig's extravagant coding of effect sizes and study features makes it

impossible to carry out accurate statistical tests or to estimate the true degree of correlation among study features. Willig's transformation of an average effect size of 0.09 to an adjusted effect size of 0.63 not only seems improbable on the face of it, but the transformation does not stand up to close scrutiny.

Other results from Willig's analysis are also puzzling. Willig reports an average unadjusted effect size of −0.46 for the 36 cases in which comparisons were based on percentile scores but an average unadjusted effect size of 0.25 for the 254 cases in which comparisons were based on raw scores. Transforming raw scores to percentile scores should leave effect sizes unchanged. The substantial difference between these two averages is therefore puzzling. One plausible explanation of this otherwise inexplicable finding is that all the percentile scores came from one or possibly two programs that produced atypical results. The reliability of the difference cannot be tested with sample sizes of 36 and 254.

Willig also reports different results for different effect size formulas. The finding is odd because most of the formulas for effect size are simply algebraic variations on a basic formula. The true significance of Willig's report becomes clear when one notices the formulas involved. Willig reports an average unadjusted mean effect size of 0.06 when the difference in gain scores is divided by the raw-score standard deviation; she reports an unadjusted mean effect size of 0.44 when the difference in gain scores is divided by the standard deviation of the gain scores. As we pointed out in Chapter 4, it is an error to standardize treatment effects on gain-score standard deviations. Willig's table shows that at least some of the effect sizes used in her analysis were incorrectly calculated.

The problems in Willig's analyses seem to be serious ones. In light of these problems, we think that it is inappropriate to take Willig's adjusted means as an indicator of the effectiveness of bilingual education programs. Even the unadjusted mean that she reports for such programs (0.09) may reflect an upper bound on program effectiveness.

Programs of Desegregation

The effects of desegregation in schools have been studied time and time again. A review in the late 1970s estimated that as many as 200 field studies of desegregation may be available in the U. S. even after inadequate studies and those which duplicate other reports are eliminated (Crain & Mahard, 1978). The studies were carried out in a variety of locations; in the first year and after several years of desegregation; and in cities where desegregation was achieved through busing, through closing of segregated schools, and through redistricting. Studies used cross-sectional designs, longitudinal designs, and combinations of both.

Krol (1978) was the first to use meta-analytic techniques to synthesize the findings from these studies. His analysis covered results from 54 studies of black achievement. Unfortunately, a good number of the studies included in Krol's review did not have a control group. A small positive effect on black

achievement (effect size = 0.11 standard deviations) was found for the 38 comparisons that did include a control group. Krol also investigated the influence on study outcome of grade level, length of desegregation, and curricular area being investigated. He concluded that none of these factors influenced the results of desegregation studies.

Crain and Mahard (1982) reviewed 93 documents that reported a total of 323 findings on Black achievement in desegregated and segregated schools. Crain and Mahard were able to code effect sizes from 268 of these comparisons, and they used 264 of these in their calculation of an average effect size. This overall average was 0.23, a value that was similar to that reported by Krol.

Crain and Mahard noted that the 264 effect sizes varied a good deal, and they concluded that the variation was largely a result of two factors: (a) grade of children when desegregation began; and (b) adequacy of control group. These analysts noted, for example, that desegregation effects seemed largest when students were in desegregated schools from the kindergarten years. Most studies, however, were done during the first year of desegregation when students were in middle and late elementary school. They also noted that many studies had no adequate control group.

When Crain and Mahard restricted their analysis to 45 comparisons that seemed to them to be most free of problems, they found that findings were positive in 40 of the 45 cases. They also found an overall average effect of desegregation in these comparisons of about 0.3 standard deviations. They concluded that, when measured properly, desegregation effects were of about this magnitude.

Crain and Mahard's adjusted effect size for desegregation efforts suffers from the problems that plague Willig's analysis (1985) of effects of bilingual education. When they include many dependent findings from a small number of studies in an analysis, researchers cannot judge the dependability of relationships they find in the data. They run the risks of capitalizing on chance in their estimates and exaggerating the reliability of their findings. Crain and Mahard's method of analysis does not provide sufficient safeguards against these possibilities.

Wortman (1983) also carried out a meta-analysis of the desegregation literature. His literature search yielded a total of 105 studies on the topic. He developed specific methodological criteria for including studies in his analysis, and found that although all studies had some serious flaws, 31 seemed acceptable for analysis. The 31 studies yielded an average effect size of 0.45 standard deviations. Wortman argued that this effect was somewhat inflated by initial subject nonequivalence in some studies. He concluded that the best evidence of the effect of school desegregation on black achievement came from studies in which subject equivalence was not a problem. He concluded that the average effect in such studies was an increase in achievement of Blacks of about 0.2 standard deviations.

Perhaps the best estimate of the effects of desegregation come from a panel convened by the National Institute of Education to review studies of

desegregation and to describe their findings. The panel reviewed Wortman's collection of 31 studies and selected 19 high-quality studies to use as a basis for conclusions. Cook (1984) reported that at least four of the panelists calculated average effect sizes independently for this set of studies, and their estimates of the average effect of desegregation were 0.04, 0.12, 0.17, and 0.14. Thus, the average effect size for the set can be represented by a figure of 0.12. Cook has pointed out that this agrees well with Krol's figure of 0.10 for the "better" desegregation studies and Crain and Mahard's average of 0.10 for the aggregate of all dependent variables for randomized experiments and studies with both pre-post measures and a control group comprised of segregated blacks. An average effect size on black achievement for desegregation thus appears to be 0.11.

Conclusion

Meta-analytic studies confirm that differences exist between sexes and among racial and ethnic groups in science knowledge and understanding during the elementary, middle, and high school years. Our estimates of the size of these differences, measured in standard deviation units, are as follows:

Comparison	Difference	
	Achievement	Attitude
Male vs. female	0.15	0.10
White vs. Black	0.40	0.10
Anglo vs Hispanic	0.30	0.05

Efforts to decrease the gap between majority and minority students have taken a number of forms. Meta-analytic studies show that several of the efforts have had positive results. Desegregation has small positive effects on the achievement of Black students, and bilingual education likewise produces small positive effects on the school achievement of minority students. College opportunity programs have had more notable effects on minority and disadvantaged students. Measured in standard deviation units, effects for these interventions appear to be as follows:

Intervention	Effect Size
College opportunity programs	0.25
Bilingual education	0.10
Desegregation	0.10

These effects are not large in absolute sense, but they may still be important effects for a society committed to increasing equity and breaking down barriers to equal opportunity.

CHAPTER 12

LESSONS FROM META-ANALYSIS

Meta-analysis is more than an educational fad. For many years reviewers have been using quantitative techniques to answer what are now called meta-analytic questions. Glass's formal definition and christening of meta-analysis in 1976 speeded up the production of quantitative reviews, and it greatly increased their sophistication. Glass's work did not, however, represent a break with tradition. His meta-analytic methodology was almost an inevitable development in the long tradition of quantitative reviews in education.

Meta-analysis today demands attention from research reviewers, and it will very likely continue to receive attention in the future. But the ultimate test of the worth of research methodologies is not their popularity nor longevity, but it is rather their contribution to our understanding. After twelve years of meta-analytic efforts in education, it is time finally to take stock. How much have we learned? Are meta-analytic efforts moving educational research and development forward? Are they producing results that can serve as a guide to action? Is it worthwhile to continue down this road?

Broad Lessons

We believe that researchers and practitioners can draw some important lessons about educational research from meta-analytic findings now available. Some of these lessons concern specific educational treatments, and some are broader lessons. Before considering specific substantive findings from meta-analysis, we will describe two of the broad lessons.

Moderate effects of most educational innovations. Meta-analysis has shown that most experimental treatments and innovative programs in education yield small to moderate positive effects. Large positive effects are the exception rather than the rule. Negative effects for experimental treatments and new programs also occur infrequently.

The largest effect that we found to be reliably associated with a special educational program was approximately 0.9 standard deviations. This is the

difference on achievement examinations of talented students in accelerated programs and equally talented students of the same age not assigned to such programs (Chapter 10). College-level programs using Bloom's Learning for Mastery approach also produce strong, positive results (Chapter 6). Students in such programs perform about 0.7 standard deviations higher on examinations than do students in conventional courses.

The meta-analyses that we reviewed showed that only a few alternative programs have a negative impact on student learning (Chapter 6). Students in open education programs, for example, scored about 0.10 standard deviations lower on examinations than did students in conventional programs. Students who learned from instructional media in elementary and secondary school classes performed slightly worse on examinations and had poorer attitudes than did students who were taught without instructional media.

Most alternative programs, however, produced small to moderate effects, raising examination scores between 0.10 and 0.55 standard deviations. This may not be a new or an extraordinary lesson, but it is an important one for researchers to remember. Because most experimental treatments do not produce large effects, educational researchers seeking statistical significance must use large sample sizes and adequate statistical and experimental controls in their studies. Evaluators likewise must remember that innovative programs seldom produce truly strong results. Special quantitative measurement is usually needed to pin down program effects. The effects cannot usually be observed with the unaided eye. Qualitative evaluations are therefore inappropriate for most educational treatments.

The frequency with which small and moderate effects are reported in educational research raises a disturbing question. Is it possible that some single factor has influenced results in all areas? For example, is it possible that evaluators of innovative programs have simply measured a pervasive human tendency to improve one's performance when given special treatment? Are the small and moderate effects so common in educational research, in other words, simply Hawthorne effects?

Anderson (1983) has argued that experimental effects in education cannot be brushed aside as Hawthorne effects. He points out, first, that some treatments do not produce positive effects but instead produce negative effects. Media-based instruction and open education are examples of such treatments. Anderson also points out that many of the comparisons that produce positive results cover lengthy periods of time and involve several teachers. In other words, positive effects are found even though the conditions of the studies are not conducive to a Hawthorne effect.

Small effects of design features. Meta-analysts have investigated the effects on study outcomes of numerous design and publication features. The influence of such features overall appears to be quite small. True experiments and quasi-experiments produce very similar results; studies using local and standardized tests usually produce similar findings; results do not often vary systematically with study year. Only small differences have been found in studies with

different publication histories, in studies of different lengths, and in studies with different amounts of control for teacher effects.

Reviewers and even researchers sometimes argue that the findings in certain studies should be ignored because the studies have suspicious features. For example, we have sometimes heard it said that the only good evidence on any issue is the most recent evidence; older evidence should be ignored. We have also been told that only true experiments are worthy of a reviewer's attention; quasi-experiments should be dismissed. And we have frequently heard it said that published results can be thrown out since they only reflect the biases of editors and journal reviewers.

Our findings on design and publication features suggest to us that reviewers and researchers should be cautious in making such statements. Moderate relations between study features and outcomes exist in some research areas, but one should not assume that the features influence study results uniformly in all areas. Reviewers must carefully examine the evidence in each research area before deciding to discount results from certain types of studies.

Specific Findings

In addition to providing some broad lessons about educational research, our review of meta-analytic results highlights some specific findings of potential importance for education.

Instructional systems. Since the introduction of programmed instruction in the late 1950s, forecasters have been predicting a revolution in education based on instructional technology. Programmed instruction, television, individualized instruction, and computers were each singled out in turn as the technology that would spark the revolution. We reviewed the evidence on effectiveness of these technologies in Chapter 6.

In terms of contributions to student learning, some instructional technologies show much more promise than others do. Mastery-based teaching systems, for example, seem to be very effective. Students learn more material with them; they generally like mastery learning systems; and the systems do not require inordinate increases in instructional time. But school systems and teachers who wish to capture such gains through mastery learning systems will have to follow the models of Bloom and Keller closely. Related approaches that seem similar on the surface (e.g., Individually Prescribed Instruction) have a far poorer record in evaluation studies.

Other technological approaches that seem worthy of exploration are computer-based instruction and peer- and cross-age tutoring programs. Students like to learn with computers, and they learn to like computers when they use them regularly in the classroom. Students also learn more when some of their instruction is presented on computer, and they apparently learn more efficiently. Tutoring programs are equally effective in helping students to

learn course material. The role that peer- and cross-age tutoring programs can play in schools is therefore an important area for further exploration.

Some programs seem to add much less to the improvement of student performance. These are individual learning packages, programmed instruction, open education, and media-based and audiotutorial programs of instruction. These instructional approaches cannot be recommended to schools solely on the basis of their contributions to student learning.

Instructional design. Studies on instructional design provide a useful supplement to these findings on instructional systems. We reviewed the evidence from such studies in Chapter 7. The studies show the importance of clear definition of learning tasks for students. Whether it is through advance organizers, behavioral objectives, or adjunct questions, lessons can be made more effective by providing cues that focus student attention on the relevant parts of the material. Another important finding from studies on instructional design concerns the use of tests in teaching. Increasing the number of conventional tests used in a class usually has only a small effect on the final level of student performance in the class. Activities that follow a test may be a more critical factor in student learning. The importance of tests as learning devices seems to be increased substantially when teachers follow tests with informative feedback and with a requirement of mastery on test items.

Curricular revision. The major curricular efforts in science education of the 1960s and 1970s emphasized inquiry teaching, science processes, and laboratory experimentation. The meta-analyses that we reviewed in Chapter 8 showed that these curricula have had generally positive effects on student learning. It is not clear, however, whether such evaluation results assure the success of other efforts at curricular revision. Perhaps success may be expected only for curricula that emphasize inquiry methods or only for the specific curricula in science education developed in the 1960s and 1970s.

Another problem in interpreting meta-analytic findings on curricular revision arises from the tests used to evaluate effects. It is difficult to find criterion tests that are equally fair to the conventional curricula that are being replaced and to the new curricula that replace them. Standardized tests tend to be unfair to new curricula because standardized tests usually reflect what is currently taught in the schools. Special tests written to capture the new emphases in alternative curricula tend to be unfair to the older curricula. Some meta-analysts have concluded that part but not all of the effectiveness of the new curricula is traceable to the better alignment of criterion tests with the new curricula. Other meta-analysts have concluded that data are simply inadequate to judge whether criterion tests are equally fair to old and new programs.

Teachers. Meta-analyses conducted so far suggest that stronger programs of teacher education and evaluation will result in substantial improvements in

educational effectiveness. We reviewed the results of these programs in Chapter 9.

The results of inservice programs of teacher education seem especially impressive. Participation in such programs affects teachers in a number of ways. Teachers learn what they are taught in the programs; their attitudes often change; and their behavior in the classroom typically changes too. Most important, student learning increases as a results of these teacher changes. In contrast to inservice programs, preservice training programs seem not to produce notable results.

Meta-analyses also indicate that colleges could increase their effectiveness by paying more attention to student evaluations of teaching. Students learn more from teachers to whom they give high ratings. Colleges can therefore add significantly to student learning by giving student ratings more weight in assigning teachers to courses and in hiring and promotion decisions. Providing teachers with feedback about their teaching or providing them with student feedback and consultation on teaching also leads to improved ratings in subsequent evaluations. Researchers have not determined whether student rating feedback improves student learning.

Classroom organization. The idea that educational effectiveness will be improved greatly by reorganizing classrooms does not get much support from the analyses that we reviewed in Chapter 10. The meta-analyses suggest that most changes in class size and homogeneity of classes would result in only small changes in student achievement. Reducing class sizes from 30 to 15 students per class, for example, would increase student achievement from the 50th to the 57th percentile. Instituting a program of comprehensive grouping would increase student achievement from the 50th to the 52nd percentile.

Classroom organization does make a difference in student learning, however, when its special focus is on providing more opportunities for talented students. Special programs for talented students usually boost the achievement level of such students by a significant amount. Whether the program consists of acceleration, instruction in special classes, or instruction in special groups in regular classrooms, the record of results is very impressive. The exceptional results of such programs for talented students seem to be a product of the special curricular adjustments made in the programs.

Results of reorganizing classrooms seem clearer in the affective domain. Class size, for example, influences students' feelings. Students generally like small classes and dislike large ones. Grouping practices seem to affect the way students feel about themselves. Bright students drop a bit in self-esteem when they move from heterogeneous to high-ability classes; they start seeing themselves as somewhat less special. The opposite happens to slow students in low-ability classes. They show a slight increase in self esteem when they are put into classes with other academically slow students.

Equity. Meta-analysis does not have anything particularly new to say about race and gender differences in school achievement (Chapter 11). Meta-analysts

have shown that boys and girls differ somewhat in their science achievement throughout the elementary and secondary school years. Boys typically outscore girls on science tests by a small amount. Meta-analytic summaries also show that racial and ethnic groups differ in science achievement. Whites outscore Blacks and also Hispanics on measures of both science knowledge and understanding through most of the school years.

Meta-analysis has been more useful in providing information about the effectiveness of programs that might help decrease these achievement gaps. Meta-analyses have shown that most large scale efforts to increase educational opportunity for minority students have positive although moderate effects on minority achievement. Desegregation has helped increase Black achievement to only a small degree. Bilingual education programs also seem to have slightly positive effects on the achievement of minority students, but this relationship needs a good deal more study before the finding can be considered firmly established. College programs for high-risk students also have had positive effects on the achievement of disadvantaged students.

Effects of these programs are small in terms of conventional standards, but the programs may still be of great social importance. We have written elsewhere about the effects of programs for high-risk students, for example:

> How important are such gains? With an *ES* between .25 and .4, GPAs of students from special programs would be about .25 points higher (on a 4-point scale) than GPAs of control students. This means that a typical student from a special program might expect to see improvement of one letter grade in one course each semester. Over the course of the student's whole career in college, perhaps eight grades that might have been C's would be B's, or eight D's would be C's. An increase in persistence from 52 to 60 percent means that 60 students rather than 52 out of 100 will be able to stay in college—an increase of 15 percent. By the usual standards of social science research, these are small effects, not visible to the naked eye without special measuring help.
>
> But for the high-risk students enrolled in these programs, the benefits might seem real enough. Such individuals would fail less often than their peers in conventional programs. Their college records might become a source of pride instead of a point of embarrassment. Participation in these special programs might even enable some potential dropouts to graduate from college. The benefits of these programs might also seem real to policymakers weighing overall societal gains. Such analysts would surely note that an increase of even a few percentage points in the effectiveness of higher education would represent a great savings because of the vast size of the educational enterprise (C. Kulik, Kulik, & Shwalb, 1983, p. 408).

The same kind of reasoning might be applied when policy makers consider the effects of desegregated education and bilingual education. In addition, policy makers should also take into account that such programs have goals that go beyond encouragement of better student performance on content examinations.

Some of the points that we have made on the basis of our review of meta-analyses are not new. Conventional reviewers have reached similar conclusions about some aspects of the educational literature. But in our view the statistical methods used in meta-analysis drive these points home. When supported by careful meta-analysis, the conclusions have a precision, clarity, and force that they lack when expressed as opinions in a narrative review.

The work that has gone into meta-analysis during the past twelve years therefore seems to us to have been worthwhile. Meta-analysis has transformed raw findings from something confusing and contradictory into something that can be meaningful for a variety of audiences. One of our hopes now is that meta-analytic conclusions will be put to use more widely. We hope that researchers will use meta-analytic findings as a guide to the design of further studies, that administrators will use them in the formulation of policy, and that teachers will use them as a guide to action.

REFERENCES

Aiello, N. C. (1981). A meta-analysis comparing alternative methods of individualized and traditional instruction in science. *Dissertation Abstracts International*, **42**, 977A. (University Microfilms 81-18682)

Anderson, R. D. (1983). A consolidation and appraisal of science meta-analyses. *Journal of Research in Science Teaching*, **20**, 497-509.

Asencio, C. E. (1984). Effects of behavioral objectives on student achievement: A meta-analysis of findings. *Dissertation Abstracts International*, **45**, 501A. (University Microfilm No. 084-12499)

Athappilly, K., Smidchens, U., & Kofel, J. W. (1983). A computer-based meta-analysis of the effect of modern mathematics in comparison with traditional mathematics. *Educational Evaluation and Policy Analysis*, **5**, 485-493.

Ausubel, D. P. (1960). The use of advance organizers in learning and retention of meaningful verbal material. *Journal of Educational Psychology*, **51**, 267-272.

Bangert-Drowns, R. L., Kulik, C.-L. C., & Kulik, J. A. (1987, April). *The impact of peekability on feedback effects.* Paper presentation at the annual meeting of the American Educational Research Association, Washington, DC.

Bangert, R. L., Kulik, J. A., & Kulik, C.-L. C. (1983a). Individualized systems of instruction in secondary schools. *Review of Educational Research*, **53**, 143-158.

Bangert, R. L., Kulik, J. A., & Kulik, C.-L. C. (1983b). Effect of coaching programs on achievement test performance. *Review of Educational Research*, **53**, 571-585.

Bangert, R. L., Kulik, J. A., & Kulik, C.-L. C. (1984, August). *The influence of study features on outcomes of educational research.* Paper presented at the annual meeting of the American Psychological Association, Toronto.

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C.-L. C. (1985). Effectiveness of computer-based education in secondary schools. *Journal of Computer-Based Instruction*, **12**, 59-68.

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C.-L. C. (1988, April). *Effects of frequent classroom testing.* Ann Arbor, MI: The University of Michigan, Center for Research on Learning and Teaching.

Becker, B. J. (1983, April). *Influence again: A comparison of methods for meta-analysis.* Paper presented at the annual meeting of American Educational research Association. Montreal, Canada.

Becker, H.J. (1988, April). *The impact of computer use on children's learning.* Baltimore, MD: The Johns Hopkins University, Center for Research on Elementary and Middle Schools.

Bloom, B. S. (1968). Learning for Mastery, *Evaluation Comment*, **1**, 2.

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

Bredderman, T. (1985). Laboratory programs for elementary school science: A meta-analysis of effects on learning. *Science Education*, **69**, 577-591.

Burns, P. K. (1981). A quantitative synthesis of research findings relative to the pedagogical effectiveness of computer-assisted mathematics instruction in elementary and secondary schools. *Dissertation Abstracts International*, **42**, 2946A. (University Microfilms No. 81-28,237)

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Clark, R. E. (1985). Confounding in educational computing research. *Journal of Educational Computing Research*, **1**, 129-139.

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

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.

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

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.

Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. *Research in Higher Education*, **13**, 321-341.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research*, **51**, 281–309.

Cohen, P. A., Ebeling, B. J., & Kulik, J. A. (1981). A meta-analysis of outcome studies of visual-based instruction. *Education Communication and Technology Journal*, **9**, 26–36.

Cook, T. D. (1984). What have Black children gained academically from school integration? Examination of the meta-analytic evidence. In National Institute of Education Panel on the Effect of School Desegregation. *School desegregation and Black achievement.* Washington, DC: National Institute of Education. (ERIC Document Reproduction Service No. ED 241 671)

Crain, R. L. (1984). Dilemmas in meta-analysis: A reply to reanalysis of the desegregation-achievement synthesis. In National Institute of Education Panel on the Effect of School Desegregation. *School desegregation and Black achievement.* Washington, DC: National Institute of Education. (ERIC Reproduction Service No. ED 241 671)

Crain, R. L., & Mahard, R. E. (1978). Desegregation and Black achievement: A review of the research. *Law and Contemporary Problems*, **42**, 17–56.

Crain, R. L., & Mahard, R. E. (1982, June). *Desegregation plans that raise Black achievement: A review of the research.* (N-1844-NIE). Santa Monica, CA: Rand Corporation. (ERIC Reproduction Service No. ED 227 198)

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.

Elashoff, J. D., & Snow, R. E. (Eds.) (1971). *Pygmalion reconsidered.* Worthington, OH: Charles A. Jones.

El-Nemr, M. A. (1980). Meta-analysis of the outcomes of teaching biology as inquiry. *Dissertation Abstracts International*, **40**, 5813. (University Microfilm No. 80-11274)

Enz, J., Horak, W. J., & Blecha. (1982). *Review and analysis of reports of science inservice projects: Recommendations for the future.* Paper presented at the annual meeting of the National Science Teachers Association, Chicago. (ERIC Document Reproduction Service No. ED 216 883)

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.

Fleming, M. L., & Malone, M. R. (1983). The relationship of student characteristics and student performance in science as viewed by meta-analysis research. *Journal of Research in Science Teaching*, **20**, 481–495.

Freeman, H. E. (1984). A meta-analysis of gender differences in mathematics achievement. *Dissertation Abstracts International*, **45**, 2039A. (University Microfilm No. 84-23,477)

Giaconia, R. M., & Hedges, L. V. (1982). Identifying features of effective open education. *Review of Educational Research*, **52**, 579–602.

Gilbert, J. P., McPeek, B., & Mosteller, F. (1977). Statistics and ethics in surgery and anesthesia. *Science*, **198**, 684–689.

Glaser, R., & Rosner, J. (1975). Adaptive environments for learning: Curriculum aspects. In H. Talmage (Ed.), *Systems of individualized education.* Berkeley, CA: McCutchan.

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.

Glass, G. V, & Smith, M. L. (1979). Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis*, **1**, 2–16.

Guskey, T. R., & Gates, S. L. (1985, March). *A synthesis of research on group-based mastery learning programs.* Paper presented at the annual meeting of the American Educational Research Association, Chicago. (ERIC Document Reproduction Service No. ED 262 088)

Hamaker, C. (1986). The effect of adjunct questions on prose learning. *Review of Educational Research*, **56**, 212–242.

Hartley, S. S. (1977). Meta-analysis of the effects of individually paced instruction in mathematics. *Dissertation Abstracts International*, **38**, 4003A. (University Microfilms 77-29926)

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. (1983). 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.

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

Hedges, L. V., Stock, W. (1983). The effects of class size: An examination of rival hypotheses. *American Educational Research Journal*, **20**, 63–85.

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

Horwitz, R. A. (1979). Psychological effects of the "open classroom." *Review of Educational Research*, **49**, 71–86.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *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.

Keller, F. S. (1968). "Good-bye, teacher...". *Journal of Applied Behavioral Analysis*, **1**, 79–89.

Keshock, J. D. (1971). An investigation of the effects of the expectancy phenomena upon the intelligence, achievement, and motivations of inner-city elementary school children. *Dissertation Abstracts International*, **32**, 01-A (University Microfilms No. 71-19,010)

Klauer, K. J. (1984). Intentional and incidental learning with instructional texts: A meta-analysis for 1970–1980. *American Educational Research Journal*, **21**, 323–339.

Krol, R. A. (1979). A meta-analysis of comparative research on the effects of desegregation on academic achievement. *Dissertation Abstracts International*, **40**, (University Microfilms No. 79-07,962)

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research*, **47**, 211–232.

Kulik, C.-L. C. (1985, August). *Effects of inter-class grouping on student achievement and self-esteem.* Paper presented at the annual meeting of the American Psychological Association, Los Angeles. (ERIC Document Reproduction Service No. ED 263 492)

Kulik, C.-L. C., & Kulik, J. A. (1982). Effects of ability grouping on secondary school students: A meta-analysis of evaluation findings. *American Educational Research Journal*, **19**, 415–428.

Kulik, C.-L. C., & Kulik, J. A. (1984, August),. *Effects of ability grouping on elementary school pupils: A meta-analysis.* Paper presented at the annual meeting of the American Psychological Association, Toronto. (ERIC Document Reproduction Service No. ED 255 329)

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, C.-L. C., & Kulik, J. A. (1986–87). Mastery testing and student learning: A meta-analysis. *Journal of Educational Technology Systems*, **15**, 325–345.

Kulik, C.-L. C., & Kulik, J. A. (1986). Effectiveness of computer-based education in colleges. *AEDS Journal*, **19**, 81–108.

Kulik, C.-L. C., & Kulik, J. A. (1988a, April). *How effective are mastery learning programs? An examination of the evidence.* Ann Arbor, MI: The University of Michigan, Center for Research on Learning and Teaching.

Kulik, C.-L. C., & Kulik, J. A. (1988b, June). *Effectiveness of computer-based education; An updated analysis.* Ann Arbor, MI: The University of Michigan, Center for Research on Learning and Teaching.

Kulik, C.-L. C., Kulik, J. A. & Shwalb, B. J. (1983). College programs for high-risk and disadvantaged students: A meta-analysis of findings. *Review of Educational Research*, **53**, 397–414.

Kulik, C.-L. C., Kulik, J. A., & Shwalb, B. J. (1986). The effectiveness of computer-based adult education: A meta-analysis. *Journal of Educational Computing Research*, **2**, 235–252.

Kulik, C.-L. C., Shwalb, B. J., & Kulik, J. A. (1982). Programmed instruction in secondary education: A meta-analysis of evaluation findings. *Journal of Educational Research*, **75**, 133–138.

Kulik, J. A. (1976). PSI: A formative evaluation. In B. A. Green, Jr. (Ed.), *Personalized instruction in higher education: Proceedings of the Second National Conference.* Washington, DC: Center for Personalized Instruction.

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. (1984). Effects of accelerated instruction on students. *Review of Educational Research*, **54**, 409–426.

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)

Kulik, J. A., & Kulik, C.-L. C. (1987). Effects of ability grouping on student achievement. *Equity and Excellence*, **23**, 22–30.

Kulik, J. A., & Kulik, C.-L. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, **58**, 79–97.

Kulik, J. A., Kulik, C.-L. C., & Bangert-Drowns, R. L. (1985a). Effectiveness of computer-based education in elementary schools. *Computers in Human Behavior*, **1**, 59–74.

Kulik, J. A., Kulik, C.-L. C., & Bangert-Drowns, R. L. (1985b). The importance of outcome studies: A reply to Clark. *Journal of Educational Computing Research*, **1**, 381–387.

Kulik, J. A., Kulik, C.-L. C., & Bangert-Drowns, R. L. (1988, May). *Effectiveness of mastery learning programs: A meta-analysis.* Ann Arbor, MI: The University of Michigan, Center for Research on Learning and Teaching.

Kulik, J. A., Kulik, C.-L. C., & Carmichael, K. (1974). The Keller Plan in science teaching. *Science*, **183**, 379–383.

Kulik, J. A., Kulik, C.-L. C., & Cohen, P. A. (1979a). A meta-analysis of outcome studies of Keller's Personalized System of Instruction. *American Psychologist*, **34**, 307–318.

Kulik, J. A., Kulik, C.-L. C., & Cohen, P. A. (1979b). Research on audio-tutorial instruction: A meta-analysis of comparative studies. *Research in Higher Education*, **11**, 321–341.

Kulik, J. A., Kulik, C.-L. C., & Smith, B. B. (1976). Research on the Personalized System of Instruction. *Journal of Programmed Learning and Educational Technology*, **13**, 23–30.

Lamb, W. K., & Whitla, D. K. (1981). *Meta-analysis and the integration of research findings: A bibliography prior to 1981.* Cambridge, MA: Harvard University.

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

Luiten, J., Ames, W., & Ackerson, G. (1980). A meta-analysis of the effects of advance organizers on learning and retention. *American Educational Research Journal*, **17**, 211–218. (University Microfilms No. 84–09065)

Lyday, N. L. (1983). A meta-analysis of the adjunct question literature. *Dissertation Abstracts International*, **45**, 129A.

Mager, R. F. (1962). *Preparing instructional objectives.* Palo Alto, CA: Fearon Publishers.

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

Malone, M. R. (1984, April). *Project MAFEX: Report on preservice field experiences in science education.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, New Orleans. (ERIC Document Reproduction Service No. ED 244 928)

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

Moore, D. W., & Readence, J. E. (1984). A quantitative and qualitative review of graphic organizer research. *Journal of Educational Research*, **78**, 11–17.

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

Mosteller, F. M., & Tukey, J. W. (1977). *Data analysis and regression.* Reading, MA; Addison-Wesley.

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

Niemiec, R. P. The meta-analysis of computer-assisted instruction at the elementary school level. *Dissertation Abstracts International*, **45**, 3330. (University Microfilms 85-01250)

Niemiec, R. P., Sikorski, M. F., & Walberg, H. J. (1987). *Comparing the cost-effectiveness of tutoring and computer-based instruction.* Digital Equipment Corporation.

Niemiec, R. P., & Walberg, H. J. (1985). Computers and achievement in the elementary schools. *Journal of Educational Computing Research*, **1**, 435–440.

Niemiec, R. P., & Walberg, H. J. (1987). Comparative effects of computer-assisted instruction: A synthesis of reviews. *Journal of Educational Computing Research*, **3**, 19–37.

Orlansky, J., & String, J. (1979). *Cost-effectiveness of computer-based instruction in military training (IDA paper P-1375)*, Institute for Defense Analysis, Science and Technology Division, Arlington, VA.

Peterson, P. L. (1979). In P. L. Peterson & H. L. Walberg (Eds.), *Research on teaching.* Berkeley, CA: McCutchan.

Postlethwait, S. W., Novak, J., & Murray, H. T., Jr. (1972). *The audio-tutorial approach to learning.* Minneapolis: Burgess Publishing Co.

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

Redfield, D. L., & Rousseau, E. W. (1981). A meta-analysis of experimental research on teacher questioning behavior. *Review of Educational Research*, **51**, 237–245.

Robin, A. R. (1976). Behavioral instruction in the college classroom. *Review of Educational Research*, **46**, 313–354.

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.

Rosenthal, R. (1976). *Experimenter effects in behavioral research.* New York: Irvington.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin*, **85**, 185–193.

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

Rosenthal, R., & Jacobson, L. (1968) *Pygmalion in the classroom.* New York: Holt, Rinehart, & Winston.

Rosenthal, R., & Rubin, D. B. (1978) Interpersonal expectancy effects: The first 345 studies. *The Behavioral and Brain Sciences*, **1**, 377–386.

Rothkopf, E. Z. (1966). Learning from written instructive materials: An exploration of the control of inspection behavior by test-like events. *American Educational Research Journal*, **3**, 241–249.

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. Validity generalization: Results for computer programmers. *Journal of Applied Psychology*, **65**, 1980, 643–661.

Schmidt, M., Weinstein, T., Niemiec, R., & Walberg, H. J. (1985). *Computer-assisted instruction with exceptional children: A meta-analysis of research findings.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Shymansky, J. A., Kyle, W. C. Jr., & Alport, J. M. (1983). The effects of new science curricula on students performance. *Journal of Research in Science Teaching*, **20**, 387–404.

Skinner, B F. (1954). The science of learning and the art of teaching. *Harvard Educational Review*, **24**, 86–97.

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

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

Slavin, R. E. (1987a). Ability grouping and student achievement in elementary schools: Best evidence synthesis. *Review of Educational Research*, **57**, 293–336.

Slavin, R. E. (1987b). Mastery learning reconsidered. *Review of Educational Research*, **57**, 175–213.

Slavin, R. E., & Karweit, N. (1984, April). *Within-class ability grouping and student achievement.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

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

Smith, M. L., & Glass, G. V. (1980). Meta-analysis of research on class size and its relationship to attitudes and instruction. *American Educational Research Journal*, **17**, 419–433.

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

Steinkamp, M. L. & Maehr,M. L. (1984). Gender differences in motivational orientations toward achievement in school science: A quantitative synthesis. *American Educational Research Journal*, **21**, 39–59.

Stone, C. L. (1983). A meta-analysis of advanced organizer studies. *Journal of Experimental Education*, **51**, 194–199.

Sweitzer, G. L. (1985). A meta-analysis of research on preservice and inservice science teacher education practices designed to produce outcomes associated with inquiry strategy. *Dissertation Abstracts International*, **46**, 128A. (University Microfilms No. 85-04089)

Sweitzer, G. L., & Anderson, R. D. (1983). A meta-analysis of research on science teacher education practices associated with inquiry strategy. *Journal of Research in Science Teaching*, **20**, 453–466.

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

Wade, R. K. (1984). What makes a difference in inservice teacher education: A meta-analysis of the research. *Dissertation Abstracts International*, **45**, 155. (University Microfilms No. 84-10,341)

Wade, R. K. (1985). What makes a difference in inservice teacher education? A meta-analysis of research. *Educational Leadership*, **42**, 48–54.

Weinstein, T., Boulanger, F. D., & Walberg, H. J. (1982). Science curriculum effects in high school: A quantitative synthesis. *Journal of Research in Science Teaching*, **19**, 511–522.

Willett, J. B., Yamashita, J. J., & Anderson, R. D. (1983). A meta-analysis of instructional systems applied in science teaching. *Journal of Research in Science Teaching*, **20**, 405–417.

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

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

Wortman, P. M. (1984). School desegregation and black achievement: An integrative review (NIE-P-82-0070). In National Institute of Education Panel on the Effect of School Desegregation. *School desegregation and Black achievement.* Washington, DC: National Institute of Education. (ERIC Document Reproduction No. 241 671)

Yeany, R. H., & Porter, C. F. (1982, April). *The effects of strategy analysis on science teacher behaviors: A meta-analysis.* Paper presented at the National Conference of the National Association for Research in Science Teaching, Chicago. (ERIC Document Reproduction Service No. ED 216 858)