COMPARISON OF METHODS FOR COMPUTATION AND CUMULATION OF EFFECT SIZES IN META-ANALYSIS

DISSERTATION

Presented to the Graduate Council of the
North Texas State University in Partial
Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

By

Sharron L. Ronco, B.A., M.S.
Denton, Texas
December, 1987

wkb

Ronco, Sharron L., Comparison of Methods for Computation and Cumulation of Effect Sizes in Meta-Analysis. Doctor of Philosophy (Educational Research), December, 1987, 102 pp., 10 tables, bibliography, 143 titles.

This study examined the statistical consequences of employing various methods of computing and cumulating effect sizes in meta-analysis. Six methods of computing effect size, and three techniques for combining study outcomes, were compared. Effect size metrics were calculated with one-group and pooled standardizing denominators, corrected for bias and for unreliability of measurement, and weighted by sample size and by sample variance. Cumulating techniques employed as units of analysis the effect size, the study, and an average study effect. In order to determine whether outcomes might vary with the size of the meta-analysis, mean effect sizes were also compared for two smaller subsets of studies.

An existing meta-analysis of 60 studies examining the effectiveness of computer-based instruction was used as a data base for this investigation. Recomputation of the original study data under the six different effect size formulas showed no significant difference among the metrics. Maintaining the independence of the data by using only one effect size per study, whether a single or averaged effect,

produced a higher mean effect size than averaging all effect sizes together, although the difference did not reach statistical significance. The sampling distribution of effect size means approached that of the population of 60 studies for subsets consisting of 40 studies, but not for subsets of 20 studies.

Results of this study indicated that the researcher may choose any of the methods for effect size calculation or cumulation without fear of biasing the outcome of the meta-analysis. If weighted effect sizes are to be used, care must be taken to avoid giving undue influence to studies which may have large sample sizes, but not necessarily be the most meaningful, theoretically representative, or elegantly designed. It is important for the researcher to locate all relevant studies on the topic under investigation, since selective or even random sampling may bias the results of small meta-analyses.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

Figure Page

CHAPTER I

INTRODUCTION

Historically, educational research has suffered from what N.L. Gage (1982) has characterized as a massive Type II error. He attributes this failure to detect true treatment effects to erroneous methods of research synthesis. For science to be cumulative, an intermediate step between past and future research is necessary: synthesis of existing research and conflict resolution. In recent years, researchers have become increasingly concerned with this intermediate step, which has been termed by different authors, "integrating findings," "research synthesis," "research summation," "synthesizing outcomes," or "combining results." In the past, attempts at integrating research evidence have relied on narrative review or vote-counting methods. However, the traditional review, in which the reviewer reads a set of studies on a given topic and attempts to derive the essence of the results, is highly subjective as well as impractical given the current sizeable body of research on almost any topic. Vote-counting, which involves determining the frequency of studies of differing direction and statistical significance, shows a primitive sensitivity to the quantification of outcomes, but will only reliably represent a treatment effect if all studies share the same sample size, use the same treatment, and have a unimodal distribution reflecting one population. For a long time, researchers felt uneasy about the reliability of qualitative methods, but no alternative was available (Kulik, 1984).

Meta-analysis, which was first introduced by Gene V. Glass in his 1976 American Educational Research Association Presidential Address, is the quantitative review of an explicitly-defined body of research, whose general purpose is to describe and explain the variability in the outcomes of that body of research (Bangert-Drowns, 1984). It is distinguishable from primary research, secondary research and narrative review by methods of sample selection, data collection and analysis. Although not a clear break with earlier methods, meta-analysis relies more heavily on quanitification and statistical techniques than its predecessors.

In the decade since Glass' introduction of meta-analysis, scores of meta-analyses relating to educational practice and policy have appeared, and the number of articles using or discussing meta-analysis has approximately doubled each year between 1979 and 1983 (Slavin, 1986). Concurrently, differing and occasionally conflicting methodologies have arisen for selecting samples, computing and cumulating effect sizes, and testing study characteristics on study outcomes. Bangert-Drowns (1986) has recently distinguished five alternative approaches to meta-analysis, each with its own philosophy, basic purpose and analytic strategy. These disagreements only obscure what was originally intended to be a tool of clarification. Thus, we have come full circle from where we began: we looked to meta-analysis to resolve the conflicts in primary research, only to discover that meta-analysis itself could produce no consensus. Yet, meta-analytical techniques show some promise for increasing the reliability and dependability of a researcher's conclusions, and it is difficult to justify a return to reviews with arbitrary and subjective study selection and analysis procedures. It is necessary to investigate and clarify the differences in meta-analytic method so that this approach to research integration can be most effectively used.

Statement of the Problem

In meta-analysis, effect sizes are computed for each study outcome and are then averaged across studies. Although a number of methods for estimating effect sizes have been developed in recent years, there is no real agreement as to which is "best" or even whether the choice of metric significantly influences the outcome of research synthesis. In addition, outcomes may be cumulated using either the effect size, the study, or the average of all effect sizes in the study as the unit of analysis.

In an effort to clarify the effects of employing the different methodologies, this study attempts to determine the degree to which the methods of computing and cumulating effect sizes will significantly affect the outcome of the meta-analysis. Because outcomes may vary depending on the number of studies included in the meta-analysis, mean effect sizes are computed for two smaller subsets of studies and these are compared to the complete study set.

Purpose of the Study

This study was concerned with a comparison of six methods for computing effect sizes from individual studies, and averaging them over the studies. Effect sizes which are compared include: 1) Cohen's $\underline{d}$, 2) Glass' Δ, 3) Hedges' unbiased $\underline{g}'$, 4) Hedges' weighted $\underline{g}^{w}$, 5) Hunter's weighted $\underline{d}^{w}$, and 6) Hunter's $\underline{d}^{r}$ corrected for unreliability of measurement. In addition, three common methods for combining studies with more than one effect size were compared. These cumulation techniques include: 1) using the study as the unit of analysis, 2) using the effect size as the unit of analysis, and 3) using a pooled effect size to represent the study.

In order to determine how smaller meta-analytic data sets may differ from larger ones, random samples of 20 and 40 studies each are drawn from the complete set of 60 studies and their properties compared with the population set.

Research Questions

To carry out the purposes of the study, the following research questions were addressed:

Research Question 1: Do the six methods of computing effect size measure study outcomes in similar ways?

Research Question 2: Do the three methods of cumulating effect sizes across studies measure study outcomes in similar ways?

Research Question 3: Will a random sampling distribution of effect size means approximate the population distribution for smaller sample sizes?

Significance of the Study

There is a growing consensus that meta-analysis can be a useful tool for synthesizing research if properly used, but, as with any other methodology, indiscriminate use can lead to abuse. Robert Slavin (1984) has gone so far as to proclaim that the way in which meta-analyses are typically conducted in education is a significant step backward in the art of research synthesis. The problem is partially attributable to the fact that the typical meta-analyst, like the primary researcher, tends to be more knowledgeable and interested in the hypothesis under consideration than in the statistical methodology for testing it. However, the researcher seeking technical details on quantitative methods for meta-analysis is confronted with a bewildering plethora of choices with virtually no consensus to guide her. Although Glass presented meta-analysis as method-free, many of his techniques became the standard for researchers seeking to cumulate research findings. Hedges and Olkin (1985) regard Glass' original formulation as outdated and present what they believe to be a more technically-adequate form. Rosenthal (1984) has independently developed his own meta-analytic techniques which have become the primary source for many researchers. Hunter, Schmidt and Jackson (1982) have referred to their meta-analytic work as "state-of-the-art" and "the most complete meta-analysis procedure now known." Glass, whose procedures are proving robust, has restated his confidence in his original formulation of meta-analysis (1983). J.A. Kulik and his colleagues have practiced a modified version of Glass' methodology, which Bangert-Drowns has termed "study effects" meta-analysis (1986).

The meta-analyst must choose from among these alternatives the most reasonable method for cumulation. There has been virtually no empirical work to compare them (Bangert-Drowns, 1984). Reynold and Day (1984) have suggested that additional study of the behavior of effect size estimates should precede a more widespread application of meta-analysis.

Although the present study directly compares different methods of effect size computation, it should be noted that these methods are not necessarily intended to be used interchangeably. Bangert-Drowns (1984) has pointed out that the different methodologies have sprung from a divergence on the general purpose of meta-analysis. On the one hand, there are those meta-analysts whose purpose is primarily to review a body of research, much like a narrative review. On the other hand, there are those meta-analysts whose purpose is to increase the sample size to test a specific hypothesis and determine a generalizable estimate of treatment effect. Glass and Kulik espouse the former method of literature review, while Hedges, Rosenthal, Hunter and their collaborators advocate the estimation of the distribution of treatment effects.

Although this difference in purpose is real and dramatic, in practice, few meta-analysts differentiate between the approaches or specify which attitude is taken toward the synthesized data. A method is chosen, based on familiarity, simplicity of computation, or the recommendation of one's colleagues. The purpose of the present study, therefore, was to examine the degree to which the researcher's choice of methods for effect size computation and combination will affect the outcome of the meta-analysis. This is not to imply that the methods are equivalent, in the case of no significant difference among them, or to advocate the use of one method over another, in the event that they differ. But it is important that the authors, consumers and critics of meta-analysis understand the statistical consequences of metric and cumulation choices so that they can make more informed evaluations of meta-analytic findings. It is hoped that the present study will add to that understanding.

Limitations

Although the present study averages effect sizes across all research reports included in the meta-analysis for purposes of examining the behavior of effect sizes obtained through different methodologies, it is not intended that statistics be mindlessly applied to any set of data. Pooling and statistical comparisons must be guided by substantive, methodological, and theoretical considerations, not conducted wholesale and interpreted according to statistical criteria alone (Slavin, 1986). The basic premise behind the use of statistics in reviews is that a series of studies have been identified that address an identical conceptual hypothesis. The reviewer must decide whether an overall quantitative summary will be useful and substantively sound. Possibly, too much attention in the past has been given to averaging effect sizes. To summarize a stream of research with an average effect size is to imply that the effect of a treatment is constant across populations of subjects, contexts, implementation of treatment, and research designs. To the extent that the treatments interact with any of these factors, averaging can be misleading. In addition to theoretical considerations and "eyeballing" the data for unusual distribution of outcomes, most authors recommend that a test of homogeneity be applied to effect sizes. Heterogeneity provides a warning that it may not be appropriate to combine and synthesize all the study results in one meta-analysis. Data in the present study have not been tested for homogeneity, and so the results themselves should not be taken at face value although the relationships among average effect sizes computed would generally hold.

In addition, it should be noted that the obtained differences (or lack of them) among average effect sizes derived from different methodologies may not hold constant when outcomes are tested against study features.

Assumptions

It is assumed that the summary statistics (sample size, mean and standard deviation of each experimental and control group), and reliability of dependent measure were correctly reported by the researchers. It is assumed that studies are of good research quality and address an identical conceptual hypothesis.

Since only parametric methods for computing effect sizes are included in this study, all of the assumptions underlying these methods are considered met; that scores are normally distributed, that treatment benefits all subjects equally, that group variances are homogeneous, and that effect size is not invariant under all monotonic transformations of scales.

The approach of averaging effect sizes in itself assumes that the size of the effect reported in each study is an estimate of the common effect of the population of studies.

CHAPTER BIBLIOGRAPHY

Bangert-Drowns, R. L. (1984). *Developments in meta-analysis: A review of five methods*. Ann Arbor, MI: Michigan University, Center for Research on Learning and Teaching. (ERIC Document Reproduction Service No. ED 248 262)

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

Gage, N. L. (1982). Future of educational research. *Educational Researcher*, *11*(8), 11-19.

Glass, G. V., & Kliegl, J. M. (1983). An apology for research integration in the study of psychotherapy. *Journal of Consulting and Clinical Psychology*, *51*, 28-41.

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

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

Kulik, J. A. (1984, April). *Uses and misuses of meta-analysis*. Paper presented at the 68th Annual Meeting of the American Educational Research Association, New Orleans, LA.

Reynolds, S. & Day, J. (1984, August). *Monte Carlo studies of effect size estimates and their approximations in meta-analysis*. Paper presented at the 92nd Annual Con vention of the American Psychological Association, Toronto, Ontario.

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

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

Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher*, *15*(9), 5-11.

CHAPTER II

SYNTHESIS OF RELATED LITERATURE

The literature review has always served the important function of "taking stock" of what is known so that future research can be directed more efficiently, policy decisions can be made more effectively, and information can be disseminated to wider audiences (Strube, 1983). Literature reviews seek to establish the "facts," those stubborn, dependable relationships that regularly occur despite any biases that may be present in particular studies because of investigator's research design, choice of measure, observation schedules, and the like. However, the precise methods used in the single studies that comprise a given literature have not generally been applied when these same studies were reviewed and integrated.

Background

Historically, the most common method of integrating the findings of a number of studies has been by means of a narrative or literary review. With this approach, the reviewer reads a set of studies on a given topic and attempts to derive general summary statements that capture the essence of the results and render the findings useful to the reader.

Although this approach may have been practical in the past, the current sizeable body of research and conflicting results will surely result in "cognitive overload" for the reviewer (Oliver & Spokane, 1983). Furthermore, the traditional literature review is highly subjective; the reviewer may choose several favorite studies based either on her judgment that they are well-designed from a classical experimental-design standpoint, or because they are carried out by investigators she respects, or even because they agree with her own hypothesis. In any case, her impressionistic conclusions will differ from those of the next well-intentioned reviewer.

A more statistically sensitive type of narrative review, the vote-counting or box-score method, involves sorting studies into three categories: those significantly favoring the experimental group, those significantly favoring the control group, and those with nonsignificant outcomes. The category with the most studies is "voted" the winner, and is assumed to give the best estimate of the true relationship between the independent and dependent variables (Light & Smith, 1971). Counting statistically significant findings has its weaknesses, however. Since statistical significance is a product of both sample size and treatment effect, large studies are more likely to show significance even where effect size is small. In addition, a bimodal or multimodal distribution could indicate that outcomes are mediated by factors other than the treatment. Hedges and Olkin (1985) have shown that the vote-counting method actually tends to make the wrong decision more often as the number of studies increases, since incorrect decisions of the same type do not cancel one another.

In 1971, Light and Smith proposed a form of integrative review called "cluster analysis." Cluster analysis resembles modern meta-analytic procedure in that studies are quantitatively represented and statistical analysis is applied to the data. This method of synthesis, however, combines original study data rather than effect sizes, and allows only studies that are methodological replications testing the same hypothesis. The point is to increase sample size where data can be grouped into natural aggregations, or clusters. The investigator can either draw conclusions from the larger samples or attribute differences in study findings to differences among the clusters.

Meta-analysis, a term coined by Gene V. Glass in 1976, is distinguished from the traditional forms of review by its application of statistical techniques to the treatment of quantitative representations of study outcomes. Although Glass terms his meta-analysis as a "tiny revolution" in the way social scientists and researchers attempt to extract information from empirical inquiry, he acknowledges that it is not a clear break with earlier integrative methods. Under the pressure of numbers, research reviewers have gradually adopted increasingly rigorous and quantitative methods of study integration. The meta-analytic attitude was a natural development of the desire for greater quantitative and statistical rigor in organizing diverse findings in social science research, and there is a growing consensus that meta-analysis can be a useful tool if properly used (Bangert-Drowns, 1984; Slavin, 1984). However, although the meta-analytic attitude is generally regarded favorably, there is no consensus regarding the methodology of meta-analysis. The introduction of inferential statistics and formal statistical hypothesis testing to a domain in which conclusions formerly rested on opinion or the crude operation of counting was a quantum leap. As a result, the application of meta-analytic methods has, in many ways, preceded the development of sound underlying mathematical theory (Kraemer, 1983; Mintz, 1983).

Meta-Analysis: Methodology

Over the years of its development, meta-analysis has been misunderstood and occasionally misrepresented. It has been characterized as "averaging effect sizes," which, says Glass, is a little like characterizing analysis of variance as "adding and multiplying." (Glass, McGaw & Smith, 1981). The research methods applied to the characteristics of findings of reports of research studies in meta-analysis include those typical of empirical research: problem selection, hypothesis formulation, definition and measurement of constructs and variables, data analysis, conclusions. Typically, the meta-analyst seeks to locate all relevant studies on a specifically defined question, such as the effects of computer-based instruction, class size on achievement, gender differences in aptitudes, open classrooms, desegregation, and so on. Although there is some debate about whether the meta-analyst should exclude methodologically flawed studies, the meta-analytic sample is not representative or random, but a close approximation to the existing population. Each selected study is represented by its features, both substantive and methodological, and its outcomes. Substantive features are specific to the problem under study and include type, length, duration and variation in treatment; race, sex and age of subjects; classroom demographics; study conditions, and so on. Methodological characteristics may be nearly the same for all meta-analyses and include sample size, reliability of measurement instruments, random or nonrandom subject selection, subject mortality, and so on. Each characteristic is given a quantitative or quasi-quantitative categorical coding.

There are two major ways to evaluate the outcomes of research studies: in terms of their statistical significance and in terms of effect size. Omnibus tests of statistical significance for combined results have been extensively examined by Rosenthal (1984). Although these can almost always be successfully applied to data collected for research synthesis, they often do not provide a test of the hypothesis of interest to the research reviewer. Such tests do not, for example, support inferences about the average magnitude of effects or about the consistency of results across studies, and they are overly dependent on sample size. These questions are addressed through the combination and comparison of effect sizes.

An effect size is a standardized index of the magnitude of effects which is independent of the scale of measurement used in the original study. It indicates the magnitude of difference between two groups, the degree of departure from the null hypothesis. An effect size of .30, for example, indicates that three-tenths of a standard deviation separates the average subjects of the two groups. Effect sizes should be expressed as a comparison between two groups, since multiple degree of freedom tests do not indicate which means differ. There are no consensually-accepted standards for evaluating what constitutes a meaningful effect size. Although Cohen (1977) has defined large, medium and small effect sizes, these designations have not achieved general acceptance. Glass argues that dissociated from a context of decision and comparison of benefits and costs, effect sizes have no inherent value and should not be assigned descriptive adjectives. "After decades of confusion, researchers are finally ceasing to speak of regions of the correlation coefficient scale as low, medium or high. The same error should not be repeated in the case of effect size metric" (Glass et al., 1981, p. 104).

Once coding is completed and effect sizes are computed and averaged, the major work of meta-analysis, the testing of study features on study outcomes, can begin. The type of analysis will depend on the meta-analyst's attitude toward the data and purpose for undertaking the review. Bangert-Drowns (1986) has developed a taxonomy of five basic types of meta-analysis, ranging from general description of a body of literature to approximate data-pooling techniques. Although all involve computing an average effect size, these may, for example, be compared in pre-established categories, tested for homogeneity, or tested for variation attributable to sampling error. Depending on the outcome of preliminary analysis, main and interaction effects can be explored using a number of techniques.

Integrative reviews undertaken using meta-analysis have repeatedly reached less conservative conclusions about the presence and magnitude of particular effects than have traditional literature reviews. Cooper and Rosenthal (1980) tested this hypothesis by randomly assigning graduate students and faculty members to review a set of related studies, using either a meta-analytic approach or the traditional qualitative method. Reviewers who used meta-analysis believed that there was more support for the phenomenon under study than did qualitative reviewers, even though they were only reviewing seven studies. Willig's reanalysis of 28 studies on the efficiency of bilingual education using Glass' meta-analytic methods found moderate differences favoring bilingual education; whereas the original review using traditional techniques concluded that the case for bilingual education was weak at best (Willig, 1985).

Advantages over Primary Research

Although the most frequently-cited virtue of synthesis is that increased sample size can increase statistical power, the interaction question can be as important as the main effects question. By capitalizing on study-level variation, meta-analysis shows its strongest advantage over even the most carefully-executed single study (Light & Pillemer, 1984). One study with a single research design, special program, geographic location and participant type cannot examine contextual effects, but a synthesis of several studies can turn up richer, more useful information. Conflicting findings offer opportunities for learning about these contextual effects on study outcomes. Light points out that synthesis can help match treatment type with recipient type; can explain which features of a treatment matter; can explain conflicting results; can evaluate the stability of treatment effects; and can assess the importance of research design (Light, 1984). Findings from a synthesis help inform policy decision-making by making a study as powerful as possible in answering a specific question or resolving a dilemma. Strube and Hartman (1983) believe that meta-analysis can also serve a predictive function, by examining the plausibility of hypotheses that have not been tested in single studies. Because each data point in a meta-analysis is a study with its own methodological and theoretical characteristics, it is possible to construct variables and test their relationship to study outcome. The use of regression analysis allows one to predict or estimate study outcome given specific values of independent variables. The values used in a regression analysis need not have existed in any one study.

Most experts agree that rather than suppressing the production of primary research, meta-analysis may contribute to the quality of subsequent studies on the topic. In a world of scarce resources, targeting the features of a treatment or program that seems to matter is a valuable endeavor.

Calculating and Cumulating Effect Sizes

Meta-analytic investigators have a veritable arsenal of statistical techniques at their disposal. A wide variety of methods exists for calculating and interpreting effect sizes, and further analysis can proceed along a number of lines. Such choice leaves open the possibility that the results of a meta-analysis can vary depending on the specific techniques used (Strube & Hartmann, 1983). Consequently, there is much controversy and doubt about the validity of results based on meta-analysis.

Although there are many meta-analytic techniques still undergoing refinement, the present study focused on methods of computing and cumulating effect sizes. Combining effect sizes is the more common approach to meta-analysis, although combining probabilities also has been widely used. The combination of probability levels across studies allows the reviewer to determine whether a set of results could have arisen by chance, whereas the combination of effect sizes is done to examine the magnitude of effects across studies. These methods are usually correlated; however, they provide different information since a statistically significant result is not necessarily a meaningful one.

A rather large number of methods for estimating effect sizes have been developed in recent years, and there is no real agreement as to which is "best" or even whether the choice of metric significantly influences the outcome of the research synthesis. When Glass introduced meta-analysis, he advocated an effect size standardized by the control group standard deviation. This was a departure from Cohen's $\underline{d}$, which uses the pooled, or within-group standard deviation. Glass reasoned that a treatment-by-subject interaction may have an effect on the means and standard deviations of the experimental group. Heterogeneous group variances cause difficulties, and standardization of mean differences by the control group standard deviation at least has the advantage of assigning equal effect sizes to equal (pre-treatment) means. Other researchers in meta-analysis disagree. Hedges and Olkin (1985) argue that in most cases, the assumption of equal population variances is reasonable, which suggests that the most precise estimate of the population variance is obtained by pooling. Hedges' modification of Cohen's d employs N-1, rather than N, as the within-group divisor for the sum of squares in the standardizing denominator. N is better used if sample sizes are equal. Hunter and his colleagues (1982) agree that since there is rarely a large difference between the control and experimental group means, it is reasonable to use the within-group standard deviation, which has only about half the sampling error of the control group standard deviation. If treatment-by-subject interaction is suspected, there are more effective procedures for addressing this problem than altering the definition of effect size. Rosenthal (1984) adds that computing the standard deviation from the control group only may cause ordinary t tests to give misleading results.

Hedges has identified both Glass' Δ as well as Cohen's d as biased estimators, demonstrating that these effect sizes have a noncentral t distribution and are therefore asymmetric, non-normal and positively skewed when the population effect is not zero. Accordingly, Hedges has formulated a correction factor, which when multiplied by d, produces an unbiased estimator of the population treatment effect. The unbiased estimator g' has a smaller mean-square error than d, and therefore less variance. The practical necessity for using the unbiased estimator is not established. Hedges has demonstrated that g' tends to d as N increases, and they are essentially the same estimators in large samples. However, since the correction for bias is easy to apply and the unbiased estimator has theoretical advantages, he recommends that the bias correction be applied routinely (Hedges, 1981). Rosenthal and Rubin (1982) conclude that d is effectively unbiased for studies where the degrees of freedom exceed ten, or where studies have approximately the same number of degrees of freedom. Cooper (1984) suggests that Hedges' correction factor be applied to effect sizes from primary research based on samples smaller than 20; whereas Slavin (1986) reports that the Hedges formula reduces estimates from studies with total sample sizes less than 50. Bangert-Drowns, Kulik, and Kulik (1983) calculated both Glass' Δ and Hedges g' in their meta-analysis of the effects of coaching programs on achievement test performance and found that the two statistics were nearly identical in every case, with a correlation of .999 over 27 studies. In a Monte Carlo study examining the behavior of effect sizes, Reynolds and Day (1984) found that both Cohen's d and Hedges g' overestimated 'true' effect size; in the case of g', the overestimation was as much as 13% for large effect sizes and small degrees of freedom.

Other measures of effect size, such as percent of variance accounted for, and percentage overlap between treatment and control conditions, no longer enjoy widespread usage. Robert Rosenthal, who has been a major contributor to the methodology and practice of research synthesis, prefers the correlation coefficient *r* to *d* as an effect size estimator, and his methods for cumulation and analysis are therefore based on *r*. Rosenthal prefers *r* because no special adjustments are needed when moving from *t* tests for independent to those for correlated observations, because it is sometimes not possible to compute *d* accurately from the research information provided, and because it is more readily interpretable as an effect size. His binomial effect size display (BESD) is a method for demonstrating the practical importance of the size of the obtained effect. Since *r* is a direct algebraic transformation of *d*, it was not further considered in this study.

There is also disagreement over how the effect sizes from the individual studies should be averaged. Glass and Kulik use unweighted averages of effect size, in order to avoid giving too great a weight to studies that may have large sample sizes but not necessarily be the most methodologically sound or theoretically representative. Other researchers prefer weighted means where studies do not share a common sample size, which is generally the case.

Hedges points out that the variance of the estimator depends on the sample size, so that effect sizes from studies with larger sample sizes are more precise than those from studies with smaller sample sizes. The weights that minimize the variance of g give weight that is inversely proportional to the variance in each study. Smaller variance (more precision) should lead to a larger weight for the study (Hedges & Olkin, 1985). Rosenthal and Hunter advocate the use of simple frequency-weighted mean effect sizes. Effect sizes may also be weighted by estimated research quality or by any other weights assigned before inspection of the data.

Some researchers such as Hedges and Hunter advocate correcting effect sizes for measurement error. To the extent that the measurement of the dependent variable is less than perfectly reliable, errors of measurement will cause the observed d value to be an underestimate of the actual effect size. In addition, if there is variation across studies in the reliability of measures, this will cause variation in observed d values, which can be eliminated by correcting each for unreliability. Smith and Glass' 1977 meta-analysis of psychotherapy outcomes was reanalyzed be Orwin and Cordray (1983), who selected a stratified random sample from the study set and corrected for unreliability. The reanalysis obtained different results from those produced by the original study.

The issue of nonindependence or multiple dependent variables is complex and important, and is just beginning to be addressed more sophisticatedly. Many studies may provide more than one effect size relevant to the hypothesis under examination in the meta-analysis; to cull out pertinent data can lead to a loss of valuable information. Glass and his colleagues have included multiple tests from the same study in a single meta-analysis, whereas most other researchers do not. Glass reasons that treating each finding in a study as independent of the others "may be a risky and untrue assumption," (Glass et al., 1981, p. 200) but it is practical since the effect of dependence is almost certain to increase the standard errors of estimate above what they would be if the same number of data points were independent, thus erring on the conservative side. Other researchers point out that by using the effect size as the unit of analysis, Glass gives greater weight to studies with more dependent measures, producing an arbitrary bias. As a result, any report, even if it is atypical or of marginal quality, can have greater influence on meta-analytic findings if it uses many dependent measures. The Educational Research Service (1980), in a review of Glass' meta-analysis of class size on achievement, pointed out that 14 of the 76 studies were considered well-controlled. The 14 well-controlled studies produced 110 effect sizes, but 73% of the 110 came from four of the 14 studies. Heavy reliance on such a small number of studies eventually defeats the purpose of meta-analysis as a literature review.

In addition, the complex interdependencies introduced into the data by including multiple findings from a single study can drastically affect the standard errors of parameter estimates and inflate the Type I error rate (Strube, 1985). Hunter believes that the problem is not severe if the number of calculated effect sizes is not large relative to the number of studies. Rosenthal and others recommend performing separate meta-analyses for each type of dependent variable rather than lumping different types of outcome measures in a single analysis. Hedges provides statistical tests for homogeneity of correlated effect sizes. If the population values of the different effect sizes are the same, then optimal weights for pooling can be derived and effect sizes can be combined (Hedges & Olkin, 1985). In practice, some researchers solve the problem by simply taking the mean effect size (Steinkamp & Maehr, 1984; Willig, 1985). All agree that independence of effect size allows the reviewer to use statistics with more confidence.

Interest in nonparametric statistics is also growing. All of the above methodologies assume that effect size is normally distributed. Use of a nonparametric estimator of effect size may be desirable when the data are skewed, non-normal or contains outliers. Kraemer and Andrews (1982) point out limitations of the parametric effect size, which include the view that the interpretation of d depends on the implicit assumption that control group scores are normally distributed, that the treatment benefits all subjects equally, and that d is not invariant under all monotonic transformations of scales. However interesting a comparison of cumulated effect sizes based on parametric and nonparametric statistics might be, the computation of nonparametric effect size requires the raw data or at least the median from each study. These data are generally unretrievable and were not available for the present study.

Judging from current meta-analyses, research reviewers do not appear to prefer one methodology over another and in many cases, apparently have chosen not to take advantage of recent advances in statistical technique. A survey of effect size computation in 24 meta-analyses published in refereed educational journals from 1983 through 1986 showed that 14 used Δ, five used d, four used r, and one used w^2. Hedges' correction for bias was applied to effect sizes in five of the 24 studies. Only two studies mentioned the use of weighted effect sizes, both employing Hunter's frequency-weighted means. All of the examined meta-analyses contained studies with more than one effect size. Seven followed Glass and considered all data points as independent; seven weighted effect sizes by the reciprocal of the number of effect sizes from each independent sample; one used Tukey's jackknife technique to adjust for interdependencies in the data; two selected a single effect size from each group; and the remaining seven did not specify any method for handling multiple dependent measures.

The differences in meta-analytic method have been largely overlooked, and it is time that they be clarified so that the limitations of this approach to research integration can be more realistically assessed. As Bangert-Drowns (1984) points out, the differences should not be taken as evidence of some inherent weakness in meta-analysis; it is merely a reflection of the natural evolution of a new social scientific tool. As with any new method, a long process of rhetoric and empiricism is needed to achieve greater clarity.

CHAPTER BIBLIOGRAPHY

Bangert-Drowns, R. L. (1984). *Developments in meta-analysis: A review of five methods*. Ann Arbor, MI: Michigan University, Center for Research on Learning and Teaching. (ERIC Document Reproduction Service No. ED 248 262)

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

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

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Cooper, H. M. (1979). Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *Journal of Personality and Social Psychology*, *37*, 131-146.

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

Cooper, H. M., & Rosenthal, R. (1980). Statistical vs. traditional procedures for summarizing research find ings. *Psychological Bulletin*, *87*, 442-449.

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

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

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

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

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

Kraemer, H. C. (1983). Theory of estimation and testing of effect sizes: Use in meta-analysis. Journal of Educational Statistics, 8, 93-101.

Kraemer, H. C., & Andrews, G. (1982). A nonparametric technique for meta-analysis effect size calculation. Psychological Bulletin, 91, 404-412.

Light, R. J. (1984). Six evaluation issues that synthesis can resolve better than single studies. In W. H. Yeaton & P. M. Wortman (Eds.), Issues in Data Synthesis (pp. 57-73). San Francisco: Jossey-Bass.

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

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

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

Oliver, L. W., & Spokane, A. R. (1983). Research integration: Approaches, problems and recommendations for research reporting. Journal of Counseling Psychology, 30, 252-257.

Orwin, R.G., and Cordray, D. S. (1983). The effects of deficit reporting on meta-analysis: A conceptual framework and reanalysis. Unpublished manuscript, Northwestern University.

Reynolds, S. & Day, J. (1984, August). Monte Carlo studies of effect size estimates and their approximations in meta-analysis. Paper presented at the 92nd Annual Con vention of the American Psychological Association, Toronto, Ontario.

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

Rosenthal, R., & Rubin, D. (1982). Further meta-analytic procedures for assessing cognitive gender differences. *Journal of Educational Psychology*, *74*, 708-712.

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

Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher*, *15*(9), 5-11.

Steinkamp, M., & Maehr, M. L. (1984). Gender differences in motivational orientations toward achievement in school science: A quantitative synthesis. *American Educational Research Journal*, *21*(1), 39-59.

Strube, M. J. (1985). Combining and comparing significance levels from nonindependent hypothesis tests. *Psychological Bulletin*, *97*, 334-341.

Strube, M.J., & Hartman, D. P. (1983). Meta-analysis: techniques, applications and functions. *Journal of Consulting and Clinical Psychology*, *51*, 14-27.

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

CHAPTER III

PROCEDURES FOR THE STUDY

The Data Base

In order to render this study as realistic and practical as possible, an existing meta-analytic data base was sought which would conform to this author's criteria for size, timeliness and accuracy. After consultation with Robert Bangert-Drowns (Appendix A), the data base used by Chen-Lin C. Kulik and James A. Kulik in their meta-analysis of studies examining the effectiveness of computer-based education in colleges, was chosen (Kulik & Kulik, in press). The Kuliks have conducted a number of meta-analyses, primarily in the area of computer-assisted instruction, and have also published extensively on the methodology of meta-analysis.

The Kuliks' meta-analysis contained 101 studies that met the following inclusion criteria: (a) the studies took place in an actual college classroom and involved real teaching, (b) the studies provided quantitative results on an outcome variable measured in the same way in both a computer-taught and a conventionally-instructed class, and (c) the studies were free of such methodological flaws as substantial pre-treatment differences between groups, differential rates of subject attrition, or unfair teaching of the criterion measure to one of the comparison groups.

In order to compute the six different effect sizes used in this study, it was necessary to obtain the means, standard deviations and sample sizes of each experimental and control group used in the study, as well as the reliability coefficient of any criterion measure. In addition, the comparison of the cumulation of effect sizes techniques required that all relevant effect sizes be retrieved from each study. Since the Kuliks' meta-analysis computed only one effect size per study and generally applied only Glass' transformation for effect size, it was necessary to consult each of the 101 individual studies to retrieve the information not used by the Kuliks. This process resulted in a considerably smaller data base than used in the original meta-analysis. Of the 101 original studies, 17 were refused through interlibrary loan, four were not retrievable from Association of Higher Education Union Journals, and two were not published in any source. Of the remaining 78 studies, 26 did not contain the information necessary to compute all six effect sizes.

In order to collect the 60 individual research reports considered to be the minimum required for this study, it was necessary to seek additional reports outside of the Kuliks' data base. Since the Kuliks collected studies from 1970 through 1983, a search was made for reports published from

1984 to the present. Through Resources in Education, the Current Index of Journals in Education, and Dissertation Abstracts International, eight additional studies were retrieved. To maintain the integrity of the data base, only studies which conformed to the Kuliks' inclusion criteria were used.

Choice of Outcome Measure

Many studies reported more than one finding for a given outcome area, some as many as 16. Such findings sometimes resulted from the use of more than one experimental and control group in a single study, or from the use of several subscales to measure a single outcome. In some instances, several measurements were made of the same group over time, or the same study was conducted with different groups at different times. Since one aspect of this investigation concerned comparing the effect size and the study as units of analysis, all reported measurements in each research report were recorded, with the following constraints:

1. Only student learning scores, as measured by achievement instruments, were used for effect size computation. Other outcomes, such as attitude toward computers, attitude toward instruction, course completion rates or amount of time needed for instruction, were not included.

2. In almost all instances, only final status scores were used. This approach was taken for several reasons.

First, although some studies also reported raw or residual gain scores, or covariate-adjusted final status scores, almost all also reported the final status outcome. Thus, choosing the final status provided for a more consistent data base. Secondly, the formulas used in the effect size computations are those recommended by the authors for final status scores. The variance of derived gain measures contains confounded measurement error which can significantly bias results if not adjusted. Since gain scores express comparisons on a scale different from that used in randomized studies with only a final scale measurement, combining these in a single meta-analysis is not recommended. Finally, since the original inclusion criteria disqualified studies with obvious nonequivalent groups from the meta-analysis, any remaining pre-existing group differences should not substantially bias the results.

For those comparisons in this report using the study as the unit of analysis, a single effect size had to be chosen to represent the study. This "primary" effect size was chosen according to the following criteria:

1. When results from both a true experiment and quasi-experiment were available from the study, the results of the true experiment were chosen to represent the study.

2. Where retention measures were made on the same group, the latest measurement was used.

3. Where standardized test instruments were employed,

these were chosen over classroom tests.

4. Where total score results and subscores were reported, the total scores were used.

5. Where studies were conducted on several groups at different times, the most recent was chosen.

6. In cases where none of the above criteria applied, the first group scores reported by the author were arbitrarily selected to represent the study.

Some authors, particularly in the dissertation literature, reported the reliability of their criterion measures. Others used standardized instruments whose reliabilities were retrievable from Buros' Mental Measurement Yearbook. No reliability coefficients were available for 32 of the 60 studies in this data base. Therefore, in order to compute effect sizes corrected for unreliability of dependent measure, it was necessary to apply Hunter's procedure to the missing data. If reliability coefficients are only given sporadically, Hunter recommends using the average of the given reliabilities to correct the remaining effect sizes (Hunter, Schmidt & Jackson, 1982).

Effect Size Computation

The summary statistics from each research study were converted into each of the six effect sizes using the formulas provided below. Effect sizes were computed using a BASIC program written by the author.

Calculation of d

The effect size d used in this study is that derived from Cohen (1977) employing the within-group, or pooled standardizing denominator, as modified by Hedges (1985) for population estimate:

$$d = \frac{\overline{Y}_e - \overline{Y}_c}{S_p}$$

where $\overline{Y}_e$ and $\overline{Y}_c$ are the respective experimental and control group means, and S_p is the pooled sample standard deviation:

$$S_p = \sqrt{\frac{(n_e - 1)(s_e)^2 + (n_c - 1)(s_c)^2}{n_e + n_c - 2}}$$

where n_e, s_e, n_c and s_c are the respective sample size and standard deviations of the experimental and control groups.

Calculation of Δ

The computation of Δ is derived from Glass and employs the control group standardizing denominator:

$$\Delta = \frac{\overline{Y}_e - \overline{Y}_c}{S_c}$$

where S_c is the control group standard deviation.

Calculation of Unbiased Estimator g'

Hedges' correction factor for bias in effect sizes is defined as:

$$g' = \left(1 - \frac{3}{4N - 9} \right) d$$

where $N = n_e + n_c$ and d is as defined above.

Computation of Weighted Estimator g^w

According to Hedges (1985), the weights that minimize the variance of g^w give weights inversely proportional to the variance in each study. This leads to weighted estimators of the form:

$$g^w = w_i g_i + \ldots\ldots + w_j g_j$$

where g is defined as in g' above, and w_i and w_j are nonnegative weights that sum to one and are defined by:

$$w_i = \frac{1}{\sigma^2(g_i)} \Big/ \sum \frac{1}{\sigma^2(g_j)}$$

and $\sigma^2(g_i)$ is estimated by:

$$\sigma^2(g_i) = \frac{n_e + n_c}{n_e n_c} + \frac{g^2}{2(n_e + n_c)}$$

Computation of Weighted Estimator d^w

Hunter's frequency-weighted effect size $\underline{d}^w$ weights the individual study effect size by its sample size and divides by the sum of all the sample sizes:

$$d^w = \frac{\Sigma[N_i d_i]}{\Sigma N_i}$$

where $\underline{N}_i$ is the total sample size of the study, and $\underline{d}$ is as defined above.

Computation of d^r Corrected for Unreliability

Several authors recommend correcting effect size for measurement error, where information on the reliability of the dependent measure can be obtained.

$$d^r = \frac{d}{\sqrt{r_{xx}}}$$

where $\underline{d}$ is as defined above, and $\underline{r}_{xx}$ is the reported reliability of the measure.

Effect Size Cumulation

The second group of comparisons in this study examined the effect of using nonindependent effect sizes in the meta-analysis. For this comparison: (a) all effect sizes from each study were averaged together as if they were independent data points, (b) one effect size was chosen to represent the study, and (c) effect sizes were weighted by the reciprocal of the number of effect sizes from each independent sample, so that only one effect size per study contributed to the average.

Finally, in order to determine whether any difference in outcome may be a function of the size of the meta-analysis, repeated random samples of 20 and 40 studies each were drawn, and their properties compared to those of the complete set of 60 studies. The "population" data base chosen for this comparison was from method two, using the study as the unit of analysis. For consistency, Glass' effect size was used in all comparisons.

CHAPTER BIBLIOGRAPHY

Cohen, J. (1977). _Statistical power analysis for the behavioral sciences_ (2nd ed.). New York: Academic Press.

Hedges, L. V., & Olkin, I. (1985). _Statistical methods for meta-analysis_. Orlando: Academic Press.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). _Meta-analysis: Cumulating research findings across studies_. Beverly Hills: Sage Publications.

Kulik, C.-L. C., & Kulik, J. A. (in press), Effectiveness of computer-based education in colleges. _AEDS Journal_.

CHAPTER IV

PRESENTATION AND ANALYSIS OF DATA

Introduction

This study was conducted in order to assess the effects of employing the different methodologies for computing and cumulating effect sizes on the outcomes of a meta-analysis. The six effect sizes computed were: (a) Cohen's $\underline{d}$, (b) Glass' Δ, (c) Hedges' unbiased $\underline{g}'$, (d) $\underline{d}^r$ corrected for unreliability of measurement, (e) Hunter's frequency-weighted $\underline{d}^w$, and (f) Hedges' variance-weighted $\underline{g}^w$. The 60 studies used for the data base produced a total of 201 effect sizes. In order to determine whether different mean effect sizes would emerge from the different cumulation techniques, these effect sizes were averaged in three different ways: (a) using the effect size as the unit of analysis ($\underline{k}$=201), (b) Using one "primary" effect size from each study ($\underline{k}$=60), and (c) using the mean of the effect sizes within the study ($\underline{k}$=60). Two smaller samples of 20 and 40 studies each were drawn at random from the population of 60 studies to determine whether outcomes would vary with the size of the meta-analysis. The results of these three investigations are reported below.

Analysis

Research Question 1

Do the six methods of computing effect size, measure study outcomes in similar ways?

Table 1 shows the mean effect sizes for each of the six methods of computing effect sizes. Not unexpectedly, the effect size corrected for unreliability of measure ($\underline{d}^r$), produced the largest mean effect size. To the extent

Table 1

Mean Effect Sizes for Three Methods of Cumulation

	Unit of Analysis		
Mean Effect Size	Effect Size $\underline{k}$=201	Study $\underline{k}$=60	Average Effect $\underline{k}$=60
d^r	.382 (.050)	.526 (.109)	.521 (.103)
d	.343 (.045)	.470 (.098)	.465 (.092)
g'	.336 (.045)	.460 (.096)	.457 (.090)
Δ	.327 (.047)	.446 (.101)	.448 (.093)
d^w	.237 (.050)	.431 (.113)	.402 (.113)
g^w	.207 (.042)	.347 (.081)	.323 (.082)

Note. Standard errors in parentheses.

that the measurement of the dependent variable is less than perfectly reliable, errors of measurement will cause the observed _d_ value to be an underestimate of the actual effect size. Thus, correcting the effect size for the known reliability of the dependent measure will necessarily raisethe size of the effect, at least slightly. As suggested by Hunter (1982), effect sizes based on dependent measures whose reliabilities were unknown were corrected by the average reliability of the other dependent measures in the data base. In this data base, the average reliability coefficient was .83.

Effect sizes computed from Cohen, Glass and Hedges' formulas yielded very similar means. Cohen's _d_, using the pooled within-group standardizing denominator, produced a slightly larger average effect size than Glass' formula using only the control group standard deviation. On the average, Hedges' unbiased _g_' lowered the estimate of _d_ by .01 or less. The literature showed some disagreement as to what size sample would benefit from Hedges' correction for bias. In this data base, Hedges' correction produced an effect size identical to _d_ carried to three decimal places until study sample size fell below 100. Differences of .03 or more were found only in sample sizes smaller than 20.

The two weighted effect size formulas produced the smallest metrics, with Hedges' estimator weighted by study variance (g^w) yielding the smallest effect size. In Table

1, the difference between the largest (*d*^{r}) and smallest (*g*^{w}) means was as much as .20. This would suggest that effect sizes from larger samples, which receive greater weight under Hunter's and Hedges' formulas, tend to be smaller than those from small samples. This was verified by correlating the sample size in individual studies with the absolute value of their effect sizes. The results, shown in Table 2, confirm that sample size and effect sizes for the four unweighted metrics were negatively correlated; i.e., the greater the number of subjects in the study, the smaller the effect size. Since sample size has already been taken into account in computing the effect size for the weighted metrics (*d*^{w} and g^w), these show a high positive correlation with sample size in Table 2.

Table 2

Correlation Between Sample Size and Absolute Effect Size

k = 201

	d	Δ	g'	d^r	d^w	g^w
Sample size	-.146	-.164	-.139	-.146	.462	.546
	p=.04	p=.02	p=.05	p=.04	p<.01	p<.01

Of the two weighted formulas, Hedges g^w showed the smaller mean effect size in all three study sets. Both weighted effect sizes take into account the size of the sample, but Hedges' weighted formula also uses the sizes of the individual groups $\underline{n}_e$ and $\underline{n}_c$, as well as the unweighted effect size. Hedges' formula assigns weights that are inversely proportional to the variance in each study. Inspection of that formula (p.39) shows that the effect size variance is at a minimum in large samples with equal group $\underline{n}$'s and small effect sizes. Therefore, $\underline{g}^w$ will show its greatest departure from $\underline{d}^w$ when effects are large and group sizes are radically different.

Outliers are frequently seen in meta-analysis and their potential influence on the mean effect size is cause for concern. In this data base, the mean sample size was 80, whereas the median was 50, indicating the presence of large sample size outliers. In order to assess the effect of these outliers on the weighted effect sizes, these were recomputed in several different subsets of the complete set of 201 effect sizes. Elimination of the three largest samples (those with more than 400 subjects), raised the mean weighted effect sizes by .04 and the unweighted means by approximately .005. The only time that the weighted and unweighted effect sizes showed essentially the same means was in a subset consisting of studies with less than 50 subjects.

Correlations among the six effect sizes are shown in Tables 3, 4 and 5 for the three methods of data cumulation. Consistent with the findings of Bangert-Drowns, Kulik and

Table 3

Intercorrelations Among Effect Sizes

Unit of Analysis: Effect Size k = 201

Effect Size	d	Δ	g'	d^r	d^w	g^w
d	1.000	.987	.999	.998	.705	.680
Δ		1.000	.986	.987	.706	.673
g'			1.000	.998	.712	.686
d^r				1.000	.702	.677
d^w					1.000	.968
g^w						1.000

Note. p<.001 for all correlations.

Table 4

Unit of Analysis: Study k = 60

Effect Size	d	Δ	g'	d^r	d^w	g^w
d	1.000	.985	.999	.998	.695	.717
Δ		1.000	.985	.987	.712	.728
g'			1.000	.998	.705	.726
d^r				1.000	.694	.715
d^w					1.000	.978
g^w						1.000

Note. p<.001 for all correlations.

Table 5

Unit of Analysis: Average Effect k = 60

Effect Size	d	Δ	g'	d^r	d^w	g^w
d	1.000	.987	.999	.998	.723	.731
Δ		1.000	.987	.991	.746	.749
g'			1.000	.998	.730	.737
d^r				1.000	.719	.727
d^w					1.000	.977
g^w						1.000

Note. $p<.001$ for all correlations.

Kulik (1983), the correlation between Glass' Δ and Hedges' unbiased g' is .99 in all study sets. The correlation matrices for all three data sets show high intercorrelation among d, Δ, g' and d^r, but lower correlation of these metrics with the weighted effect sizes; whereas the weighted estimates showed a .97 correlation with each other. In order to examine whether the weighted and unweighted effect sizes might represent different underlying factors, a factor analysis was performed. Given a criterion of two factors, principal-axes factor analysis produced the rotated factor matrix shown in Table 6. Because the intercorrelation matrix showed that the individual effect size variables were substantially related, however, oblique factor solutions

Table 6

Rotated Factor Matrix: Varimax Rotation

Effect size	Factor 1	Factor 2
d^r	.926	.376
d	.926	.378
g'	.921	.387
Δ	.910	.382
g^w	.360	.913
d^w	.393	.905

Table 7

Factor Pattern Matrix: Oblique Rotation

Effect size	Factor 1	Factor 2
d^r	1.005	-0.009
d	1.004	-0.006
g'	0.994	0.008
Δ	0.982	0.007
g^w	-0.022	0.996
d^w	0.025	0.969

were also sought. Table 7 reports the pattern matrix for the oblique rotation. The factors correlated .705 with each other. Thus, the factors representing the weighted and unweighted effect sizes do not define subsets of variables which are conceptually different. Confirmatory factor analysis could not be performed on this data because the two-factor model was not identified for only six variables.

Finally, the differences among the six different methods of computing effect size were examined by constructing 95-percent confidence intervals for each mean, shown in Figure 1. When considered with their confidence bands, none of the effect sizes differed reliably from each other.

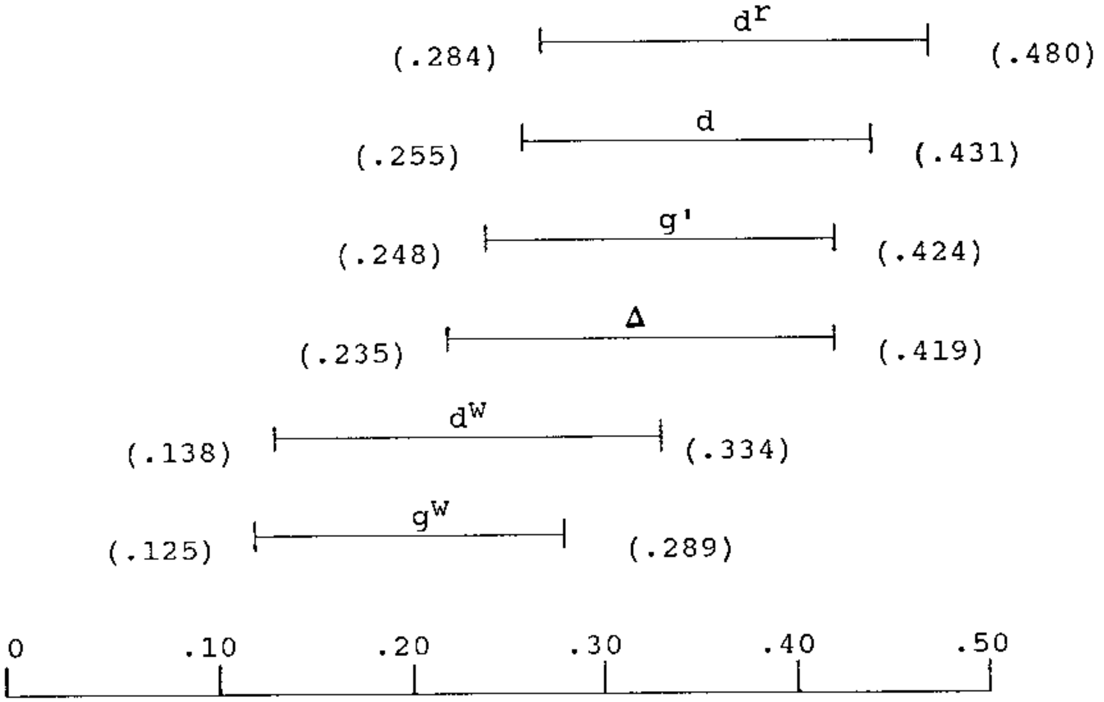

Figure 1. 95-percent confidence intervals for mean effect sizes.

Research Question 2

Do the three methods of cumulating effect sizes across studies measure study outcomes in similar ways?

Table 1 shows the mean effect sizes for each of the three methods of cumulating effect sizes: (a) using the effect size as the unit of analysis, (b) using one primary effect size to represent the study, and (c) using the simple mean of all effect sizes within the study to represent the study. Method one employed 201 effect sizes while methods two and three used 60 effect sizes each.

Larger mean effect sizes were produced by using the study or average effect as the unit of analysis (methods two and three), rather than the effect size itself (method one). Thus, in this data base, the most representative outcome in each study (as defined in Chapter III) tended to be larger than other effect sizes produced by the study.

Using the effect size as the unit of analysis gives greater weight to studies with more dependent measures, producing an arbitrary bias. This data base contained an average of 3.35 effect sizes per study, with the largest study yielding 16 effect sizes. While not as lopsided as Glass' meta-analysis of class size and achievement, where 80 of the "most valid" effect sizes came from four of the studies (Educational Research Service, 1980), this data base contains a rather large number of calculated effect sizes relative to the number of studies. Hunter and others warn that this situation can create serious bias in the outcome of meta-analysis (Hunter et al., 1982). When the data were reanalyzed without the three studies containing more than twelve effect sizes apiece, the resulting mean effect sizes were changed by less than .02. Thus, in this data base, the meta-analysis was not seriously biased by the inclusion of studies with many outcome measures.

When arranged according to their magnitude, the order of the six different methods of computing effect size did not change with the methods of cumulating the effect sizes. Regardless of the unit of analysis, $\underline{d}^r$ showed the largest mean effect size and $\underline{g}^w$ the smallest. Intercorrelations among the effect sizes remained fairly constant across all three methods of cumulation (Tables 3, 4, and 5).

The factor analyses shown in Tables 6 and 7 were repeated using the effect sizes from methods two and three. Results showed the same configuration of factors as with method one, and almost identical factor loadings.

Figure 2 depicts the effect sizes with their 95-percent confidence intervals for each method of cumulation. When confidence bands are included, the mean effect sizes cumulated under each of the three methods do not differ reliably from each other for any of the six methods of computing effect size.

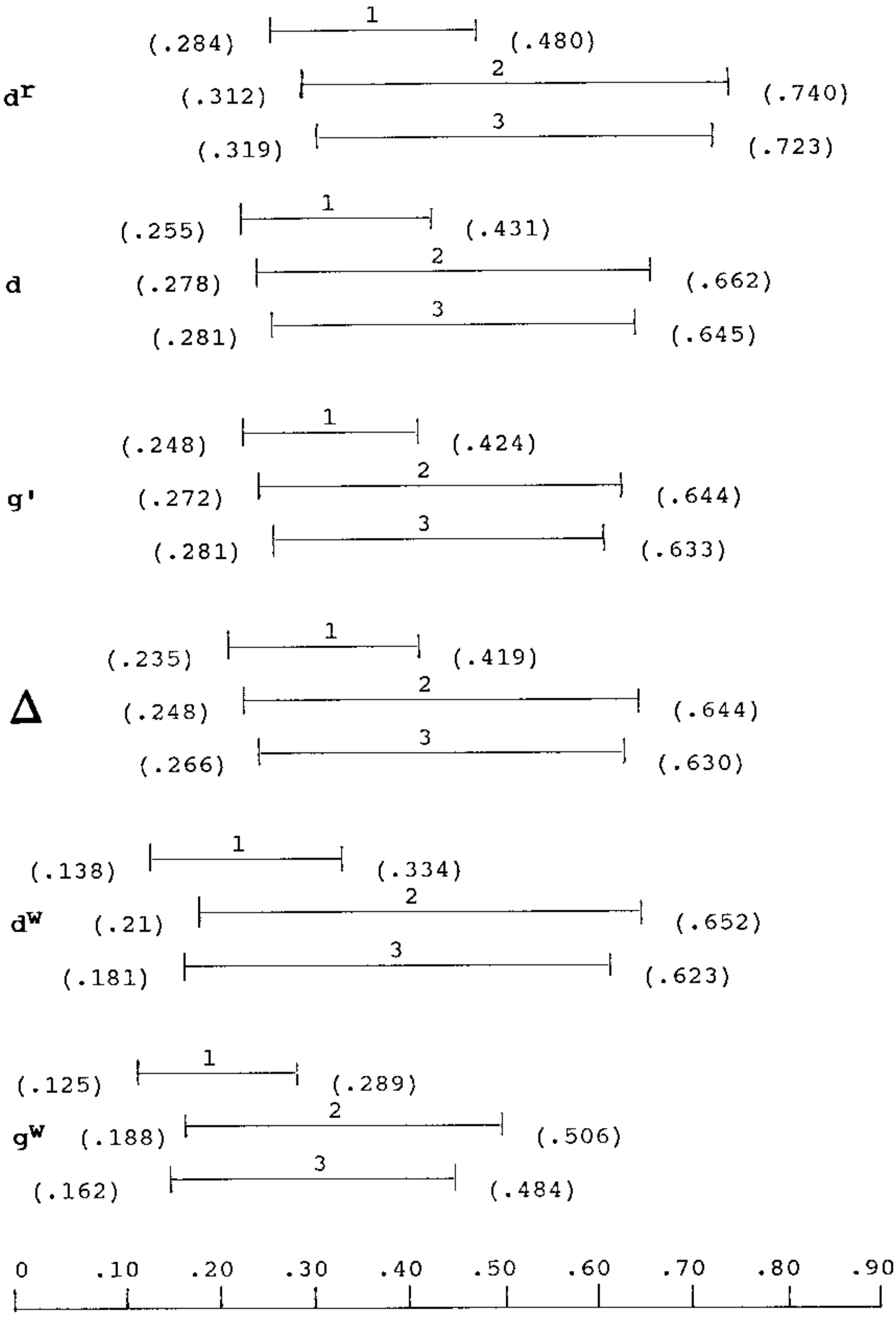

Figure 2. 95-percent confidence intervals for mean effect sizes by unit of analysis.

Research Question 3

Will a random sampling distribution of effect size means approximate the population distribution for smaller sample sizes?

Theoretically, meta-analysis can be performed on any number of studies, and it is not uncommon for as few as 20 studies to constitute the data base for a meta-analysis. In the current study, 60 individual reports were cumulated to produce the mean effect size. It was theorized that using fewer studies in the data base might produce results that depart significantly from those given by the complete set of 60 studies.

In order to assess the odds of choosing a sample representative of the population data base of 60 studies, repeated random samples of 20 and 40 studies each were drawn, and their properties compared to those of the population data base. The population data base chosen for this comparison was from method two, using the study as the unit of analysis. For consistency, Glass' effect size Δ was used in all comparisons. The population mean was .446, with a positive skew (.709) and kurtosis (.792).

Table 8 shows statistics descriptive of the distributions of the smaller sample sizes. A Kolmogorov-Smirnov one-sample test was performed on the sample data, using the mean and standard error associated with the population of 60 studies as a basis for comparison. For the samples

Table 8

Comparison of Population and Sample Distributions

Sample Size	No. of Samples	Mean	Skew	Kurtosis	K-S z	p*
Population		.446	.709	.792		
		(.101)	(.309)	(.608)		
40	30	.445	.415	.991	.864	.444
		(.016)	(.427)	(.883)		
20	60	.467	-.412	-.220	1.675	.007
		(.018)	(.309)	(.608)		

Note: Standard errors in parentheses.

*two-tailed.

of 40, the H_0 was not rejected, indicating that the observations could reasonably have come from the same distribution. Such was not the case, however, for the sample size of 20. Although the mean of .467 approximated the population mean, the distribution of means was flatter than the population distribution and negatively skewed (Figure 3).

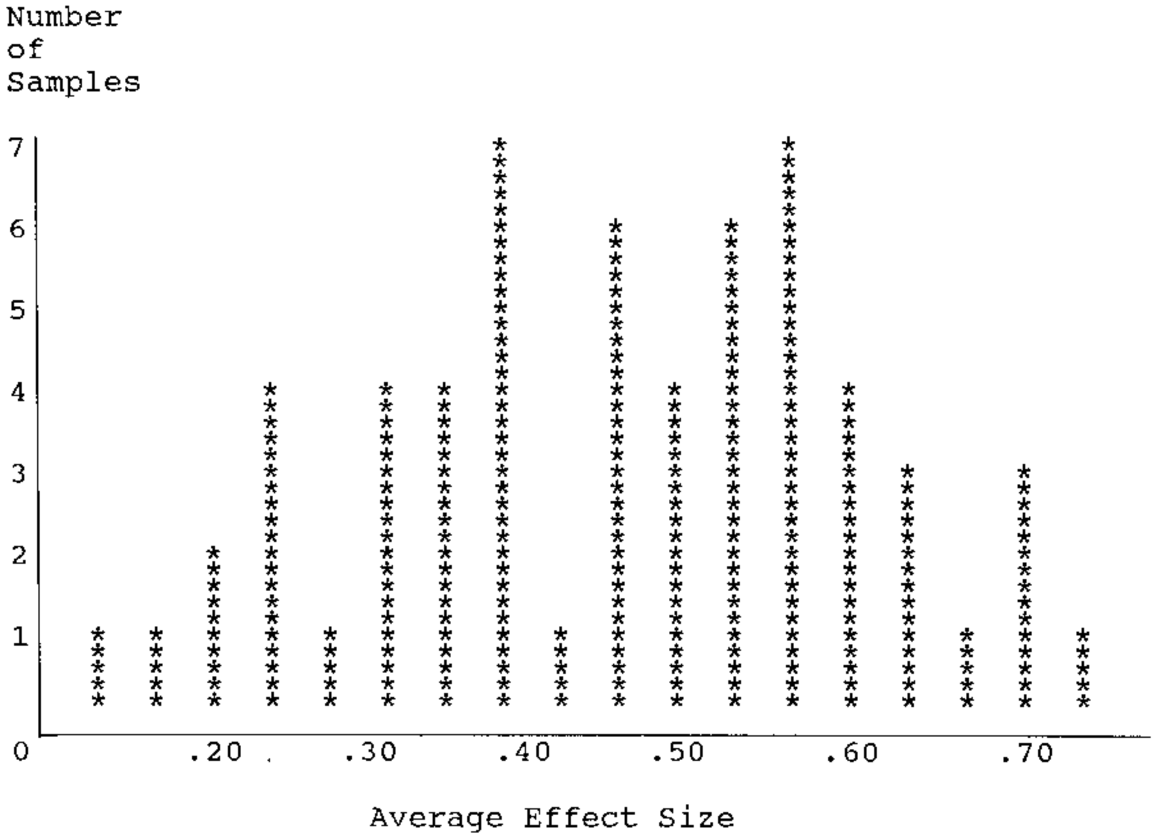

Figure 3. Distribution of sample means for samples of 20 studies each.

CHAPTER BIBLIOGRAPHY

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

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

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

CHAPTER V

FINDINGS, CONCLUSIONS AND RECOMMENDATIONS FOR RESEARCHERS

Findings

Research Question 1

Do the six methods of computing effect size measure study outcomes in similar ways?

Two distinct groups of effect sizes emerged from this analysis, one composed of the unweighted effect sizes and the other of the weighted effect sizes. The four unweighted metrics d, Δ, g' and d^r measure study outcomes in very similar ways. The high correlation (.99) between d and Δ in all study sets confirms Hedges' and Olkin's assertion that equal population variances in experimental and control groups can generally be assumed. Hedges' correction for the bias of d, incorporated in the metric g', lowered the estimate of d by no more than .01 in the three study sets. As suggested by other researchers, d appears to be effectively unbiased except in very small samples. In this data base, differences of .03 or more between d and g' were found only in studies using fewer than twenty total subjects. When d is corrected for unreliability of the dependent measure, the resulting effect size will necessarily be higher than the uncorrected metric. The amount of difference will depend on the reliability of the measurement instrument. Although Hunter and his colleagues recommend using the average of known reliabilities to correct effect sizes whose reliabilities are unknown, they do not specify the percentage of reliabilities in a data base which should be known before applying the average to all other effect sizes. Thus, the researcher interested in determining the improvement in average effect size achieved through correction for unreliability of measure when few or no reliabilities are known must make a judgment as to the probable reliability of measuring instruments. Happily, the absence of a significant difference between the mean $\underline{d}^r$ and other unweighted mean effect sizes in this investigation suggests that some unreliability of measures, within acceptable limits, is unlikely in itself to bias the results of the meta-analysis. Although the literature on tests and measurements does not provide any standard reliability coefficient for deciding whether or not a test is reliable, it does give indications that a reliability coefficient of .80 or higher is evidence that a test is adequately reliable for classroom use. In this meta-analysis, the average reliability of the measuring instruments was .83.

The main consequence of employing one of the weighted formulas is to magnify the study's effect size in large samples and attenuate it in small samples. For example, the one-point posttest difference between experimental and control groups in Saul's study produced an effect size of $\underline{d} = -0.28$ (Saul, 1975). However, the study used 510 subjects, resulting in weighted effect sizes of -1.55 ($\underline{d}^{w}$) and -2.09 ($\underline{g}^{w}$). Likewise, in Homeyer's study, the effect $\underline{d} = 1.39$ was reduced to 0.16 in the weighted formulas because the sample size was only ten (Homeyer, 1970).

In this particular data base, the weighted and unweighted metrics formed two distinct groups, at least from a standpoint of practical importance if not statistical significance. Would this condition hold true in other meta-analytic data bases? Weighted effect sizes reflect the characteristics of their data bases. The distribution of effect sizes in this meta-analysis is positively skewed, indicating that larger frequencies were found among the smaller effect sizes. Smaller effect sizes were more likely to be associated with larger sample sizes, and both of the weighted effect sizes depend primarily on the sample size of the study with which they are associated. Thus, weighted effect sizes may or may not be smaller in other data bases, depending on the size of effects contained in the studies having the largest samples.

If weighted effect sizes are to be used, care must be taken to avoid giving too much weight to studies which may have large sample sizes but not necessarily be the most meaningful or theoretically representative. In the example from Saul's study discussed above, -2.09 is a large effect size and probably not justified by a one-point difference between experimental and control groups on a single dependent measure. Although not the case in this particular meta-analysis, in the aggregate these outliers can unduly influence the mean effect size, particularly in a meta-analysis with fewer studies. Avoiding the pitfalls of large but nonrepresentative studies while recognizing the greater precision of larger samples may well lead the meta-analyst to a technique characterized by Slavin as "Best-Evidence Synthesis" (1986). Rather than include virtually all studies on a given topic, best-evidence synthesis summarizes only those studies that meet strict inclusion criteria for germaneness, external validity and methodological adequacy, considered the "best evidence" on the topic. Pooled effect sizes are reported as adjuncts to the literature review, not its primary outcome. Best-evidence seeks to incorporate the quantification and systematic literature methods of meta-analysis with the detailed analysis of critical issues and study characteristics of the best traditional reviews.

Research Question 2

Do the three methods of cumulating effect sizes across studies measure study outcomes in similar ways?

Larger mean effect sizes emerged from using the study or average effect size as the unit of analysis, rather than the effect size itself. Although confidence intervals did not show a statistically significant difference among the methods of cumulation, the difference in mean effect size of approximately .15 to .20 between method one and methods two or three could be of theoretical or practical importance. The largest mean effect sizes came from method two, where a single effect size was chosen to represent the study. This is the cumulation strategy espoused by most meta-analysts other than Glass. In most cases, this "most representative" study outcome was not chosen arbitrarily; it was in some way the best measure of the treatment's effectiveness, and so is considered the "truest" outcome. Method three, using the average effect size as the unit of analysis, could be considered the method of compromise. It makes use of all study outcomes while still maintaining the independence of the data. The average effect sizes produced by method three were, for the most part, only slightly lower than those of method two.

Regardless of the method of cumulation, the order of magnitude of the mean effect sizes for the six different methods of computation were constant. Nor were mean effect sizes unduly influenced by those studies with a large number of effects.

The lower mean effect size in method one may confirm Glass' reasoning that nonindependent data errs on the conservative side. Concerns expressed by Strube (1985) that using the effect size as unit of analysis will drastically inflate the Type I error rate appear unfounded for this particular meta-analysis. The lowest effect sizes were produced by this method, suggesting that the possibility of finding effects where none exist is more likely with methods two and three. The overall effect of nonindependence of data was to lower the mean effect size, although not significantly so.

Research Question 3

Will a random sampling distribution of effect size means approximate the population distribution for smaller sample sizes?

The central limit theorem dictates that the accuracy of estimation improves as $\underline{N}$ increases. For $\underline{N} \geq 30$, the sampling distribution of means is approximately a normal distribution irrespective of the population so long as the population mean and variance are finite and the population size is at least the sample size. If the population is normally distributed, the sampling distribution of means is also normally distributed, even for small values of $\underline{N}$ (Blalock, 1979).

The sampling distributions of means from this meta-analysis behaved as would be expected under the central limit theorem. For sample sizes of 40, the distribution of means approached the normal distribution, even though the population from which it was drawn showed a positive skew and kurtosis.

The smaller sample size of 20, however, departed significantly in its sampling distribution from the population. Since the population was not normally distributed, it cannot be expected that samples of less than 30 will show a normal distribution of means.

The properties of these sampling procedures depend on the availability of unrestricted samples of effect size estimates. Unfortunately, there is often reason to believe that nonrepresentative sampling is prevalent in research synthesis (Hedges, 1985). This is the well-known problem of publication bias, where only research results that are statistically significant are reported, and the effect size estimates that correspond to nonsignificant mean differences are not available for inclusion in the meta-analysis. Evidence of publication bias can be seen in this study itself. The Kuliks' meta-analysis of the effectiveness of computer-based education, which served as a source for this investigation, used 101 studies and found an overall average effect of .26 favoring CBE. The corresponding effect size computed in this meta-analysis from 60 studies was .45 (Glass' Δ, study as unit of analysis; see Table 1). In the Kuliks' study, however, the average effect was .42 for studies found in professional journals, .16 in the dissertation studies, and .11 in unpublished technical reports. Whereas nearly all of the published studies were located for this data base, fewer dissertations and almost none of the unpublished technical reports were included; hence, the higher average effect. The problem of selective sampling of studies coupled with sampling error renders meta-analyses using fewer than 20 studies of questionable validity, due to the large potential for systematic and unsystematic error.

Conclusions

<u>Research Question 1</u>

Do the six methods of computing effect size measure study outcomes in similar ways?

In this investigation, the six methods of computing effect size did not produce significantly different results. The weighted formulas gave smaller mean effect sizes because in this particular data set, studies with larger sample sizes tended to have smaller effects. Weighted effect sizes reflect the nature of their data base; therefore, this phenomenon may or may not hold true in other meta-analyses.

<u>Research Question 2</u>

Do the three methods of cumulating effect sizes across studies measure study outcomes in similar ways?

Maintaining the independence of the data by using only one effect size per study, whether a single or averaged effect size, produced a higher mean effect than averaging all effect sizes together, although the difference did not reach statistical significance. Regardless of the method of cumulation, the order of magnitude of the mean effect sizes for the six different methods of computation were constant.

Research Question 3

Will a random sampling distribution of effect size means approximate the population distribution for smaller sample sizes?

In this meta-analysis, the sampling distribution of effect size means approached that of the population of 60 studies for samples consisting of 40 studies, but not for samples of 20 studies. This behavior is consistent with the central limit theorem, which specifies that the accuracy of approximation improves as the sample size increases.

Recommendations for Researchers

The purpose of the present study was to examine the degree to which the researcher's choice of method for effect size computation and combination would affect the outcome of the meta-analysis. It was hoped that the information gained from this investigation would assist authors and consumers of meta-analysis in making more informed decisions. Accordingly, I have included several suggestions regarding choice of effect size and cumulation technique, based on the information resulting from this study.

1. For an effect size metric, the researcher may choose among Δ, d, g' or d^r without fear of biasing the outcome of the meta-analysis.

2. Correcting for unreliability of measure will not appreciably alter the results of the meta-analysis unless the measurement instruments are dramatically unreliable, in which case one would probably not be interested in the study in the first place.

3. If weighted effect sizes are to be used, care must be taken to avoid giving too much weight to studies which may have large sample sizes but not necessarily be the most meaningful, theoretically representative, or elegantly designed.

4. Although outcomes did not differ significantly with method of effect size cumulation, most researchers favor maintaining independence of data; i.e., using only one effect size per study. Where effect sizes can be grouped conceptually, the method of using the average of all effects to represent the study is appealing since it makes use of all study outcomes while maintaining independence of the data.

5. It is important for the meta-analyst to locate all relevant studies on the topic under investigation, since selective or even random sampling will bias the results of the meta-analysis. This is particularly true in data sets containing 20 or fewer studies.

CHAPTER BIBLIOGRAPHY

Blalock, H. M. (1979). Social statistics. New York: McGraw-Hill.

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

Homeyer, F.C. (1970). Development and evaluation of an automated assembly language teacher. Austin, TX: Univer sity of Texas. (ERIC Document Reproduction Service No. ED 053 531)

Saul, W.E. (1975). An experimental study of the effects of computer-augmented instruction on achievement and attrition in beginning accounting at Miami-Dade Community College, North Campus. Dissertation Abstracts Interna tional, 35, 4757A. (University Microfilms No. 75-363A)

Slavin, R. E. (1986). Best-evidence synthesis: An alter native to meta-analytic and traditional reviews. Educational Researcher, 15(9), 5-11.

APPENDICES

APPENDIX A

LETTER FROM R. L. BANGERT-DROWNS

The University of Michigan
Center for Research on Learning and Teaching
109 East Madison Street
Ann Arbor, Michigan 48109

November 10, 1986

Sharon Ronco
2717 South Llewellyn
Dallas, Texas 75224

Dear Sharon,

As we discussed in our recent telephone conversation, I am sending you a package of information which I hope will be helpful to you in your research. I have enclosed a copy of an article I recently published in Psychological Bulletin outlining what I believe are the chief differences among contending types of meta-analysis. I believe that each approach has something positive to offer, that none of them are fully developed yet, and that there are serious problems in some aspects of these approaches. Some of my concerns are outlined in the pages I've copied from my dissertation. My colleagues, the Kuliks, have also recently completed a paper describing other potential problem areas in meta-analytic work, and I am forwarding a copy of their paper to you too.

You asked me to recommend a meta-analysis in any area of education that includes at least 60 studies. I am listing five possible such meta-analyses; hopefully, one of them will be useful to you. If not, let me know and I could recommend others. As a general suggestion, dissertations often report more detailed information than published articles, so they may be easier to use if you intend to do a reanalysis. Here is the list:

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

Curbelo, J. (1984). Effects of problem-solving instruction on science and mathematics student achievement: A meta-analysis of findings. Dissertation Abstracts International, 46, 23A.

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

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.

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

You'll notice that some of these meta-analyses review the same area. I recommended them so you could see how replication succeeds or fails in meta-analysis. I am also including a recent article by the Kuliks as a meta-analysis for your consideration.

This is an exciting time to be working in the area of meta-analysis. There's a lot of interest in the method, and a lot of disagreement over how it should best be done. Maybe your investigation can shed some new light on the issues.

Let me know if I can be of further help. I am especially interested to know what you decide to do and what your results are, so please keep in touch.

Sincerely,

Robert L. Bangert-Drowns

APPENDIX B

DESCRIPTIVE DATA FROM STUDIES USED IN META-ANALYSIS

Table 9

Descriptive Data from Studies Used in Meta-Analysis

Author	$\overline{Y}_E$	$\overline{Y}_C$	N_E	N_C	S_E	S_C	r_{xx}
Alderman	26.67	24.50	21	68	4.3	6.3	0.78
(English study)	28.89	28.13	28	30	6.1	5.8	0.81
	27.63	25.08	82	149	5.1	6.1	0.79
	27.07	27.28	41	46	7.6	6.4	0.86
	2.36	2.14	21	66	0.7	0.8	
	2.18	1.95	28	30	0.7	0.7	
	2.14	2.03	77	157	0.7	0.6	
	1.96	2.34	40	46	0.7	0.8	
	32.18	29.23	57	39	5.6	6.0	0.84
	22.83	31.12	84	164	9.2	7.0	0.92
	28.75	28.82	122	84	6.7	5.9	0.84
	2.72	2.07	58	38	0.8	0.8	
	2.10	2.68	83	165	0.8	0.8	
	2.42	2.27	113	83	0.8	0.8	
Alderman	44.26	38.58	27	155	6.7	10.4	0.92
(Math study)	40.57	38.58	35	146	6.9	8.5	0.86
	36.41	30.33	34	67	4.6	7.6	0.85
	35.29	32.14	17	91	6.5	8.7	0.86
	40.50	36.73	10	44	7.6	8.7	0.88
	39.25	39.63	12	43	8.3	8.3	0.85
	24.71	24.54	14	57	7.7	9.2	0.90
	45.33	38.19	9	42	6.2	5.9	0.90
	40.00	36.76	25	21	7.6	5.8	0.91
	40.36	40.73	11	11	8.0	6.6	0.94
	21.00	20.62	2	13	5.7	5.3	0.88
	25.67	18.62	6	26	8.0	5.3	0.86
Andrews	159.94	161.28	17	21	25.31	29.16	0.94
	166.77	168.18	17	21	20.09	28.61	0.94
	151.44	150.72	17	21	27.77	29.49	0.94
Axeen	8.90	10.00	32	34	6.22	6.58	0.86
Boen	6.44	5.38	16	16	0.73	1.15	
Boyson	84.72	77.06	18	18	11.90	12.32	

Table 9--*Continued*

Author	$\overline{Y}_E$	$\overline{Y}_C$	N_E	N_C	S_E	S_C	r_{xx}
Brum	2.87	2.42	38	32	0.78	0.65	
Cartwright	65.59	52.78	27	87	4.68	5.89	
Caruso	14.32	14.25	37	50	1.14	1.22	0.82
	23.14	22.90	37	50	3.55	2.26	0.82
Castleberry	75.00	41.30	99	99	18.40	15.80	0.78
Conklin	55.40	43.30	13	12	9.50	6.20	0.55
Crawford	20.52	20.01	319	64	4.11	4.98	
	20.45	18.86	319	64	3.55	4.15	
	6.19	6.97	319	64	1.83	2.25	
Daughdrill	20.35	20.00	34	32	4.24	5.79	0.84
Diem	58.30	68.50	11	14	18.00	12.30	0.96
	59.90	68.50	14	14	13.80	12.30	
	67.50	68.50	14	14	21.10	12.30	
DuBoulay & Howe	22.50	21.80	6	6	5.08	3.87	
	9.50	7.83	6	6	3.93	3.96	
	26.17	28.20	6	6	4.89	4.32	
Durgin	11.92	11.58	40	38	3.16	3.17	0.57
	12.05	11.58	38	38	2.93	3.17	0.57
	8.57	8.53	40	38	2.69	2.98	0.60
	9.05	8.53	38	38	3.22	2.98	0.60
	7.97	8.05	40	38	3.58	2.90	0.69
	8.63	8.05	38	38	3.27	2.90	0.69
	7.10	7.18	40	38	2.64	3.19	0.65
	8.13	7.18	38	38	3.43	3.19	0.65
Fiedler	26.50	26.95	24	24	3.86	4.40	
	16.74	17.53	19	19	4.32	4.33	
	22.70	21.00	20	20	3.91	3.46	
	21.00	19.28	18	18	4.12	4.84	
Friesen	14.07	14.43	62	75	3.96	3.74	0.85

Table 9--Continued

Author	$\overline{Y}_E$	$\overline{Y}_C$	N_E	N_C	s_E	s_C	r_{xx}
Grandey	15.23	10.46	13	13	7.41	5.50	
	16.39	13.46	13	13	2.60	3.02	
	13.85	11.38	13	13	2.27	4.91	
	13.00	7.23	13	13	5.99	3.68	
Gray	31.30	30.10	44	25	5.00	5.10	
Green	38.30	35.70	10	10	3.63	5.44	
	35.30	36.70	6	7	6.24	3.35	
	24.40	27.90	10	15	3.71	4.02	
	24.40	23.80	9	9	6.69	5.51	
Hartig	21.30	17.75	23	139	3.71	4.03	
Henry & Ramsett	18.80	18.00	310	110	4.47	4.47	0.76
Hofstetter	86.00	75.00	17	16	12.40	14.40	
	83.00	75.00	17	16	13.30	16.10	
Holoien	70.00	74.00	15	14	15.00	12.00	0.90
	79.00	77.00	15	14	10.00	7.00	0.88
	77.00	68.00	14	16	10.00	15.00	0.88
	72.00	77.00	15	14	14.00	11.00	0.89
	76.00	67.00	14	16	10.00	13.00	0.86
	66.00	75.00	15	14	18.00	15.00	0.94
	69.00	60.00	14	16	19.00	18.00	0.95
	65.00	59.00	14	16	21.00	19.00	0.95
Homeyer	90.72	88.44	5	5	1.55	1.73	
	87.14	91.20	5	5	12.66	7.82	
	89.10	88.74	5	5	10.51	10.25	
	92.70	94.14	5	5	7.99	1.53	
Hong	77.625	66.636	8	11	8.65	11.59	
	58.900	49.643	10	14	14.36	11.45	
	65.158	59.563	19	16	10.75	17.31	
	61.556	59.563	9	16	17.14	17.31	
Hu & Saunders	82.19	78.85	48	48	4.42	4.56	
	59.00	52.27	48	48	5.92	5.90	
Huckabay et al.	8.36	7.17	14	17	0.93	1.78	0.87
	10.07	8.59	14	17	1.44	2.21	0.91

Table 9--Continued

Author	$\bar{Y}_E$	$\bar{Y}_C$	N_E	N_C	S_E	S_C	r_{xx}
Hughes	39.60	38.40	45	46	5.52	6.94	0.94
Jim	73.80	72.80	36	30	11.00	13.00	
Johnson et al.	19.10	17.10	35	30	3.31	4.00	
	18.90	17.10	37	30	3.15	4.00	
	16.40	17.10	28	30	2.50	4.00	
	22.20	19.40	35	30	4.90	2.70	
	23.30	19.40	37	30	2.80	2.70	
	19.60	19.40	28	30	2.80	2.70	
Lang	21.08	19.08	48	36	4.48	4.12	0.84
	15.44	13.80	48	36	6.62	6.83	0.80
	13.38	11.50	48	36	4.24	4.48	0.82
	14.77	14.61	48	36	4.48	4.12	0.87
	22.78	19.74	36	27	5.96	7.31	0.84
	16.44	14.33	36	27	6.73	7.27	0.80
	14.53	11.85	36	27	3.98	4.19	0.82
	16.06	15.56	36	27	4.04	4.40	0.87
	22.17	19.43	24	21	5.27	7.57	0.84
	15.96	14.33	24	21	6.11	7.59	0.80
	14.67	11.48	24	21	3.43	4.18	0.82
	16.08	14.24	24	21	4.10	4.65	0.87
	24.00	20.83	12	6	7.24	6.85	0.84
	17.42	14.33	12	6	8.04	6.65	0.80
	14.25	13.17	12	6	5.07	4.31	0.82
	16.00	15.66	12	6	4.09	3.50	0.87
Larson	1.292	2.000	24	24	0.908	1.180	
	2.790	2.630	24	24	0.588	0.875	
	0.833	0.708	24	24	0.381	0.464	
	2.917	2.833	24	24	0.282	0.637	
	2.875	2.917	24	24	0.448	0.282	
Lawler	87.75	73.85	40	41	11.49	11.21	0.85
Lee	27.09	28.76	23	21	8.02	5.34	
Liu	71.13	52.19	35	17	12.70	20.15	
	74.14	68.17	14	10	10.26	12.78	
Lorber	31.05	23.5	20	20	6.4	5.84	0.72

Table 9--Continued

Author	$\overline{Y}_E$	$\overline{Y}_C$	N_E	N_C	S_E	S_C	r_{xx}
Lozano	92.10	88.30	69	64	12.30	17.40	
	30.40	26.70	69	64	4.80	6.40	
	9.97	9.88	69	64	2.32	2.62	
	10.30	10.25	69	64	2.03	2.20	
	41.40	41.40	69	64	7.61	11.00	
	54.30	55.80	69	64	16.80	15.10	
McKay	71.70	60.30	16	22	11.40	11.00	0.85
Meyer & Beaton	83.2	82.3	25	23	7.8	10.7	
	67.7	65.5	25	23	13.9	12.5	
	77.7	78.0	25	23	13.7	13.7	
Mitzell	83.28	95.71	7	7	16.69	9.18	0.93
	49.50	47.18	7	7	8.09	4.58	0.94
Murphy	23.69	23.12	13	12	5.35	5.27	0.90
	28.13	29.00	15	12	5.22	4.45	0.93
	33.35	35.93	88	60	8.70	9.58	0.90
Oates	83.0	85.0	61	38	6.1	7.5	0.94
	62.0	52.0	11	20	11.0	14.0	0.94
Paden	17.10	16.60	79	99	3.25	3.48	
	23.70	22.50	79	99	3.34	3.30	
	32.00	30.20	79	99	4.91	5.89	
	23.00	21.70	79	99	3.72	4.54	
	46.52	45.43	318	64	7.87	9.79	
	223.09	221.47	33	30	31.80	30.00	
Proctor	81.8	78.3	10	10	10.26	10.48	0.93
	92.5	85.9	10	10	9.49	14.13	0.93
	94.6	85.9	10	10	12.48	14.13	0.93
	88.9	78.3	10	10	12.52	10.48	0.93
Romaniuk	73.06	71.78	18	18	18.51	23.11	
Rota	18.8	20.0	28	19	3.9	2.5	0.65
	20.0	19.6	28	19	4.0	2.5	0.65
Saul	17.428	16.574	86	250	5.24	5.84	
	15.157	16.574	260	250	3.97	5.84	
	13.107	12.489	86	250	4.10	4.43	
	11.340	12.489	260	250	3.38	4.43	

Table 9--Continued

Author	$\bar{Y}_E$	$\bar{Y}_C$	N_E	N_C	S_E	S_C	r_{xx}
Skavaril	60.00	57.00	70	50	16.45	19.06	
Smith	70.94	72.14	87	56	16.04	14.16	
	67.63	64.44	97	61	12.76	14.91	
	68.69	64.44	100	61	13.12	14.91	
	72.81	72.14	95	56	15.32	14.16	
Swigger	21.00	16.77	14	13	1.00	2.26	
Thompson	60.6	56.7	33	16	16.1	20.0	0.76
	62.7	56.7	35	16	17.3	20.0	0.76
	67.5	68.0	46	21	14.5	15.0	
	70.8	68.0	40	21	13.3	15.0	
	66.3	58.8	39	18	15.3	17.5	
	67.3	58.8	38	18	15.0	17.5	
	70.3	66.1	34	17	15.8	17.7	
	71.3	66.1	35	17	13.8	17.7	
	73.8	63.5	33	16	14.3	19.0	
	74.5	63.5	35	16	15.0	19.0	
Tira	17.74	15.06	27	27	2.07	2.09	0.72
Tollefson	13.65	12.81	29	51	1.29	1.26	
	12.37	10.69	29	51	2.34	4.43	
Torop	30.50	16.03	30	30	6.03	5.34	
Tsai	85.53	86.93	15	15	11.20	4.04	
	85.46	83.33	15	15	8.19	6.79	
	90.27	86.97	15	15	9.57	11.36	
	83.17	80.10	15	15	7.09	5.80	
Vaughn	44.60	30.90	20	20	8.29	7.12	0.88
	46.20	28.90	20	20	5.97	6.30	0.88
	18.80	12.75	20	20	3.90	4.01	0.84
	19.20	12.30	20	20	2.55	2.79	0.84
	9.70	6.80	20	20	3.06	2.84	0.72
	11.30	6.45	20	20	3.51	2.54	0.72
	16.10	11.35	20	20	3.29	3.51	0.67
	15.70	10.15	20	20	2.49	3.35	0.67
Wolcott	17.00	23.00	22	22	7.165	7.74	0.91

Table 9--Continued

Author	$\bar{Y}_E$	$\bar{Y}_C$	N_E	N_C	S_E	S_C	r_{xx}
Wood	19.94	19.63	17	24	4.23	4.33	
	88.25	90.49	20	30	7.95	6.78	
	84.75	93.15	20	30	5.80	5.76	
	84.75	90.82	20	30	6.00	6.36	
	89.50	88.46	20	30	8.00	7.31	
	85.75	90.25	20	30	7.65	4.78	

APPENDIX C

COMPUTED EFFECT SIZES FOR STUDIES USED IN META-ANALYSIS

Table 10

Computed Effect Sizes for Studies Used in Meta-Analysis

Author	d	Δ	g'	d^r	d^w	g^w
Alderman (English study)	0.368	0.344	0.365	0.416	0.383	0.337
	0.128	0.131	0.126	0.142	0.095	0.106
	0.442	0.418	0.441	0.498	1.207	1.328
	-0.030	-0.033	-0.030	-0.032	-0.036	-0.038
	0.283	0.275	0.280	0.310	0.299	0.258
	0.329	0.329	0.324	0.360	0.238	0.270
	0.173	0.183	0.173	0.190	0.536	0.518
	-0.503	-0.475	-0.499	-0.552	-0.510	-0.603
	0.512	0.492	0.508	0.558	0.590	0.664
	-1.061	-1.184	-1.058	-1.106	-3.670	-3.042
	-0.011	-0.012	-0.011	-0.012	-0.031	-0.032
	0.813	0.813	0.806	0.891	0.975	1.000
	-0.725	-0.725	-0.723	-0.795	-2.247	-2.197
	0.188	0.188	0.187	0.206	0.459	0.518
Alderman (Math study)	0.571	0.546	0.568	0.595	1.242	0.746
	0.242	0.234	0.241	0.261	0.530	0.395
	0.901	0.800	0.894	0.977	1.010	1.078
	0.375	0.362	0.372	0.404	0.489	0.308
	0.442	0.433	0.436	0.472	0.292	0.204
	-0.046	-0.046	-0.045	-0.050	-0.031	-0.025
	0.019	0.018	0.019	0.020	0.016	0.012
	1.200	1.210	1.182	1.265	0.771	0.463
	0.474	0.559	0.466	0.496	0.321	0.301
	-0.050	-0.056	-0.049	-0.052	-0.015	-0.016
	0.071	0.072	0.067	0.076	0.013	0.007
	1.208	1.330	1.177	1.302	0.532	0.302
Andrews	-0.049	-0.046	-0.048	-0.050	-0.022	-0.026
	-0.056	-0.049	-0.055	-0.058	-0.023	-0.030
	0.025	0.024	0.025	0.026	0.012	0.013
Axeen	-0.172	-0.167	-0.170	-0.185	-0.138	-0.162
Boen	1.101	0.922	1.073	1.207	0.369	0.437
Boyson	0.632	0.622	0.618	0.694	0.280	0.309

Table 10--Continued

Author	d	Δ	g'	d^r	d^w	g^w
Brum	0.622	0.692	0.615	0.682	0.606	0.594
Cartwright	2.274	2.175	2.259	2.495	3.098	1.855
Caruso	0.059	0.057	0.058	0.065	0.062	0.072
	0.083	0.106	0.083	0.092	0.115	0.102
Castleberry	1.965	2.133	1.958	2.225	5.277	3.816
Conklin	1.495	1.952	1.446	2.016	0.610	0.417
Crawford	0.120	0.102	0.119	0.131	0.490	0.370
	0.435	0.383	0.434	0.477	1.834	1.330
	-0.409	-0.347	-0.408	-0.449	-1.659	-1.254
Daughdrill	0.069	0.060	0.068	0.076	0.050	0.066
Diem	-0.678	-0.829	-0.656	-0.692	-0.259	-0.223
	-0.658	-0.699	-0.639	-0.722	-0.245	-0.248
	-0.058	-0.081	-0.056	-0.064	-0.028	-0.023
DuBoulay & Howe	0.155	0.181	0.143	0.170	0.027	0.025
	0.423	0.422	0.391	0.464	0.063	0.067
	-0.440	-0.470	-0.406	-0.483	-0.070	-0.070
Durgin	0.107	0.107	0.106	0.142	0.105	0.121
	0.154	0.148	0.152	0.204	0.141	0.168
	0.014	0.013	0.014	0.018	0.013	0.016
	0.168	0.174	0.166	0.216	0.166	0.183
	-0.024	-0.028	-0.024	-0.029	-0.027	-0.028
	0.188	0.200	0.186	0.226	0.190	0.205
	-0.027	-0.025	-0.027	-0.034	-0.024	-0.031
	0.287	0.298	0.284	0.356	0.283	0.311
Fiedler	-0.109	-0.102	-0.107	-0.119	-0.061	-0.075
	-0.183	-0.182	-0.179	-0.200	-0.087	-0.099
	0.460	0.491	0.451	0.505	0.246	0.256
	0.383	0.355	0.374	0.420	0.160	0.193
Friesen	-0.094	-0.096	-0.093	-0.102	-0.165	-0.184

Table 10--Continued

Author	d	Δ	g'	d^r	d^w	g^w
Grandey	0.731	0.867	0.708	0.802	0.282	0.252
	1.040	0.970	1.007	1.141	0.315	0.338
	0.646	0.503	0.625	0.708	0.163	0.226
	1.161	1.568	1.124	1.273	0.509	0.368
Gray	0.238	0.235	0.236	0.261	0.203	0.217
Green	0.562	0.478	0.538	0.617	0.119	0.151
	-0.287	-0.418	-0.267	-0.315	-0.068	-0.050
	-0.897	-0.871	-0.867	-0.984	-0.272	-0.278
	0.098	0.109	0.093	0.107	0.024	0.024
Hartig	0.890	0.881	0.886	0.977	1.783	0.972
Henry & Ramsette	0.179	0.179	0.179	0.205	0.939	0.842
Hofstetter	0.821	0.764	0.801	0.900	0.315	0.356
	0.543	0.497	0.530	0.596	0.205	0.246
Holoien	-0.293	-0.333	-0.285	-0.309	-0.121	-0.119
	0.230	0.286	0.224	0.245	0.104	0.094
	0.697	0.600	0.678	0.742	0.225	0.279
	-0.395	-0.455	-0.384	-0.419	-0.165	-0.159
	0.769	0.692	0.748	0.829	0.260	0.304
	-0.541	-0.600	-0.526	-0.558	-0.217	-0.215
	0.487	0.500	0.474	0.500	0.187	0.201
	0.301	0.316	0.293	0.309	0.118	0.126
Homeyer	1.388	1.318	1.254	1.523	0.165	0.153
	-0.386	-0.519	-0.349	-0.423	-0.065	-0.050
	0.035	0.035	0.031	0.038	0.004	0.005
	-0.250	-0.941	-0.226	-0.275	-0.118	-0.033
Hong	1.048	0.948	1.001	1.150	0.225	0.241
	0.728	0.808	0.703	0.798	0.242	0.225
	0.396	0.323	0.387	0.435	0.141	0.192
	0.116	0.115	0.112	0.127	0.036	0.037
Hu & Saunders	0.744	0.732	0.738	0.816	0.879	0.966
	1.139	1.141	1.130	1.249	1.368	1.362
Huckabay	0.814	0.669	0.793	0.873	0.259	0.329
	0.777	0.670	0.757	0.815	0.259	0.316

Table 10--Continued

Author	d	Δ	g'	d^r	d^w	g^w
Hughes	0.191	0.173	0.190	0.197	0.197	0.250
Jim	0.084	0.077	0.083	0.092	0.063	0.079
Johnson	0.549	0.500	0.542	0.602	0.406	0.492
	0.506	0.450	0.501	0.556	0.377	0.468
	-0.208	-0.175	-0.205	-0.228	-0.127	-0.172
	0.693	1.037	0.685	0.760	0.842	0.609
	1.415	1.444	1.399	1.552	1.209	1.087
	0.073	0.074	0.072	0.080	0.054	0.061
Lang	0.046	0.485	0.458	0.504	0.510	0.535
	0.244	0.240	0.242	0.273	0.252	0.288
	0.433	0.420	0.429	0.478	0.440	0.502
	0.037	0.039	0.037	0.040	0.041	0.044
	0.463	0.416	0.457	0.505	0.327	0.400
	0.303	0.290	0.299	0.339	0.228	0.266
	0.658	0.640	0.650	0.727	0.504	0.556
	0.119	0.114	0.118	0.128	0.089	0.106
	0.425	0.362	0.418	0.464	0.204	0.267
	0.238	0.215	0.234	0.266	0.121	0.152
	0.840	0.763	0.825	0.928	0.429	0.496
	0.422	0.396	0.414	0.452	0.223	0.265
	0.445	0.463	0.424	0.486	0.104	0.097
	0.405	0.465	0.386	0.453	0.105	0.088
	0.223	0.251	0.212	0.246	0.056	0.049
	0.087	0.097	0.083	0.093	0.022	0.019
Larson	-0.672	-0.600	-0.661	-0.738	-0.360	-0.438
	0.215	0.183	0.211	0.235	0.110	0.147
	0.294	0.269	0.290	0.323	0.162	0.200
	0.171	0.132	0.168	0.187	0.079	0.117
	-0.112	-0.149	-0.110	-0.123	-0.089	-0.077
Lawler	1.225	1.240	1.213	1.328	1.255	1.208
Lee	-0.243	-0.313	-0.239	-0.266	-0.172	-0.151
Liu	1.224	0.940	1.205	1.342	0.611	0.692
	0.525	0.467	0.507	0.577	0.140	0.167
Lorber	1.232	1.293	1.208	1.452	0.646	0.595

Table 10--<u>Continued</u>

Author	d	Δ	g'	d^r	d^w	g^w
Lozano	0.084	0.077	0.083	0.092	0.063	0.079
	0.658	0.578	0.654	0.721	0.961	1.200
	0.036	0.034	0.036	0.040	0.057	0.070
	0.024	0.023	0.024	0.026	0.038	0.045
	0.000	0.000	0.000	0.000	0.000	0.000
	-0.094	-0.099	-0.093	-0.103	-0.165	-0.180
McKay	1.021	1.036	0.999	1.107	0.492	0.481
Meyer & Beaton	0.097	0.084	0.095	0.106	0.050	0.066
	0.166	0.176	0.163	0.182	0.106	0.114
	-0.022	-0.022	-0.022	-0.024	-0.013	-0.015
Mitzell	-0.923	-1.354	-0.864	-0.957	-0.237	-0.161
	0.353	0.507	0.330	0.364	0.089	0.066
Murphy	0.107	0.108	0.104	0.113	0.034	0.038
	-0.178	-0.196	-0.172	-0.184	-0.066	-0.067
	-0.285	-0.269	-0.283	-0.300	-0.498	-0.583
Oates	-0.300	-0.267	-0.298	-0.309	-0.330	-0.402
	0.767	0.714	0.747	0.791	0.277	0.290
Paden	0.148	0.144	0.147	0.162	0.320	0.376
	0.362	0.364	0.360	0.397	0.809	0.907
	0.329	0.306	0.327	0.361	0.680	0.826
	0.310	0.286	0.308	0.340	0.637	0.780
	0.133	0.111	0.132	0.145	0.531	0.410
	0.052	0.054	0.052	0.057	0.043	0.047
Proctor	0.337	0.334	0.323	0.350	0.083	0.093
	0.548	0.467	0.525	0.569	0.117	0.148
	0.653	0.616	0.625	0.677	0.154	0.174
	0.918	1.011	0.879	0.952	0.253	0.234
Romaniuk	0.061	0.055	0.060	0.067	0.025	0.031
Rota	-0.352	-0.480	-0.346	-0.437	-0.282	-0.225
	0.115	0.160	0.113	0.143	0.094	0.074
Saul	0.150	0.146	0.150	0.165	0.614	0.557
	-0.285	-0.243	-0.284	-0.312	-1.546	-2.090
	0.142	0.140	0.142	0.156	0.586	0.527
	-0.292	-0.259	-0.292	-0.321	-1.653	-2.144

Table 10--Continued

Author	d	Δ	g'	d^r	d^w	g^w
Skavaril	0.171	0.157	0.170	0.187	0.236	0.287
Smith	-0.078	-0.085	-0.078	-0.086	-0.151	-0.154
	0.234	0.214	0.233	0.257	0.422	0.505
	0.307	0.285	0.306	0.337	0.573	0.668
	0.045	0.047	0.045	0.049	0.089	0.092
Swigger	2.454	1.872	2.379	2.692	0.631	0.547
Thompson	0.224	0.195	0.220	0.257	0.119	0.137
	0.330	0.300	0.325	0.379	0.191	0.206
	-0.034	-0.033	-0.034	-0.037	-0.028	-0.028
	0.201	0.187	0.199	0.221	0.142	0.159
	0.468	0.429	0.462	0.514	0.305	0.324
	0.537	0.486	0.529	0.589	0.340	0.365
	0.255	0.237	0.251	0.280	0.151	0.165
	0.343	0.294	0.338	0.376	0.191	0.222
	0.646	0.542	0.635	0.708	0.332	0.382
	0.674	0.579	0.663	0.739	0.369	0.405
Tira	1.288	1.282	1.270	1.518	0.865	0.831
Tollefson	0.661	0.667	0.655	0.725	0.666	0.672
	0.440	0.379	0.436	0.483	0.379	0.460
Torop	2.541	2.710	2.508	2.787	2.032	1.227
Tsai	-0.166	-0.347	-0.162	-0.182	-0.130	-0.070
	0.283	0.314	0.275	0.311	0.118	0.119
	0.314	0.290	0.306	0.345	0.109	0.132
	0.474	0.529	0.461	0.520	0.198	0.196
Vaughn	1.773	1.924	1.738	1.890	0.962	0.735
	2.819	2.746	2.763	3.005	1.373	0.823
	1.530	1.509	1.499	1.669	0.754	0.682
	2.582	2.473	2.530	2.817	1.236	0.819
	0.982	1.021	0.963	1.158	0.510	0.503
	1.583	1.909	1.552	1.866	0.954	0.695
	1.396	1.353	1.369	1.706	0.676	0.646
	1.880	1.657	1.843	2.297	0.828	0.754
Wolcott	-0.805	-0.775	-0.790	-0.843	-0.426	-0.470

Table 10--Continued

Author	d	Δ	g'	d^r	d^w	g^w
Wood	0.072	0.072	0.071	0.079	0.037	0.041
	-0.308	-0.330	-0.303	-0.338	-0.206	-0.210
	-1.454	-1.458	-1.431	-1.595	-0.911	-0.803
	-0.976	-0.954	-0.961	-1.071	-0.596	-0.604
	0.137	0.142	0.135	0.150	0.089	0.094
	-0.740	-0.941	-0.728	-0.812	-0.588	-0.479

BIBLIOGRAPHY

Books

Blalock, H. M. (1979). Social statistics. New York: McGraw-Hill.

Bryant, F. B., & Wortman, P. M. (1984). Methodological issues in the meta-analysis of quasi-experiments. In W. H. Yeaton & P. M. Wortman (Eds.), Issues in Data Synthesis (pp. 5-24). San Francisco: Jossey-Bass.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences (2nd ed). New York: Academic Press.

Cook, T. D. (1974). The potential and limitations of secondary evaluations. In M. W. Apple, M. J. Subkoviak & H. S. Lufler, Jr. (Eds.) Educational evaluation: Analysis and responsibility (pp. 155-222). Berkeley, CA: McCutchan Publishing Co.

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

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

Hedges, L. V. (1984). Advances in statistical method for meta-analysis. In W. H. Yeaton & P. M. Wortman (Eds.), Issues in Data Synthesis (pp. 25-42). San Francisco: Jossey-Bass.

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

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

Kirk, R. E. (1982). Experimental design: Procedures for the behavioral sciences. Monterey: Brooks/Cole Publishing Co.

Light, R. J. (1983). Evaluation studies review annual. Vol. 8. Beverly Hills: Sage Publications.

Light, R. J. (1984). Six evaluation issues that synthesis can resolve better than single studies. In W. H. Yeaton & P. M. Wortman (Eds.), Issues in Data Synthesis (pp. 57-73). San Francisco: Jossey-Bass.

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

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

Wolf, F. M. (1984). Meta-analysis: Quantitative methods for research synthesis. Beverly Hills: Sage University Paper.

Articles

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

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

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C.-L. C. (1985). Effectiveness of computer-based instruction in secondary schools. Journal of Computer-Based Instruction, 12(3), 59-68.

Becker, B. J., & Hedges, L. V. (1984). Meta-analysis of cognitive gender differences: A comment on the analysis by Rosenthal and Rubin. Journal of Educational Psychology, 76, 583-587.

Carlberg, C. G. (1984). Meta-analysis in education: A reply to Slavin. Educational Researcher, 13(8), 16-23.

Cook, T. D., & Leviton, L. C. (1980). Reviewing the literature: A comparison of traditional methods with meta-analysis. Journal of Personality, 48, 449-469.

Cooper, H. M. (1979). Statistically combining independent studies: A meta-analysis of sex differences in conformity research. Journal of Personality and Social Psychology, 37, 131-146.

Cooper, H. M. (1981). On the significance of effects and the effects of significance. Journal of Personality and Social Psychology, 41, 1013-1018.

Cooper, H. M. (1982). Scientific guidelines for conducting integrative research reviews. Review of Educational Research, 52, 291-302.

Cooper, H. M., & Arkin, R. M. (1981). Quantitative reviewing. Journal of Personality, 49, 225-229.

Cooper, H. M., & Rosenthal, R. (1980). Statistical vs. traditional procedures for summarizing research findings. Psychological Bulletin, 87, 442-449.

Cotton, J. L., & Cook, M. S. (1982). Meta-analysis and the effects of various reward systems. Psychological Bulletin, 92, 176-183.

Crain, R. L., & Mahard, R. E. (1983). The effects of research methodology on desegregation achievement studies: A meta-analysis. American Journal of Sociology, 88, 839-855.

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

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

Fiske, D. W. (1983). The meta-analytic revolution in outcome research. Journal of Consulting and Clinical Psychology, 51, 65-70.

Gage, N.L. (1982). Future of educational research. Educational Researcher, 11(8), 11-19.

Gallo, P. S. (1978). Meta-analysis: A mixed meta-phor? American Psychologist, 33, 515.

Glass, G. V. (1976). Primary, secondary and meta-analysis of research. Educational Researcher, 5(9), 3-8.

Glass, G. V. (1978). Reply to Eysenck. American Psychologist, 33, 517-518.

Glass, G. V. (1978). Reply to Mansfield and Busse. Educational Researcher, 7, 3.

Glass, G. V. (1982). Meta-analysis: An approach to the synthesis of research results. Journal of Research in Science Teaching, 19, 93-112.

Glass, G. V., & Kliegl, J. M. (1983). An apology for research integration in the study of psychotherapy. Journal of Consulting and Clinical Psychology, 51, 28-41.

Hedges, L. V. (1981). Distribution theory for Glass' estimator of effect size and related estimators. Journal of Educational Statistics, 6(2), 107-128.

Hedges, L. V. (1982). Fitting categorical models to effect sizes from a series of experiments. Journal of Educational Statistics, 7(2), 119-137.

Hedges, L. V., & Olkin, I. (1986). Meta-analysis: A review and a new view. Educational Researcher, 15(8), 14-21.

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

Hsu, L. M. (1980). Tests of differences in p levels as tests of differences in effect sizes. Psychological Bulletin, 88, 705-708.

Hyde, J. S. (1981). How large are cognitive gender differences? A meta-analysis using w2 and d. American Psychologist, 36, 892-901.

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

Joyce, B. (1987). A rigorous yet delicate touch: A response to Slavin's proposal for 'Best-Evidence' reviews. Educational Researcher, 16(4), 12-14.

Kraemer, H. C. (1983). Theory of estimation and testing of effect sizes: Use in meta-analysis. Journal of Educational Statistics, 8, 93-101.

Kraemer, H. C., & Andrews, G. (1982). A nonparametric technique for meta-analysis effect size calculation. Psychological Bulletin, 91, 404-412.

Krauth, J. (1983). Nonparametric effect size calculation: A comment on Kraemer and Andrews. Psychological Bulletin, 94, 190-192.

Kulik, C.-L. C., & Kulik, J. A. (in press). Effectiveness of computer-based education in colleges. AEDS Journal.

Landman, J., & Dawes, J. M. (1982). Psychotherapy outcomes: Smith and Glass conclusions stand up to scrutiny. American Psychologist, 37, 504-516.

Light, R. J., & Pillemer, D. (1982). Numbers and narrative: Combining their strength in research reviews. Harvard Educational Review, 52(2), 1-26.

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

Mansfield, R. S., & Busse, T. V. (1977). Meta-analysis of research: A rejoinder to Glass. Educational Researcher, 10(8), 3.

McGaw, B., & Glass, G. V. (1980). Choice of the metric for effect size in meta-analysis. American Educational Research Journal, 17, 325-337.

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

Oliver, L. W., & Spokane, A. R. (1983). Research integration: Approaches, problems and recommendations for research reporting. Journal of Counseling Psychology, 30, 252-257.

Pillemer, D. B., & Light, R. J. (1980). Synthesizing outcomes: How to use research evidence from many studies. Harvard Educational Review, 50, 176-195.

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical bayes meta-analysis. Journal of Educational Statistics, 10, 75-97.

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

Rosenthal, R., & Rubin, D. (1979). A note on percent variance explained as a measure of the importance of effects. Journal of Applied Social Psychology, 9, 395-396.

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

Rosenthal, R., & Rubin, D. (1982). Further meta-analytic procedures for assessing cognitive gender differences. Journal of Educational Psychology, 74, 708-712.

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

Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. Educational Researcher, 15(9), 5-11.

Slavin, R. E. (1987). Best-evidence synthesis: Why less is more. Educational Researcher, 16(4), 15-16.

Steinkamp, M., & Maehr, M. L. (1984). Gender differences in motivational orientations toward achievement in school science: A quantitative synthesis. American Educational Research Journal, 21(1), 39-59.

Stock, W. A. et al. (1982). Rigor in data synthesis: A case study of reliability in meta-analysis. Educational Researcher, 11(5), 10-15.

Strube, M. J. (1985). Combining and comparing significance levels from nonindependent hypothesis tests. Psychological Bulletin, 97, 334-341.

Strube, M. J., & Hartman, D. P. (1983). Meta-analysis: techniques, applications and functions. Journal of Consulting and Clinical Psychology, 51, 14-27.

Viana, M. A. (1980). Statistical methods for summarizing independent correlational results. Journal of Educational Statistics, 5, 83-104.

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

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

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

ERIC Reports

Bangert-Drowns, R. L. (1984). Developments in meta-analysis: A review of five methods. Ann Arbor, MI: Michigan University, Center for Research on Learning and Teaching. (ERIC Document Reproduction Service No. ED 248 262)

Bangert-Drowns, R. (1985, April). The meta-analytic debate. Paper presented at the 69th Annual Meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 262 095)

Becker, B. J. (1984, April). Power differences among tests of combined significance. Paper presented at the 68th Annual Meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 249 266)

Kulik, J. A. (1984, April). Uses and misuses of meta-analysis. Paper presented at the 68th Annual Meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 247 270)

Reynolds, S. & Day, J. (1984, August). Monte Carlo studies of effect size estimates and their approximations in meta-analysis. Paper presented at the 92nd Annual Convention of the American Psychological Association, Toronto, Ontario. (ERIC Document Reproduction Service No. ED 253 567)

Schmidt, F. L. (1984, August). Meta-analysis: Implications for cumulative knowledge in the behavioral and social sciences. Invited address at the 92nd Annual Convention of the American Psychological Association, Toronto, Ontario. (ERIC Document Reproduction Service No. ED 251 722.

Unpublished Materials

Jackson, S.E. (1984, August). Can meta-analysis be used for theory development in organizational psychology? Paper presented at the 92nd Annual Meeting of the American Psychological Association, Toronto, Ontario.

Orwin, R.G., and Cordray, D. S. (1983). The effects of deficit reporting on meta-analysis: A conceptual framework and reanalysis. Unpublished manuscript, Northwestern University.

Tracz, S. M. (1984). The effect of the violation of the assumption of independence when combining correlation coefficients in a meta-analysis. Unpublished doctoral dissertation, Southern Illinois University at Carbondale.

Studies Used in Meta-Analysis

Alderman, D. R. (1978). Evaluation of the TICCIT computer-assisted instructional system in the community college. Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction Service No. ED 167 606)

Andrews, C. S. (1974). An investigation of the use of computer-assisted instruction in French as an adjunct to classroom instruction. Dissertation Abstracts International, 34, 5900A. (University Microfilms No. 74-6710)

Axeen, M. E. (1967). Teaching library use to undergraduates: Comparison of computer-based instruction and the conventional lecture. Urbana, IL: University of Illinois. (ERIC Document Reproduction Service No. ED 014 316)

Boen, L. L. (1983). Traditional and computer-directed instruction (CDI) of study skills in college students. Dissertation Abstracts International, 43, 3213A.

Boysen, J. P., & Francis, P. R. (1982). An Evaluation of the instructional effectiveness of a computer lesson in biomechanics. Research Quarterly for Exercise and Sport, 53, 232-235.

Brum, Joseph Jr. (1983). Effects of computer-assisted instruction on students' final grades. Nova University: Doctoral Practicum Paper. (ERIC Document Reproduction Service No. ED 263 832)

Cartwright, C. A., Cartwright, G. P., & Robine, G. C. (1972). CAI course in the early identification of handicapped children. Exceptional Children, 38, 453-459.

Caruso, D. E. F. (1970). An experiment to determine the effectiveness of an interactive tutorial program, implemented on the time-sharing IBM system 360, model 50, in teaching a subject-oriented user to formulate inquiry statements to a computerized on-line information retrieval system. Dissertation Abstracts International, 30, 3484A. (University Microfilms No. 70-2051)

Castleberry, S. J., Montague, E. J., & Lagowski, J. J. (1970). Computer-based teaching techniques in general chemistry, Journal of Research in Science Teaching, 7, 197-208.

Conklin, D. N. (1983). A study of computer-assisted instruction in nursing education. Journal of Computer-Based Instruction, 9, 98-107.

Crawford, A. M. et al. (1978). Evaluation of a computer-based course management system. Final report. Urbana, IL: University of Illinois. (ERIC Document Reproduction Service No. ED 165 790)

Daughdrill, R. W. (1978). A comparative study of the effectiveness of computer-assisted instruction in college algebra. Dissertation Abstracts International, 39, 3431A. (University Microfilms No. 7824040)

Diem, D. C. (1982). The effectiveness of computer-assisted instruction in college algebra. Dissertation Abstracts International, 43, 1456A.

DuBoulay, J. B., & Howe, J. A. (1982). Logo building blocks: Student teachers using computer-based mathematics apparatus. Computers and Education, 6, 93-98.

Durgin, M. W. (1979). The effects of teaching beginning college mathematics with a business emphasis by computer aided instruction. Dissertation Abstracts International, 39, 5380A. (University Microfilms No. 7904951)

Fiedler, L. A. (1969). A comparison of achievement resulting from learning mathematical concepts by computer programming versus class assignment approach. Dissertation Abstracts International, 29, 3910A. (University Microfilm No. 69-8595)

Friesen, V. E. (1977). The relationship of affective and cognitive variables to achievement and attitude under lecture-discussion and computer-assisted instruction. Dissertations Abstracts International, 37, 4095A. (University Microfilm No. 76-29,997)

Grandey, R. C. (1971). The use of computers to aid instruction in beginning chemistry. Journal of Chemical Education, 48, 791-794.

Gray, C. F. (1973). Expressed student attitude toward conventional versus computer supplemented instruction. Decision Sciences, 4, 141-148.

Green, C., & Mink, W. (1973). Evaluation of computer simulation of experiments in teaching scientific methodology. St. Paul, MN: Macalester College. (ERIC Document Reproduction Service No. ED 082 475)

Hartig, G. (1984). Implementing CAI in a university learning center. Journal of Computer-Based Instruction, 11, 113-116.

Henry, M., & Ramsett, D. (1978). The effects of computer-aided instruction on learning and attitudes in economic principles courses. The Journal of Economic Education, 10, 26-34.

Hofstetter, F. T. (1975). GUIDO: An interactive computer-based system for improvement of instruction and research in ear-training. Journal of Computer-Based Instruction, 1, 100-106.

Holoien, M. O. (1971). Calculus and computing: A comparative study of the effectiveness of computer programming as an aid in learning selected concepts in first-year calculus. Dissertation Abstracts International, 31, 4490A.

Homeyer, F. C. (1970). Development and evaluation of an automated assembly language teacher. Austin, TX: University of Texas. (ERIC Document Reproduction Service No. ED 053 531)

Hong, S. T. (1973). An empirical study of the effectiveness of programmed instruction and computer-assisted instruction in elementary accounting. Dissertation Abstracts International, 33, 4589A. (University Microfilm No. 73-5299)

Hu, M. Y., & Saunders, G. (1986). Integrating the microcomputer into managerial accounting classes: An experimental study. The Woman CPA, 48(1), 29-31.

Huckabay, L. et al. (1979). Cognitive, affective, and transfer of learning consequences of computer-assisted instruction. Nursing Research, 28, 228-233.

Hughes, R. J. (1977). An experimental study in teaching mathematical concepts utilizing computer-assisted instruction in business machines. Dissertation Abstracts International, 37, 6911A.

Jim, L. K. et al. (1984). A computer-assisted instructional approach to teaching applied therapeutics. American Journal of Pharmaceutical Education, 48, 21-25.

Johnson, C. W., & Plake, B. S. (1981, April). Interactive study lessons to complement ANOVA. Paper presented at the 65th Annual Meeting of the American Educational Research Association, Los Angeles, CA. (ERIC Document Reproduction Service No. ED 204 361)

Lang, M. T. (1974). Computer extended instruction in introductory calculus. Dissertation Abstracts International, 34, 5662A. (University Microfilm No. 74-5271)

Larson, D. E. (1982). The use of computer-assisted instruction to teach calculation and regulation of intravenous flow rates to baccalaureate nursing students. Dissertation Abstracts International, 42, 3459A. (University Microfilm No. 8202467)

Lawler, R. M. (1971). An investigation of selected instructional strategies in an undergraduate computer-managed instruction course. Tallahassee, FL: Florida State University, Computer-Assisted Instruction Center. (ERIC Document Reproduction Service No. ED 054 652)

Lee, A. L. (1973). A comparison of computer-assisted instruction and traditional laboratory instruction in an undergraduate geology course. Dissertation Abstracts International, 34, 2273A. (University Microfilm No. 73-26,036)

Liu, H. C. (1975). Computer-assisted instruction in teaching college physics. Dissertation Abstracts International, 36, 1411A. (University Microfilm No. 75-18,862)

Lorber, M. A. (1970). The effectiveness of computer-assisted instruction in the teaching of tests and measurements to prospective teachers. Dissertation Abstracts International, 31, 2775A. (University Microfilm No. 70-24,434)

Lozano, A. G. (1985). Educational technology and language training. International research and studies program. Boulder, CO: University of Colorado. (ERIC Document Reproduction Service No. ED 263 786)

McKay, A. B., & Jackson, R. A. (1984). The provision of management CAI through commercially available pharmacy computer systems. American Journal of Pharmaceutical Education, 48, 11-19.

Meyer, J. H., & Beaton, G. R. (1974). An evaluation of computer-assisted teaching in physiology. Journal of Medical Education, 49, 295-297.

Mitzel, H. E. (1967). The development and presentation of four college courses by computer teleprocessing. University Park, PA: Pennsylvania State University. (ERIC Document Reproduction Service No. ED 016 377)

Murphy, R. T., & Appel, L. R. (1977). Evaluation of the PLATO IV computer-based education system in the community college. Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction Service No. ED 146 235)

Oates, W. R. (1983). Effects of computer-assisted instruction in writing skills on journalism students in beginning newswriting classes. Dissertation Abstracts International, 43, 2822A.

Paden, D. W., Dalgaard, B. R., & Barr, M. D. (1977). A decade of computer-assisted instruction. The Journal of Economic Education, 9, 14-20.

Proctor, W. L. (1969). A comparison of two instructional strategies based on computer-assisted instruction with a lecture-discussion strategy for presentation of general curriculum concepts. Dissertation Abstracts International, 29, 2075A. (University Microfilm No. 69-591)

Romaniuk, E. W. (1978). A summative evaluation of the CAI course 'COMPS'. Edmonton, Canada: Alberta University, Division of Educational Research Services. (ERIC Document Reproduction Service No. ED 153 604.

Rota, D. R. (1982). Computer-assisted instruction, lecture instruction, and combined computer-assisted/lecture instruction: A comparative experiment. Dissertation Abstracts International, 42, 4809A. (University Microfilm No. DA8208685)

Saul, W. E. (1975). An experimental study of the effects of computer-augmented instruction on achievement and attrition in beginning accounting at Miami-Dade Community College, North Campus. Dissertation Abstracts International, 35, 4757A. (University Microfilm No. 75-363A)

Skavaril, R. V. (1974). Computer-based instruction of introductory statistics. Journal of Computer-Based Instruction, 1(1), 32-40.

Smith, R. B. (1976). The effects of computer-assisted feedback on students' performance in a televised college course. Dissertation Abstracts International, 36, 5163A. (University Microfilm No. 76-2516)

Swigger, K. M. (1976). Automated Flanders interaction analysis. Journal of Computer-Based Instruction, 2(2), 63-66.

Thompson, F. A. (1977). TIPS teaching information processing systems implementation at Riverside City College, 1976-77: An experiment in educational innovation. Final report. Riverside City College, CA. (ERIC Document Reproduction Service No. ED 142 248)

Tira, D. E. (1977). Rationale for and evaluation of a CAI tutorial in a removable partial prosthodontics classification system. Journal of Computer-Based Instruction, 4(11), 34-42.

Tollefson, N. (1978). A comparison of computerized and paper-pencil formative evaluation. College Student Journal, 12, 103-106.

Torop, W. (1975, March). An analysis of individualized learning system chemistry. Paper presented at the 48th Annual Meeting of the National Association for Research in Science Teaching, Los Angeles, CA. (ERIC Document Reproduction Service No. ED 110 321)

Tsai, S.-Y. W., & Pohl, N. F. (1977). Student achievement in computer programming: Lecture vs. computer-aided instruction. Journal of Experimental Education, 46, 66-70.

Vaughn, A. C. Jr. (1977). A study of the contrast between computer-assisted instruction and the traditional teacher/learner method of instruction in basic musicianship. Dissertation Abstracts International, 38, 3357A. (University Microfilm No. 77-25,414)

Wolcott, J. M. (1976). The effect of computer-assisted instruction, traditional instruction, and locus-of-control on achievement of beginning typewriting students. Dissertation Abstracts International, 37, 1942A. (University Microfilm No. 76-22,070)

Wood, L. J. (1976). Computer assisted test construction in the BYU Library School. Brigham Young University, UT. (ERIC Document Reproduction Service No. ED 144 602)