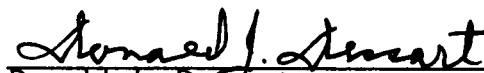
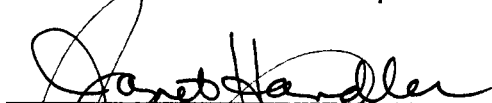


To the Graduate Council:

I am submitting herewith a dissertation written by Ray Hembree entitled "Model for Meta-Analysis of Research in Education, with a Demonstration in Mathematics Education: Effects of Hand-Held Calculators." I have examined the final copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Education.


Donald J. Desart
Major Professor

We have read this dissertation
and recommend its acceptance:



H T Matthews

Accepted for the Council:


The Graduate School

MODEL FOR META-ANALYSIS OF RESEARCH IN EDUCATION,
WITH A DEMONSTRATION IN MATHEMATICS EDUCATION:
EFFECTS OF HAND-HELD CALCULATORS

A Dissertation
Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Ray Hembree

August 1984

Copyright © Ray Hembree, 1984

All rights reserved

ACKNOWLEDGMENTS

To the committee, Dr. Janet Handler, Dr. H. T. Mathews, and Dr. John Peters, for kindness, courage, and counsel . . .

To Dr. Donald J. Dessart, committee chairman, for eternal willingness to share his expertise and wisdom, not only in working this study but throughout my graduate experience . . .

To three good ladies at the Main Library, The University of Tennessee, Knoxville--Mary Frances Crawford, Margaret Humphries, and Flossie Wise--for patiently scouring for and securing reports of research . . .

To the family, present and not . . .

Many thanks.

R. H.

ABSTRACT

During the past quarter century, there has occurred the "knowledge explosion," in which the findings of empirical research have increased exponentially. A demand has thus been created for the combining of results from kindred studies. The method of meta-analysis seems best suited for the work of research integration. However, the process is new, its nature unique. Its theory has developed piecemeal from a number of researchers, whose descriptions of methodology have appeared in diverse locations. There has existed no reference source which unifies all elements of the method. There have appeared no set of procedures for actually doing a meta-analysis and no evaluative guidelines for self-critical appraisal of the effort being accomplished.

The present study was designed to bridge these voids. A model for meta-analysis of research in education was constructed, and explicit guidelines were derived for reporting and evaluating reports of meta-analysis. These tasks proceeded essentially through the synthesis of methodological literature, with particular attention to the works of Glass, who invented the method, and of Hedges, who provided statistical foundation for the process. Verbal text and procedural flow charts specify the model and describe an orderly process of meta-analysis from inception through conclusion.

To challenge the model, its procedures were applied to the synthesis of research on the effects of hand-held calculators in mathematics education, Grades K-12. Seventy-nine research reports yielded

529 effect magnitudes related to student achievement and attitude. Findings indicated that, except in Grade 4, a use of calculators in concert with traditional mathematics instruction will improve the average student's basic skills with paper and pencil, both in working exercises and in problem solving. Sustained calculator usage in Grade 4 appears to damage students' basic skills. Across all grade and ability levels, students using calculators will possess a better attitude toward mathematics and an especially better self-concept in mathematics than noncalculator students.

The model's procedures were clearly effective in the calculator study. It was considered that the model is a tool of significant value in the synthesis of studies, which, in light of the Knowledge Exploration, seems an area of major growth in future research.

TABLE OF CONTENTS

CHAPTER	PAGE
I. INTRODUCTION	1
Statements of Purpose	7
Organization of the Study	7
II. REVIEW OF RELATED LITERATURE	9
Traditional Narrative Review	9
Traditional Voting Method	11
The Cluster Approach	13
Cumulation of p -Values	13
Summary of Previous Methods	15
Meta-Analysis	15
The Glass Approach	17
Refinements	21
The Hunter-Schmidt Approach	23
Critique	27
Criticism	29
Survey of Applications	32
Conclusions	36
III. MODEL DEVELOPMENT	39
Purpose and Orientation	39
Definitions of Terms	40
Initial Assumptions	41
Conceptual Framework	42
Topic Selection	43
Framing Research Questions	44
Collecting the Body of Studies	46
Coding the Studies	49
Representing Independent Variables	50
Representing Dependent Variables--Experiments	54
Representing Dependent Variables--Correlations	62
Organizing the Coding	63
Paradigm for Data Collection	66
Data Analysis--Experiments	68
Definitions of Terms	69
Assumptions	70
Preparing for Data Analysis	72
Simple Description of Effect Magnitudes	77
Testing the Homogeneity of Effect Size Groups	77
Calculating Means and Confidence Intervals	81
Testing Homogeneity Within Effect Size Classes	82
Analyzing Heterogeneous Effect Size Groups	84
Comparing Means of Effect Size Groups	87

CHAPTER	PAGE
Data Analysis--Correlations	88
Summary of Calculations	91
Paradigm for Data Analysis	94
Presenting and Interpreting Results	97
Preparing the Report	99
Model Specification	100
IV. REPORT OF META-ANALYSIS: EFFECTS OF HAND-HELD CALCULATORS IN MATHEMATICS EDUCATION	104
Introduction	104
Problem Statement	108
Organization	108
Review of Related Reviews	109
Operational Definitions	115
Research Questions	118
Method	122
Delimitations	123
Limitations	123
Collecting the Body of Studies	124
Coding the Studies	126
Features of the Data	134
Results	138
Overview of Procedures	138
Presentation of Findings	142
Summary and Discussion	171
Conclusions and Recommendations	177
Recommendations for Classroom Usage	178
Recommendations for Curriculum Development	179
Recommendations for Future Research	180
Evaluation	182
V. LOOKING BACK AND LOOKING FORWARD	184
REFERENCES	186
APPENDICES	195
A. BIBLIOGRAPHY OF STUDIES USED IN THE META-ANALYSIS	196
B. RESULTS OF DATA COLLECTION	205
C. MICROCOMPUTER PROGRAM FOR DERIVING STATISTICS RELATED TO EFFECT SIZE GROUPS IN META-ANALYSIS	223
VITA	229

LIST OF TABLES

TABLE	PAGE
1. Summary of Basic Calculations	92
2. Distinctions Among Achievement Effects	119
3. Numbers of Effect Sizes in Subsets for Achievement Dependent Variables	137
B-1. Data Tabulations for Independent Variables by Study Number	207
B-2. Effect Sizes by Study Number for Achievement Dependent Variables Related to Skills Acquisition (Equivalent Instruction Except for Calculators)	211
B-3. Effect Sizes by Study Number for Achievement Dependent Variables Related to Skills Retention (Equivalent Instruction Except for Calculators)	216
B-4. Effect Sizes by Study Number for Achievement Dependent Variables Related to Skills Transfer (Equivalent Instruction Except for Calculators)	217
B-5. Effect Sizes by Study Number for Achievement in Acquiring Estimation Skills (Equivalent Instruction Except for Calculators)	217
B-6. Effect Sizes for Attitude Dependent Variables, Study Nos. 4-24 (Equivalent Instruction Except for Calculators)	218
B-7. Effect Sizes for Attitude Dependent Variables, Study Nos. 27-53 (Equivalent Instruction Except for Calculators)	219
B-8. Effect Sizes for Attitude Dependent Variables, Study Nos. 56-77 (Equivalent Instruction Except for Calculators)	220
B-9. Effect Sizes by Study Number for Achievement Dependent Variables Related to Skills Acquisition (Special Instruction with Calculators versus Traditional Instruction without Calculators)	221
B-10. Effect Sizes by Study Number for Attitude Dependent Variables (Special Instruction with Calculators versus Traditional instruction without Calculators)	222

LIST OF FIGURES

FIGURE	PAGE
1. Example of Data for Meta-Analysis	20
2. Preliminary Checklist for Evaluating Reports of Meta-Analysis	34
3. Individual Coding Sheet	64
4. Flow of Procedures in Data Collection	67
5. Ratio of Biased/Unbiased Estimators as a Function of Degrees of Freedom	75
6. Example Box-and-Whiskers Plot for Effect Size and Various Outcome Constructs	78
7. Master Coding Sheet	93
8. Flow of Procedures in Data Analysis	95
9. Checklist for Evaluating Reports of Meta-Analysis	101
10. Specification of Procedures in Meta-Analysis	103
11. Example Coding Sheet for Meta-Analysis	133
12. Format of Data Analysis Summary Figures	144
13. Analysis Summary: Maintenance Effects Regarding Achievement in Acquiring Composite Operational Skills	148
14. Analysis Summary: Extension Effects Regarding Achievement in Acquiring Composite Operational Skills	150
15. Analysis Summary: Maintenance Effects Regarding Achievement in Acquiring Computational Skills	152
16. Analysis Summary: Extension Effects Regarding Achievement in Acquiring Computational Skills	153
17. Analysis Summary: Maintenance Effects Regarding Achievement in Acquiring Conceptual Skills	155

FIGURE	PAGE
18. Analysis Summary: Maintenance Effects Regarding Achievement in Acquiring Composite Problem Solving Skills	156
19. Analysis Summary: Extension Effects Regarding Achievement in Acquiring Composite Problem Solving Skills	158
20. Analysis Summary: Extension Effects Regarding Achievement in Acquiring Problem Solving Productivity	160
21. Analysis Summary: Maintenance Effects Regarding Achievement in Acquiring Problem Solving Selectivity	160
22. Analysis Summary: Extension Effects Regarding Achievement in Acquiring Problem Solving Selectivity	161
23. Analysis Summaries: Effects Regarding Achievement in Retaining Operational Skills	163
24. Analysis Summary: Extension Effects Regarding Achievement in Retaining Composite Problem Solving Skills	164
25. Analysis Summary: Maintenance Effects Regarding Achievement in Transferring Composite Operational Skills	166
26. Analysis Summary: Effects Regarding Achievement in Acquiring Estimation Skills	166
27. Analysis Summary: Effects Regarding Attitude Toward Mathematics	167
28. Analysis Summary: Effect Regarding Anxiety Toward Mathematics	169
29. Analysis Summary: Effect Regarding Self-Concept in Mathematics	169
30. Effects with Special Calculator Instruction: Achievement in Acquiring Skills; and, Attitude Toward Mathematics	170

FIGURE	PAGE
31. Summary of Achievement Effects Regarding Skills Acquisition, Retention, and Transfer	172
32. Checklist for Evaluating Report of Meta-Analysis on Effects of Hand-Held Calculators	183

CHAPTER I

INTRODUCTION

The tale has been told of a traveler who one day found a stone beside the road (Shah, 1968, pp. 110, 111). Inscribed on the stone was a cryptic suggestion, "TURN ME OVER AND READ." The traveler looked at the other side and found there written, "WHY DO YOU SEEK MORE KNOWLEDGE WHEN YOU PAY NO HEED TO WHAT YOU KNOW ALREADY?"

This story is related to the current state of educational research. During the past quarter century, there has occurred the "knowledge explosion," in which the findings of empirical research have increased exponentially. In 1974, Maccoby and Jacklin reviewed the existing research on the psychology of sex differences--1,600 works in print before 1973. In the field of educational research, a multitude of topics could be named whose literature numbers hundreds of studies, e.g., reading instruction, class size, ability grouping, and individualized instruction (Glass, 1978). From 1978 to date, selected subject areas in the annual bibliographies of the Journal for Research in Mathematics Education display the following numbers of studies: achievement evaluation--351; attitudes--160; calculators/computers--206; diagnosis/remediation--217; organizing and grouping--464; and, problem solving--203. Purely with regard to high school algebra, Dessart has identified six bodies of research containing significant numbers of studies: classroom instructional practices--88; attitudes and affective factors--

40; problem solving--42; computers and calculators--39; and, testing and prediction--36 (Dessart and Suydam, 1983).

Such an abundance of literature should stand to resolve at least some issues; else, what is the point of performing research? Then why are "final words" on any topic seldom heard? Pillemer and Light (1980) suggest that reports of research but rarely provide unequivocal answers. They follow instead a predictable format:

A statement of the problem is followed by a narrative literature review. Then the study results are reported, often independently of this review. Finally a conclusion is offered, with the familiar caveat that "more research is needed." (p. 176)

Despite the volume of research literature, the knowledge base of the behavioral sciences has failed to expand accordingly (Hedges, 1982d). Hence, amid what ought to be plenty, researchers do often appear to be starving, impelled to seek more knowledge (as suggested by the stone), while paying scant heed to what has already been found.

Why should this be true?

An answer appears to reside with two basic problems, difficulties common to research in years gone by, but magnified more and increasingly more by the Knowledge Explosion. The first involves a becoming aware of all the research that bears on a topic. This problem accrues from the way research in education progresses. The research enterprise is built like a house that sprawls across a vast terrain, rough-hewn, haphazard, and variegated, having no planned architectural scheme; like Topsy, it "just grew." Reports of research are its furnishings, which are often hard to find, "strewn among the scree of

a hundred journals" or lying about in the "unsightly rubble of a million dissertations" (Glass, McGaw, and Smith, 1981, p. 11) or tucked away unpublished and uncatalogued in file drawers, cubbyholes, and niches. Locating studies amid such clutter seems always a challenge, looming as more of a dare as knowledge expands. Still, the problem gives mostly discomfort. Through effort and patience, reviewers may find the literature related to their purposes. Then the second problem emerges: how to synthesize the findings of perhaps a hundred studies, all bearing upon a particular topic.

Here exists a problem that is critically compounded by the Knowledge Explosion. Within a single collection of studies, an awesome array of outcome statistics may be encountered--t-tests, variance analyses, point-biserial and rank-order correlations, phi coefficients, contingency tables, chi squares, and more. A rigorous integration of findings would necessitate transforming such varied statistics to some sort of common standard. This, of course, is no small chore, and so reviewers of research have usually by-passed the outcome statistics in favor of significance levels as the common metric (Glass, 1978). If these appear consistent across studies, then the synthesis of findings is said to be done.

However, such cases seem rare indeed and stand to be more precious still as knowledge increases. More likely, the findings will fail to agree; sometimes they even contradict each other. Then the reviewer is charged with resolving why this lack of correspondence should exist. Reasons and causes will be submerged within the matrix

of research designs, measuring instruments, subject samples, and other properties of the studies. These characteristics, Glass assures us, vary with "bewildering irregularity" from study to study, even though the basic topic is the same (Glass et al., 1981, p. 12).

There may be observed a paradox regarding knowledge in expansion, offering closer approaches to truth while hiding truth within amassing quantities of evidence. The limits of time and level of effort may often preclude evaluating all the features of each study in a set of many studies. Under such limits, as Jackson (1980) suggests, reviewers may consider only a part of the full set of studies, resulting in an arbitrary and misleading synthesis of findings. Perhaps they will be led to perform more research, even when their questions have already been answered. Walberg and Haertel (1980) have warned that concentrating energies on additional primary research without a systematic integration of previous findings is "scientifically and educationally wasteful" (p. 5). Substantial groups of kindred studies now exist in education. It seems to behoove the field of research to move those groups toward data condensation. The problem remains: how to do it.

Our subject of concern then is research integration, the synthesis of conclusions from clusters of studies. This subject appears as no area newly discovered; reviews of research are fundamental in the behavioral sciences. Researchers presenting fresh evidence by tradition introduce their work by surveying the literature. Reviews are also presented as independent scholarly efforts, what Jackson (1980)

has called "integrative reviews" (p. 438). These offer no new evidence but construct overarching conclusions from a set of different but related studies.

To serve the conduct of such integrative efforts, a number of methods have been devised, whose order of emergence has clearly been swayed by the mass of research literature. During the 1940s and '50s, reviewers were likely to find maybe 10 or 12 studies that bore on a topic (Glass et al., 1981, p. 12). With so few studies, a rhetorical review seemed equal to the task of research integration. But by the late 1960s, swelling bodies of literature demanded methods more systematic than the narrative review. Techniques aligned with statistics appeared--voting, adding p's, adding z's, adding t's, and others. None of these techniques has been widely accepted. Hedges and Olkin (1980) have proved traditional voting severely flawed statistically and logically. In the other prominent methods, the scope seems too narrow for general value. Compounding these adverse conditions is a lack of formally presented methodology for research reviews. This void was noted as long ago as 1971 (Feldman; Light and Smith). It still exists (Jackson, 1980; Hedges, 1982d).

In 1976, a new approach to research integration was proposed, heralded as more powerful and more scientifically sound than the previous methods. Thus arose meta-analysis, used with apparent success in diverse areas of research, including the educational field. The process of meta-analysis seems to offer a best solution to the twice-repeated problem: how to integrate research by methods surmounting the Knowledge Explosion.

Large-scale reviews of research have been termed a "major scientific challenge" (Pillemer and Light, 1980, p. 176) and are considered a type of research whose significance at least equals the import accorded original studies (Feldman, 1971; Glass, 1978). The present dissertation's central thrust initially was projected as the meta-analysis of research upon a topic in mathematics education. Reasons issued from what is perceived as a comprehensive need for integration of educational research, using a method statistically sound and able to handle large bodies of studies.

That method appears to be meta-analysis. However, study of the method has shaped concerns for its validity in practice. (Precisely how concerns emerged will be detailed within the dissertation's literature review.) The process is new, its nature unique. Its theory has developed piecemeal from a number of researchers; thus, reports of methodology are found scattered in numerous sources. Voids exist within the method's formal presentation. There exists no synthesis of inputs from the various contributors, no common reference which unifies all elements of the process. There exist no explicit procedures for actually doing a meta-analysis. And there exists no comprehensive set of guidelines for report evaluation. This lack of specific techniques and procedures must aggravate the likelihood of inadvertent mischievous results from meta-analysis. Maybe the gaps should be closed.

Bridging the voids in research methodology seems consistent with a dissertation effort. Such a bridge may be projected as a general

model for meta-analysis, to specifically embrace the integration of educational research. This model, artfully presented, would enhance the understanding of and facility with the process of meta-analysis, render the process more accessible to rank-and-file researchers, and develop standards for reporting and evaluating the meta-analysis of educational research.

The intention to conduct a meta-analysis of research upon a topic in mathematics education was retained. "Effects of hand-held calculators" was selected as the topic. Although its central thrust was modified, it stands important in its own right.

Thus, the total study has tandem foci: bridging the voids in research methodology; and, doing a meta-analysis to challenge the bridge.

Statements of Purpose

The study was constructed to achieve the following purposes:

1. Develop a model of meta-analysis for integrating educational research;
2. Derive explicit guidelines for reporting and evaluating reports of meta-analysis;
3. Demonstrate the model by performing and reporting a meta-analysis of research upon the topic, "effects of hand-held calculators in mathematics education."

Organization of the Study

Results of the study are organized into five chapters.

Chapter I describes the rationale by which the study and its purposes were established.

Chapter II reviews the literature on the general theme of research integration. All prominent methods are described, with emphasis to the process of meta-analysis and its previous applications in educational research. The need of a model for meta-analysis has evolved from this literature survey. Thus, the review may be considered an elaboration of the study's rationale.

Chapter III develops the model and derives evaluative guidelines. Definitions, assumptions, and limitations related to the process of meta-analysis are presented in the context of the model.

Chapter IV reports the meta-analysis of effects of hand-held calculators in mathematics education. The rationale for selecting this topic introduces the report. Specific research questions are developed, and findings in their regard are presented as outcomes of meta-analysis.

Chapter V assesses the model in light of the meta-analytic report and forms concluding remarks with an eye to the future.

CHAPTER II

REVIEW OF RELATED LITERATURE

The general context surrounding the study is the synthesis of educational research. This literature survey overviews the older techniques most widely employed and contrasts their procedures with the newer process of meta-analysis.

Given the common occurrence of research reviews, it could be presumed that literature describing the subject would be extensive. Regarding the process of meta-analysis, such a presumption seems true. However, regarding the previous methods, a well-developed literature does not apparently exist. In his study of approaches to research integration, Jackson (1980) examined 39 books on general methodology in educational, psychological, and sociological research. He found only three which provided criteria for the integrative study, and no discussion exceeded two pages in length. A similar dearth appears in the journals (Jackson, 1980). Descriptions of older techniques must evolve substantially from the writings of researchers who have helped to develop the meta-analysis.

Traditional Narrative Review

The oldest and most popular means of research integration is the narrative review, also described as "literary," "qualitative," "non-quantitative," and "verbal" (Hunter, Schmidt, and Jackson, 1982, p. 129). Within this procedure, the reviewer accepts each study at

face value and, from his impressions, composes a theory which appears to reconcile the set of findings. No treatment by statistics is applied. The person engaged in such a review essentially becomes a "one-man jury" who subjectively determines the preponderance of evidence.

Where studies to be synthesized are small in number, the narrative review has sufficed well enough and will no doubt continue to be used. Especially in clinical settings, an impressionistic approach is often useful, providing descriptive richness which is difficult to capture in more quantitative summaries (Pillemer and Light, 1980). Nonetheless, where studies are many, problems arise with regard to the brain's ability to summarize multiple findings (Glass et al., 1981, p. 14). Studies will seldom precisely compare in subjects, measures, and designs. Findings will also frequently vary in contradictory ways. Such diversities tend to render a subjective integration too complex and interreactive for the unassisted human mind to process.

Facing such a prospect, reviewers in the past have usually opted for one of three approaches (Hunter et al., 1982, pp. 129, 130). The first is what Glass has called "pedestrian reviewing," where verbal synopses string themselves out in "dizzying lists" (Glass, 1976, p. 4). Second is an easing of the integrative burden by basing conclusions on only a subset of studies, one which may not represent the full collection. Third, the reviewer may actually try to integrate findings mentally and may incorrectly postulate their interactions. In a Cooper-Rosenthal experiment (1980), two independent groups synthesized the same set of studies, one by way of narrative approaches and

the other by statistical techniques. Conclusions by the two groups notably differed, even though the set of studies numbered only seven.

The nature of the narrative review seems inconsistent with the spirit of scientific tradition. Such reviews show little concern for standardized procedures. Conclusions seem guided, perhaps in large measure, by stereotyping and personal bias (Glass et al., 1981, p. 18). The explicitness required for replication seems hardly commanded. Thus, narrative reviews do not provide the systematic data integration that researchers will need in their designs of powerful future studies (Pillemer and Light, 1980).

Traditional Voting Method

Research integration imposes a problem of data processing upon the reviewer. The larger the number of studies involved, the larger the problem. To compensate, numerical methods have been devised, more or less related to statistical analysis.

The method least related yet the one most commonly used has been referred to as the voting method. In its traditional form, the method is a simple tabulation of the findings across studies in three categories: significantly positive, significantly negative, and nonsignificant. The largest group is then declared the winner (Light and Smith, 1971).

Vote-counting is attractive because the technique is easy to apply, requires only minimal data from each study to be integrated, and permits a merging of findings from the different studies (Hedges

and Olkin, 1980). However, voting is said to contain three critical deficiencies (Glass et al., 1981, pp. 94, 95; Hunter et al., 1982, pp. 131-133).

First defect: Sample size is disregarded. Voting treats all studies alike, and the fact is ignored that different sample sizes impose completely different meanings on the word "significant."

Second defect: Voting fails to produce effect size information. Effect size, commonly expressed as a descriptor of the difference between two means, is a statistic which depicts the magnitude or strength of a relation between variables. Even where voting correctly reveals the presence of a true effect, the method gives no inkling of the size of that effect, say, whether one treatment surpasses another by a nose or in a walk.

Third defect: Voting can and does encourage false conclusions. Hedges and Olkin (1980) have proved that, when phenomena under study produce small effects, the vote-counting method systematically fails to detect those effects. This result is caused by the low statistical power of significance tests when effects are small. In educational research, small effects seem more the rule than the exception (Hedges, 1982d). If teaching behaviors, for instance, relate to achievement, the magnitude of each effect is likely to be small (Gage, 1978). Moreover, under such conditions, the probability that voting will yield false conclusions actually increases as the number of studies increases (Hedges and Olkin, 1980). Contrary to expectations, the more data that we may examine, the greater the chance that we

shall be wrong. "Clearly," says Glass, "there is much that is unacceptable in research integration by means of vote counting" (Glass et al., 1981, p. 97).

The Cluster Approach

Light and Smith (1971) were pioneers in their demand for serious attention to the methodological/technical problems of research integration. They generalized an approach from cluster sampling to derive an analysis strategy for a set of similar experiments. The cluster approach requires the use of original data from all studies to be synthesized. This requirement for raw data limits the practical value of the Light and Smith solution (Hedges, 1982d). Wolins (1962) once requested raw data from 37 researchers whose works had been published during the previous three years. Five authors failed to respond, 21 claimed the data lost or destroyed, and two attached proprietary rights. Restriction to the use of original data may discard informative studies whose data cannot be accessed, though summary findings of the studies probably will remain (Glass et al., 1981, p. 14).

Cumulation of p-Values

Proposals to combine p-values as a means of research integration go back at least three decades (Baker, 1952; Jones and Fiske, 1953). Recent advocates of the approach are Rosenthal and his colleagues (Rosenthal, 1978; Rosenthal and Rubin, 1979). This procedure

cumulates significance levels across studies to produce an overall p-value for the set of studies as a whole. The basic technique, called the "Edgington method," derives a joint or combined probability from the equation,

$$P = \frac{(\sum_{i=1}^k p_i)^k}{k!} .$$

Here, k is the number of studies, and p_i is an observed p-level. If the value of P is "small," the reviewer concludes that true effects exist.

A better known model is the Fisher method, which uses the equivalence,

$$-2 \sum_{i=1}^k \log_e p_i \sim \chi_{2k}^2 .$$

Here, k is the number of independent studies with significance levels p_i through p_k . Chi square is then used to test the null hypothesis for $2k$ degrees of freedom and a desired significance level.

These approaches seem statistically defensible. Two major advantages are their computational simplicity and low informational requirements. They also allow a combination of p-value and effect size analyses. However, both models are limited with regard to the number of studies permitted (Rosenthal, 1978). Adding p's may only apply when the sum of observed p-levels is less than or equal to unity.

Summing the logs of p -values restricts the number of studies to five or less.

Summary of Previous Methods

Researchers appear in consensus that the older techniques of research integration contain significant flaws and limitations. The narrative review is deemed unscientific to a marked degree. Problems in voting, the cluster approach, and cumulation of p -values issue from statistical infirmity, failing to produce effect size measures, requiring original data, or being confined to small numbers of studies. It also may be observed that voting and joint p 's require a set of findings to be viewed as replications of the same experiment or correlational study. Hence, no effects may be assessed regarding features which may differ across studies, e.g., research design quality, sample sizes, school grade levels, and study durations. Use of the previous methods seems bound to surface rather than in-depth conclusions.

The approach called "meta-analysis" seeks to overcome the bulk of these deficiencies. This process, its procedures, strengths, and problems, and its applications in educational research will now be surveyed.

Meta-Analysis

The process, broadly defined, is a "statistical analysis of the findings of many individual studies" (Glass, 1978, p. 352). It differs from previous methods by assigning to each finding a standardized

effect size. From this tabulation are derived effect size means, related to the various questions posed for the meta-analysis. The process uses findings, not raw data. Thus, it is distinguished from primary analysis, the original scrutiny of data in a study, and from secondary research, the reanalysis of data from a primary study.

Since the mid-1970's, at least four researchers have worked on procedures under the rubric of meta-analysis. In 1976, Glass and his associates at the University of Colorado integrated the findings of some 400 studies in psychotherapy research (Smith and Glass, 1977). Also in 1976, Rosenthal of Harvard University cumulated evidence from several hundred studies of experimenter expectancy effects in behavioral research. Within the same time frame, Schmidt of George Washington University and Hunter of Michigan State University collaborated on the integrative analysis of multiple validation studies in personnel testing (Schmidt and Hunter, 1977). Glass, Rosenthal, and Schmidt-Hunter worked independently and without knowledge of the others' efforts. Despite their emphasis on different fields of interest, the techniques they developed and published seem remarkably complementary in direction and methodology.

The process by Glass was introduced first and has apparently taken precedence over Rosenthal's similar method. Hence, Glass may be acknowledged as, if not the parent, at least the godfather of meta-analysis. Hunter and Schmidt view their approach as an extension of Glassian meta-analysis, dealing with areas Glass omits (Hunter et al., 1982, p. 139). Hedges of the University of Chicago has become chief architect of supporting methods and theory.

The Glass Approach

Meta-analysis first was advanced in Glass's presidential address to the American Educational Research Association in 1976. After its introduction, Glass described his novel approach in journal articles (e.g., 1978), which he compiled with much elaboration in his monograph with McGaw and Smith (Glass et al., 1981). This survey of Glassian meta-analysis was extracted from that monograph.

Glass (Glass et al., 1981) describes his approach as "nothing more than the attitude of data analysis applied to quantitative summaries of individual experiments" (p. 21). This "attitude" embraces those considerations normally addressed within empirical research: problem selection; hypothesis formation; definition and measurement of constructs and variables; sampling; and, data analysis. However, most research proceeds with methods of data analysis clearly in view. Meta-analysis invites the reviewer to use the full power of statistical methods deemed appropriate to the task. Thus, to Glass, it is not a "technique," but rather a "perspective" which may utilize many techniques of analysis at the option of the reviewer (p. 21).

The process begins when a reviewer selects some topic of interest, whose body of research exists in journals, theses, project reports, dissertations, and unpublished literature. These studies need not replicate each other and indeed may address different questions, so long as the questions clearly bear on the topic of interest. [For example, Glass's study of class size effects concerned such questions as: "How does the effect of class size on achievement vary with the

ages of pupils?"; and, "How does this effect vary between reading and mathematics instruction?" (Glass and Smith, 1979).] The reviewer secures all available studies and assumes the resulting anthology is a sample from the topic's theoretical population.

Thus, the units of meta-analysis are findings from all studies. Systematic research integration demands that these findings be statistically combined. Since many findings have undoubtedly been obtained by different measuring instruments, worry that incommensurable outcomes are to be united may be foreseen. The novelty of Glass's approach derives from expressing each experimental finding as a "standardized mean difference" between a pair of treatment conditions (Glass et al., 1981, p. 102). This mean difference is directly an effect size. If different measures or tests are linearly equatable (highly correlated on a true-score basis), the standardized mean difference is a scale-invariant measure of treatment effect. That is, if the same population of test scores should be represented by tests that are linearly equatable, the resulting effect sizes would be identical.

The assumption, linear equatability of measures, is crucial to the process of meta-analysis. It seems a valid assumption, provided that the measures all apply to the same construct (say, achievement). Then the standardized mean difference is an index of effect magnitude independent of the particular test used to measure the construct.

As his standardized mean difference, Glass proposed an effect size delta, computed from study outcome data as the difference between experimental and control group means divided by a within-group standard deviation (Glass et al., 1981, p. 102):

$$\Delta = \frac{\bar{X}_E - \bar{X}_C}{s_x} .$$

The denominator term becomes the standardizing element for the numerator's mean difference. For s_x , Glass offers options. He prefers to use the control group standard deviation, but on occasion he has employed the square root of the treatment groups' pooled variance (p. 38). Thus, each finding from experimental studies must be cast in terms of means, standard deviations, and sample sizes; these are the data which must be derived. In some of the studies, data for computing " Δ " may not be directly provided, e.g., where study results are presented as no more than t or F statistics. Techniques exist for deriving " Δ " from such outcomes (pp. 114-147).

Glass (Glass et al., 1981) has also observed, the product moment correlation is a scale-invariant measure of relationship between two variables, i.e., the correlation does not change as a result of linear rescaling. Pearson's r , r^2 , or Fisher's z transformation may be used as indices of "effect" magnitude to integrate correlational studies (p. 147). Glass has derived convenient formulas for translating various metrics to Pearson's r (pp. 149, 150).

The meta-analysis thus may subsume experimental or correlational studies. Once the reviewer has secured all available reports of either kind, effect magnitudes for those studies are computed or otherwise derived. Light and Smith (1971) have argued that contradictions in similar studies are often a result of varying properties across studies.

In anticipation of data analysis to determine relationships between such properties and effect size, significant features for each study are recorded, e.g., research design quality, school grade level, study duration, and characteristics of measuring instruments. Figure 1 presents a fabricated example tabulation. This record ends the phase of data collection for meta-analysis.

Study No.	Independent Variables					Basic Data			Dependent Variable, Δ
	V_1	V_2	V_3	--	V_n	X_E	X_C	s_C	
1	36	2	0.7	--	52	9.7	6.2	4.1	.85
2	49	1	0.9	--	31	5.3	9.1	3.7	-1.03
k	9	3	0.6	--	19	8.3	4.5	4.6	.83

V_1 : Sample size

V_2 : Design quality (high = 3, med = 2, low = 1)

V_3 : Instrument reliability

|

V_n : Duration of treatment (hours)

Figure 1. Example of data for meta-analysis.

There then exists a sample of data analogous to the results from a single experiment. Analysis proceeds accordingly. Glass believes that the purpose of research integration is more descriptive than inferential; thus, the process should be directed toward the magnitude of effects, that is, effect size means. Initial products of data analysis

are the mean and standard deviation of effect sizes across studies. Where different but related questions are considered, a clustering of data into subsets will provide effect size means that correspond to the separate questions. Within each group, the various effect size values are deemed estimators of a fixed population parameter, and the mean effect size for its "sample" gives best estimate of that fixed parameter. A feature of Glassian meta-analysis is acceptance of effect size variance at face value. The assumption is implicit that this variance is real and therefore has a substantive or methodological explanation. Reasons are sought in the varying characteristics across studies, e.g., sample size, research design quality, or properties of the measuring instruments. This explains why Glass demands extensive coding of study properties a priori. The covariation of such study features as independent variables and effect size as the dependent variable may be analyzed by regression, factor analysis, correlation, or any method useful for extracting meaning from data. Results may be presented by direct comparison of means, use of percentiles, fitted curves, or other techniques selected by the reviewer.

Refinements

Upon its introduction, meta-analysis struck the fancies of a host of research integrators, and reports of meta-analysis began to appear.

Despite such immediate application of the techniques suggested by Glass, too little heed had been accorded fundamental statistical issues (Hedges, 1982a). Glass had proposed the estimation of a standardized

mean difference but had presented no sampling theory for his procedures. Hedges (1981) derived the exact distribution of Glass's estimator of effect size:

$$ES = \frac{\bar{X}_E - \bar{X}_C}{s_x} .$$

Hedges then used the exact distribution to show that Glass's estimator has a systematic bias dependent on df, the number of degrees of freedom for the standard deviation in the ES definition. The bias approaches zero when df is large but can be substantial when df is small (<10). Its result is an inflating of ES values. Hedges (1981) obtained a correction factor for converting a biased ES to an unbiased value (pp. 111, 115).

Hedges also has noted that, during the process of meta-analysis, reviewers have not always distinguished sample estimates of effect size from the population parameters they wish to estimate. Meta-analysis involves a combining of ES values into means, often derived from a carving of data into subsets. Each mean ES has usually been taken as an estimated population value, a defensible assumption when all studies in the subset share a common effect size parameter, that is, when ES values are homogeneous in the subset. However, if those ES's are heterogeneous, two or more population values are embedded within the collection, and a need to search for explanations (i.e., moderating variables) is then indicated. Tests for homogeneity have been published (Hedges, 1982b; Rosenthal and Rubin, 1982). These tests

examine whether an observed ES variance is larger than would be expected from sampling error. The Hedges technique permits assessing homogeneity both within and between ES subsets. It also provides equations for setting confidence intervals about ES means of homogeneous subsets.

While most of the work in meta-analysis has been devoted to experimental studies, correlational integrations are entirely proper. Techniques have been derived for aggregating correlations across studies and testing the groups for homogeneity (Hedges and Olkin, 1983). This article also gives an analogue to regression analysis, useful in the search for moderating variables after a test for homogeneity fails.

Hedges (1982d) has collected his meta-analytic studies into a monograph for convenience of reference.

The Hunter-Schmidt Approach

Efforts in meta-analysis by Hunter and Schmidt, like those of Glass, first appeared in journals (e.g., Schmidt and Hunter, 1977; Schmidt, Hunter, Pearlman, and Shane, 1979). Their early work was confined to correlational studies, but this domain was later extended to include experiments as well. The Hunter-Schmidt method is also presented in monograph form (with Jackson, 1982), from which the present survey is extracted.

Hunter and Schmidt (Hunter et al., 1982) retain effect size as the core statistic for meta-analysis. In correlational synthesis, Pearson's r

again becomes the common metric. For experimental studies, each finding is transformed to an effect size (p. 98):

$$d = \frac{\bar{Y}_E - \bar{Y}_C}{s} .$$

Corresponding to Glass's ES, the numerator is the difference between experimental and control group means. Denominator s expressly is the standard deviation from the pooled experimental and control group variances. From a tabulation of effect size metrics may be calculated means and standard deviations as desired. These statistics, unlike those of Glass, are weighted a priori for sample size variations across studies.

Also in contrast with Glass's approach, Hunter-Schmidt meta-analysis does not accept effect size variance at face value (Hunter et al., 1982, p. 139). Early in their work with data cumulation, Hunter and Schmidt discovered that much or most effect size variance is a result of man-related errors, rather than variations in study properties. Such errors have been termed statistical artifacts, seen to accrue from five sources: (1) sampling error; (2) differences across studies in the reliability of measuring instruments; (3) study differences in range restriction; (4) differences in instrument validity; and, (5) computational, typographical, and transcription errors. The latter two cannot be easily corrected. Hunter-Schmidt developed formulas for subtracting from the total observed ES variance contributions due to sampling error, instrument unreliabilities, and range restriction

differences. It ought to be noted, when all three artifacts were corrected in actual studies, sampling error was found to contribute more than 90 percent of artifactual variance (Schmidt, Hunter, and Caplan, 1981). Thus, the researchers were led to declare that if only sampling error were considered and the other artifacts ignored, conclusions from meta-analysis would not be changed.

In terms of population parameters, total ES variance is corrected for sampling error by the equation (Hunter et al., 1982, p. 102):

$$\sigma_{\delta}^2 = \sigma_d^2 - \sigma_e^2 .$$

Here, the population variance σ_{δ}^2 , associated with the true effect size δ , is predicted as the difference between the (estimated) total observed ES variance σ_d^2 and the (estimated) sampling variance σ_e^2 . For simplicity, let us consider variance contributions from the other correctable artifacts as part of σ_e^2 . Then, once all effect sizes have been computed, the first order of analysis is a test of the hypothesis that effect size variance across studies entirely accrues from artifactual errors, that is, $\sigma_{\delta}^2 = 0$. Two guidelines prevail (Hunter et al., 1982, pp. 140, 141):

1. Hypothesis is accepted if the three correctable artifacts contribute 75 percent or more of the total observed effect size variance σ_d^2 . The uncorrectable artifacts are then judged the cause of residual variance σ_{δ}^2 , and the true effect size variance equals zero. Effect size is taken as constant across

studies at the mean. Thus, the effect size mean becomes the sole and final product of meta-analysis.

2. Hypothesis is rejected if the three correctable artifacts contribute less than 75 percent of the observed ES variance. To discover the cause of residual variance, properties that vary across studies are coded and correlated with effect size. Two outcomes are possible:
 - a. All such correlations may be trivial in magnitude. The residual variance is then considered the result of uncorrectable artifacts. This residual is used to construct a confidence interval about the effect size mean. The resulting distribution is the final product of meta-analysis.
 - b. Some correlations may imply that variations in study properties are related to the residual variance. By regression, these effects are isolated and interpreted. An estimated true mean effect size is then computed, with its standard deviation corrected for artifactual errors and for variations of study properties from their average (quantified) values. A confidence interval is constructed about the estimated true mean effect size, and this distribution is the final product.

The Hunter-Schmidt monograph, like the similar work by Glass, provides no sampling theory for its procedures. Their method seems related to a model of "random effects" proposed by Cronbach (1980).

In the context of this model, each treatment site (or study) is viewed as a sample realization from a universe of related treatments. When reviewers examine "replications" of a treatment across studies, they actually are observing a variety of treatments, differing as a result of different study characteristics. Each treatment site is considered as having its own unique true effect. An effect size group then estimates, not one fixed parameter as the Glass approach presumes, but rather a distribution of true effects. The variance of this distribution may be designated σ_{δ}^2 , with δ representing true effect. Whether this variance differs from zero is examined by correcting the total observed ES variance for sampling error, as in the Hunter-Schmidt equation,

$$\sigma_{\delta}^2 = \sigma_d^2 - \sigma_e^2 .$$

If σ_{δ}^2 indeed is zero, the set of ES values share a common effect size δ , and the model of random effects is equivalent to one of fixed effect. If $\sigma_{\delta}^2 \neq 0$, the set of ES values is heterogenous and suggests a population distribution of true effects.

Hunter and Schmidt do not provide a test for the hypothesis, $\sigma_{\delta}^2 = 0$. However, Hedges (1983) has derived a statistical test for this hypothesis, based on the model of random effects.

Critique

The Glass and Hunter-Schmidt methods differ mainly in their attitudes toward population parameters. Glass presumes that, in a set of ES values, all estimators relate to one fixed true population effect.

The Hunter-Schmidt method appears to consider any ES group as estimators for a distribution of true effects. However, the two approaches yield analogous results regarding ES means. A test for homogeneity under the fixed-effect criterion assesses whether an observed ES variance could have arisen by chance (sampling error). This test directly corresponds to testing the hypothesis that $\sigma_{\delta}^2 = 0$ in the random-effects paradigm (Hedges, 1983). Where any ES group is heterogeneous, both approaches rely on a search for moderating variables to explain the residual variance.

Regarding statistical soundness, the Hunter-Schmidt approach appears less rigorously proved than Glass's method as refined by Hedges. Hunter and Schmidt have based their approach on the hypothesis that ES variance across studies is due entirely to artifactual errors. The cutoff point for acceptance/rejection is set at 75 percent, that is, accept the hypothesis if the correctable artifacts contribute 75 percent or more of the total observed ES variance; otherwise, reject. This seems an arbitrary figure. It may be wondered whether ES variance should not be examined more rigorously. Moreover, confidence intervals for ES means in the Hunter-Schmidt method must be derived from an assumed distribution of the population effect size δ . This parameter symbolizes true effect. Since true effect in any treatment site is never known, it must be predicted from sample data and of course is prone to sampling variations. Confidence intervals based on an assumed distribution of δ thus become suspect.

Both the Glass and Hunter-Schmidt approaches fail to explicate procedures for the actual performance of meta-analysis. This void

exists almost by design. Since the inception of meta-analysis, Glass has emphasized the process as an "attitude" or "perspective" and not a technique (Glass et al., 1981, p. 21). Thus, he avoids prescribing procedures. His monograph, though extensive, is essentially a collection of disjoint routines, e.g., for locating studies, computing an effect size when group means are missing, transforming various metrics to a standardized effect size, and expanding dichotomous variables to continuous outcomes. Potential techniques are described for data analysis, but Glass expressly remands the treatment and interpretation of effect size aggregations to the judgment of the analyst-reviewer (Glass et al., 1981, p. 21). As a consequence, reviewers have been forced to rely on ad hoc methods (Cooper, 1979; Cohen, 1981). Hunter and Schmidt are more definitive. Still, they limit explicit procedures to those for handling statistical artifacts. Other procedures, they generalize.

Criticism

There seems a widespread positive regard for research integration by meta-analysis (e.g., Cahen, 1980; Leviton and Cook, 1981). Problems nonetheless attend the process, and hatchets of censure have often been raised. Major criticisms may be grouped within four classes: "apples and oranges"; poor research designs; publication bias; and, "lumpy" data.

The first of these reflects a concern that meta-analysis, in trying to make different studies resolve the same issue, must equalize the unequal, that is, the process mixes "apples and oranges." This worry

issues from a belief that data aggregation may only be done legally with studies which are the "same" in certain regards (Light and Smith, 1971). There surely can be no claim that studies must be the same in all regards, for then there would be no need to integrate studies; all would display the same effect within statistical error (Glass et al., 1981, pp. 218, 219). Therefore, only "different" studies need integration. Glass points out, analysis is habitually performed where data cumulates from different persons in a sample. These subjects in behavior also seem like "apples and oranges." Therefore, says Glass, the critics of meta-analysis must present convincing arguments that pooling data across persons is qualitatively different from cumulating findings across studies (Glass et al., 1981, p. 220). Apparently, no such arguments have appeared.

A use of all available studies is the blood and bone of meta-analysis. Concern has accrued from that usage, and the process is charged with relying on indiscriminating volume rather than the quality of findings (Bandura, 1978; Eysenck, 1978). Would not a better approach reject in advance all studies perceived as methodologically flawed (Mansfield and Bussey, 1977)? Glass and Hunter-Schmidt believe this suggestion impractical in two regards (Glass et al., 1981, pp. 220-226; Hunter et al., 1982, pp. 150, 151). First, not many studies will remain if every less-than-perfect report is rejected. Second, rejecting studies a priori follows an assumption that flawed research must surely produce flawed conclusions. This assumption may well be false and rarely is subjected to empirical testing.

Both Glass and Hunter-Schmidt have strongly advised investigating weak research designs, but a posteriori. If a significant relation is found between research design quality and effect size, weaker studies should then be rejected.

Most disciplines of research are charged with showing a somewhat political bias in what they choose to publish, leaning toward strong and/or popular findings (Mansfield and Bussey, 1977). Hence, it is expected that systematic differences will exist between corresponding results from journals and from dissertations. Since meta-analysis largely depends on information from these sources, critics maintain that its conclusions must be similarly biased. Moreover, the bias may be quite large. A study by Mary Lee Smith (1980) has shown corresponding effects in journals exceeding those in dissertations by 33 percent. However, the existence of such bias serves less to criticize meta-analysis than to indict publication practices. Meta-analysis handles the problem by correlating mode of publication and effect size. If that correlation proves to be significant, the literature may then be divided for separate meta-analyses by mode of publication.

Of all the technical criticisms of meta-analysis, Glass feels only one to be cogent, that nonindependent or "lumpy" data are frequently used (Glass et al., 1981, p. 229). Lumpy data will issue from studies with multiple findings, if more than one effect size derive from the same set of subjects. The use of dependent data implies a possible reduction in the reliability of ES means. Its consequences may

statistically be reduced, though not erased. A simple approach employs the calculation of an average effect size for each group of dependent data. This solution seems appropriate if the number of clusters of dependent data is small compared to the overall body of studies. However, if that number seems proportionately large, Hunter-Schmidt suggest a combining technique such as confirmatory factor analysis (Hunter et al., 1982, p. 117). For each cluster, this procedure yields an equivalent effect size for use in the meta-analysis.

Survey of Applications

Hunter-Schmidt meta-analysis and refining procedures by Hedges and Rosenthal have appeared but recently, so their acceptance remains to be tested. Most applications have used the Glass approach, with published reports occurring from such diverse research domains as psychotherapy, aggressive reaction patterns, transcendental meditation, and effects of television on social behavior.

The process has also impacted educational research, in topic areas such as class size, advance organizers, open classrooms, reading instruction, special education treatment techniques, peer influences, individualized instruction, and teacher effectiveness. Meta-analytic educational studies are reported in about two dozen dissertations and perhaps three dozen articles in journals.

The service of meta-analysis in mathematics educational research has been examined in detail. Reports were found by way of a computer search of the ERIC (Educational Resources Information Center) database and by manual searches of such abstracting services as

Dissertation Abstracts International. Two journal reports were located, with topics as follows:

Individualized instruction (Horak, 1981);

Computer assisted instruction (Burns and Bozeman, 1981).

Three dissertations were located, with topics as follows:

Individually paced instruction (Hartley, 1977);

Modern versus traditional mathematics (Athappilly, 1978);

Problem solving (Marcucci, 1980).

To survey the status of meta-analysis in mathematics educational research, general reviews of these reports would be required. Such reviews need guidelines for report evaluation, if judgments of adequacy are to rise above the subjective. However, no guidelines unique to meta-analysis have appeared. Indeed, as Jackson (1980) has noted, evaluative guidelines for the previous means of research integration have also failed to appear in published form.

To bridge the void, a preliminary set of guidelines for report evaluation was constructed from the writings of developers and critics of meta-analysis (Glass et al., 1981; Hunter et al., 1982; Jackson, 1980). As shown in Figure 2, each guiding item pertains to a criterion identified as essential to the meta-analytic process. Since most applications have used the Glass approach in the absence of Hedges' refinements, the list is restricted to items aligned with Glass's basic process: finding and using all studies that bear on a topic; quantifying common properties (independent variables) across studies; investigating correlations between the independent and outcome

Evaluative Criteria	Study Reference Number							
	1	2	3	4	5	6	7	8
1. Are questions/hypotheses clearly stated?	N	?	?	✓	?	N	N	✓
2. Is prior evidence (e.g., reviews) critically examined?	✓	?	✓	✓	✓	?	?	✓
3. Is the literature search process clearly reported?	✓	N	N	✓	N	✓	✓	✓
Is a bibliography of studies made available?	✓	N	✓	N	✓	✓	N	✓
Are reasons given for excluded studies?	✓	N	✓	N	N	✓	✓	N
Do the studies represent the full body of research?	?	?	?	?	?	?	✓	✓
4. Are study characteristics clearly coded?	N	?	N	N	N	✓	✓	✓
Are intercoder reliabilities reported?								
5. Is design quality quantified for each study?				✓			✓	✓
Are criteria for quality reported? Appropriate?	N	N	N	✓	N	?	?	?
6. Is the basic effect size (ES) specifically defined?	✓	N	✓	✓	✓	✓	✓	✓
Are all ES's listed in both magnitude and direction?	N	N	N	✓	✓	N	?	N
Are transformations to basic ES reported?	N	N	N	NA		✓	✓	✓
7. Were attempts made to locate missing data?					✓		✓	
8. Does analysis account for sample size variations?	N	?	N	N	✓	N	✓	N
9. Does analysis account for data dependence?	N	✓	N	NA	N	✓	N	N
10. Does analysis correlate study characteristics and ES?	N	✓	N	N	N	✓	✓	✓
Were independent variables surveyed for interactions?	N	N	N	N	N	✓	N	N
Was action taken to control for significant correlation?	N	N	N	N	N	✓	✓	N
11. Was action taken to control for poor research designs?	N	N	N	✓	N	N	✓	N
Were there at least a few good designs?				✓		✓	✓	✓
12. Is data cross-cutting adequately described?	✓	✓	✓	✓	✓	✓	✓	✓
13. If some aggregations were done with small numbers of ES, was capitalization on chance appropriately noted?	N	✓	N	N	N	N	N	✓
14. Does analysis account for mean ES variance?	N	N	N	✓	N	N	N	N
15. Are results of analysis clearly and fully presented?	?	✓	✓	✓	✓	✓	✓	✓
Are analytic methods appropriate?	?	?	?		?	?	?	✓
16. Are conclusions related to study questions?	?	✓	✓	✓	✓	✓	?	✓
17. Does the meta-analysis determine what the body of research reveals/does not reveal about the topic?	N	✓	N	?	✓	?	✓	✓
18. Does the report propose additional questions that need to be answered? Suggest future meta-analyses?	N	✓	N	N	N	✓	N	N
19. Does the report provide enough information so that a reader may critically examine the evidence?	N	N	✓	✓	N	N	N	✓

✓ means "yes" and implies strength

N means "no" and implies possible defect

? means "questionable"

blank means "information missing/not reported"

NA means "not applicable"

Figure 2. Preliminary checklist for evaluating reports of meta-analysis.

(dependent) variables; applying appropriate methods of data analysis; and, general research concerns.

For convenience in report evaluation, the guidelines of Figure 2 were formed as a checklist, with questions checked "yes" implying strength and other responses suggesting possible defects. The studies first examined by way of the checklist were the two journal reports of mathematics educational meta-analyses previously cited (Horak, 1981; Burns and Bozeman, 1981). Results are recorded in Figure 2 for study reference numbers 1 and 2. Both studies exhibit severe deficiencies with regard to central elements of meta-analysis: using full bodies of studies; coding characteristics across studies; examining the effects of poor research designs; accounting for sample size variations across studies; and, clearly presenting results of meta-analysis.

It was wondered if this observation would extend to areas of educational research besides mathematics. To investigate the wonder, a convenience sample of five journal articles was reviewed, whose topics included advance organizers, effects of coaching on SAT scores, educational seduction, computer based college teaching, and pretest sensitization effects in experimental designs. Results are displayed in Figure 2 for study reference numbers 3-7 (respectively, Luiten, Ames, and Ackerson, 1980; DerSimonian and Laird, 1983; Abrami, Leventhal, and Perry, 1982; Kulik, Kulik, and Cohen, 1980; Willson and Putnam, 1982). These studies also show basic deficiencies. Study 6 seems reasonably strong, performed by researchers who had conducted at least three previous meta-analyses.

Study 8 of Figure 2 is Hartley's (1977) dissertation. (The other two dissertations related to mathematics education were not available from their parent universities and thus were not reviewed.) Hartley's work seems well developed and presented, an apparent indication that meta-analysis, handled with care, can integrate educational research with sound results. However, reviews of the other reports suggest a general lack of facility with the process. Deficiencies may be partly explained by the newness/uniqueness of meta-analysis, and its usage should, like wine, improve with age. A weightier cause seems likely as well: the lack of explicit procedures for conducting a meta-analysis.

A general model for meta-analysis of educational research would perhaps relieve both causes. No such model has yet been described in methodological literature, the two-dozen dissertations of educational meta-analysis, or the three-dozen meta-analyses reported in educational journals.

Conclusions

Five methods have now been surveyed, attending the task of research integration: narrative reviews; vote-counting; clustering; cumulation of p -values; and, meta-analysis.

The older techniques appear doomed by their flaws. They seem statistically unsound or fail to yield effect magnitudes or need original data or lack the scope to handle large bodies of studies.

The newest approach has problems as well. A number of researchers have contributed to the process; thus, its methodological

literature is found in numerous sources, perhaps where educators seldom tread (e.g., Psychological Bulletin and American Statistician). There exists no single reference description which synthesizes the inputs of the various contributors. Explicit procedures do not exist for applying the process. There exist no guidelines for evaluating meta-analytic studies.

These voids may mask what Hunter has called the "quantum improvement" of meta-analysis over the previous methods (Hunter et al., 1982, p. 143). Jackson (1980) reviewed its sources of power, as follows:

First, [the meta-analysis] is a systematic, clearly articulated, and replicable approach to integrating findings from a set of studies. Second, it can be used with information from both the best and the less-than-best studies on a topic, but with controls for possible biases caused by various flaws in the available studies. Third, it can provide estimates of the population parameters. Fourth, when using multivariate statistical procedures, it provides a method for simultaneously investigating the relationships among studies' methods, population of subjects, scope conditions, duration of treatments, and findings. No approach used to date for integrative reviews has been capable of doing this. (p. 452)

It also may be noted that most of the problems encountered in meta-analysis are common to the older approaches. Researchers appear to concede that the process renders an important contribution to behavioral research, which will often prove valuable "when applied and interpreted with care" (Jackson, 1980, p. 455). It appears that these actions--application and interpretation--have caused some reviewers confusion, at least in educational research. Their plights are viewed as no indictment of the process. It appears essentially complete,

although fragmented in the literature. Work apparently needs to be done in better describing the method.

Another word on value may be in order. According to Glass (1978), the "fiction" dies hard that science may only progress along a string of dramatic experiments (p. 353). As a result, research integration has not yet reached the pinnacle of warm regard. In light of the Knowledge Explosion, maybe priorities ought to be changed. The best word on "value" seems a paraphrase of mathematician David Hilbert: "The importance of a scientific work can be measured by the number of previous publications it makes superfluous to read."

CHAPTER III

MODEL DEVELOPMENT

A meta-analysis of research is "the statistical analysis of the findings of many individual analyses" (Glass, 1978, p. 352). Its process begins with selecting a topic, passes through the phases of data collection and treatment of data, and ends with reporting conclusions. This chapter develops a model which describes and explicates that process as related to the integration of educational research.

Purpose and Orientation

Light (1980) has suggested, both science and art are involved with research integration. "Science" concerns the collection of data--securing the body of studies and recording their common properties, including that outcome data required for computing effect magnitudes. Within this domain, a model for meta-analysis may prescribe explicit procedures, since these seem quite standard.

In contrast, the aspect of "art" concerns analysis and interpretation of the study outcomes. Glass desires these actions to be left as much as possible to the option of the analyst-reviewer (Glass et al., 1981, p. 21). However, the treatment of data has been an area of hazard in past applications of meta-analysis, shaping concern that reviewers have not fully grasped the process. Pillemer and Light (1980) have declared that data synthesis, if uncritically applied, may

actually undermine the refinement of knowledge, stating conclusions that may not be sound but nonetheless discourage further study.

Thus, the model's basic purpose is to foster self-critical uses of meta-analysis. It develops paradigms for both the collection and treatment of data, while respecting Glass's wish for flexibility. The model is presented less to fix procedures than to describe in one source a statistically sound and orderly process from inception to conclusion.

The model's orientation is the synthesis of educational research. Restrictions will be applied to the model accordingly. First stress is put upon experimental studies, but the model will subsume correlational studies as well. A leaning toward simplicity will pervade the model. The Glass and Hunter-Schmidt approaches, while basically complementary, still involve some difference of opinion. Where conflict exists, the model will use those procedures which appear most logically or statistically persuasive. Where no such imbalance is seen to exist, the simpler or more practical method will be applied.

Definitions of Terms

Since Glass first coined the term in 1976, meta-analysis has been used in reference to research integration by almost any statistical method, including vote-counting (e.g., Johnson, Johnson, and Merayuma, 1983). This model will employ the definition:

1. Meta-analysis--The process of research integration established by Glass and extended by other contributors, present and

future. The process may subsume either experimental or correlational studies.

Its variables will be classified according to the broad definitions of Kleinbaum and Kupper (1978, p. 8):

2. Independent variable--One used to describe another variable. In either experimental or correlational meta-analysis, the independent variables are defined to be the common characteristics across studies, e.g., type of treatment, sample sizes, or study duration.
3. Dependent variable--One to be described by other variables. For experimental studies, the dependent variables are the types of constructs measured across studies, e.g., achievement or attitude. Each structural outcome is expressed as an effect size, defined as the difference between experimental and control group means divided by within-group standard deviation. In correlational meta-analysis, the dependent variables are relations between constructs, e.g., teacher effectiveness and student achievement, or teacher effectiveness and student attitude. These relations are described by correlation coefficients, taken (loosely) as effect magnitudes.

Initial Assumptions

General guidelines for model development have been selected for consistence with accepted statistical theory and with common properties of research in education.

1. It is assumed that study outcome data will possess the following forms:

- a. Experimental means and standard deviations or other statistics (e.g., gain scores) translatable to an effect size.
- b. Pearson's product-moment correlation coefficients or other statistics translatable to Pearson's r .

2. It is assumed that study outcomes which produce dependent variables have been drawn from continuous data.

This restriction seems justified because of: (1) the elemental occurrence of continuous outcome data in educational research; and, (2) the need to treat polychotomous data by exotic methods such as probit/logit transformations.

Independent variables may be either continuous or discrete.

3. It is assumed that all sample sizes of studies are equal to or greater than 10.

This restriction permits the use of large-sample approximations, enormously helpful in data analysis. Excluding small-sample studies may be further justified in its own right; the resulting effect magnitudes appear so imprecise as to have little value (Hedges, 1982d, p. 19).

Conceptual Framework

Jackson (1980) has identified six basic tasks for research integration:

1. Selecting the questions/hypotheses for the review;
2. Securing the studies to be reviewed;
3. Representing each study's properties and findings;
4. Analyzing the findings;
5. Interpreting results;
6. Reporting the review.

Cooper (1982) has listed analogous duties. These tasks correspond to the efforts pursued in primary research, and so the model will be developed within this conceptual framework, perceived as entailing three phases: data collection; data analysis/interpretation; and, reporting. Retaining the spirit of meta-analysis, procedures in each of these phases will be synthesized from methodological literature, with homage to Glass, Hedges, Hunter-Schmidt, and Jackson.

Topic Selection

The choice of an appropriate topic launches the process of meta-analysis. Two bases for topic selection exist. A subject area may become so clearly crammed with evidence, integration seems demanded by the field (or maybe by editors). This is the case of a topic in quest of a meta-analysis. On the other hand, a reviewer may wish simply to condense research findings in some area of interest and must therefore choose a subject. This is a meta-analysis in quest of a topic. Whatever the motive, three guidelines for topic selection are stated:

1. The reviewer should be reasonably acquainted with the topic being proposed (Pillemer and Light, 1980).

It seems deceptively simple to collect a group of studies and combine their outcomes. However, risk arises with the interpretation of findings. The more removed is the reviewer from topical context, the greater the danger.

2. The topic should be consistent with constraints of time and level of effort.

3. The literature must contain enough studies to represent the topic.

This latter condition must lead to the query, how many studies are needed? No specific answer can be stated. The number depends on how many concerns are being addressed; a one-question topic may require no more than 8 or 10 studies. A preliminary search of the literature seems wise. Mainly based on how much is there, the reviewer may determine the extent to which a topic is constrained by time and level of effort and judge if available evidence seems to represent the topic.

Framing Research Questions

Establishing the topic leads to problem formulation, where concepts become "operationalized" (Cooper, 1982, p. 292). In primary research, a statement of the problem is consistently the first component of a study's plan and its final report. This problem statement gives direction to the study. It generates tentative hypotheses which are refined to final form as a result of reviewing the literature.

Within the integrative review, problem definitions seem less formal but are nonetheless as crucial for study direction. Hypotheses typically are cast as research questions, which may be asked at two levels (Jackson, 1980). The first is comprehensive; it relates to a phenomenon itself, e.g., what is the effect of hand-held calculators on mathematics achievement? The second level considers variations in the properties of studies, which may account for variations in study results, e.g., how does the effect of hand-held calculators on achievement vary across Grades 1 through 12? If findings should prove consistent across studies, then the second-level question is superfluous. However, if findings should display apparent inconsistencies, second levels become essential for productive inquiry.

Regarding question formation, Jackson (1980) has noted four sources that will need to be considered:

1. Available theory that bears on a topic. Refining theory is of course a central purpose of research. In addition, theory may submit significant questions on relations among variables, framing not only the first-level query but also a subquestion series at the second level.

2. Previous reviews. A topic's research in full or in part may have already been reviewed in other integrations. Prior investigations will have asked important questions on the topic or related topics.

3. Intuition, insight, and ingenuity. The gauntlet hurled at research integration is finding order in apparent chaos. The resourcefulness of the reviewer can aid by suggesting new slants or connections. Intuition and insight thus are notable askers of

questions. It ought to be remembered, they become suspect when used to give answers.

4. The body of studies to be reviewed. Following topic selection, the reviewer usually starts with one or more broad questions and may then enlarge the set by way of the previous sources. However, the primary studies themselves become a vital source of questions and should therefore be reviewed before the question list is finalized.

No attention need be paid to whether the data will actually answer all questions. "A good integration, at the same time that it shows how much is known in an area, also shows how little is known" (Feldman, 1971, p. 100). Unanswered questions will guide the perception of needs in future research.

Collecting the Body of Studies

Preparations for meta-analysis--choosing a topic and asking its questions--have now been considered. Next is the task of study collection, which proceeds within two stages: (1) finding all studies with potential value; and, (2) selecting the final group or body for data analysis.

Finding all studies of possible worth becomes a test of fortitude and patience. The evidence of research must be plucked from a landscape of literature studded well with places to hide: journal articles, dissertations, theses, project reports, papers from scholarly meetings, reports to foundations, and unpublished studies. Still, the effort in meta-analysis corresponds to the literature search in primary studies,

differing mainly in degree. Cooper (1982) has recommended four techniques that researchers may use to retrieve information.

1. Computer searches. Libraries linked with nationwide computer networks have access to stored databases which contain research citations and abstracts. In educational research, two such databases seem particularly useful, Dissertation Abstracts International and ERIC (Educational Resources Information Center). Through entry of keyword descriptors for a particular database, the computer scans its indices and prints a bibliography and/or abstracts of relevant documents.

2. Abstracting services. Reviewers may search for document abstracts directly, e.g., in volumes of the Current Index to Journals in Education, Resources in Education, and Dissertation Abstracts International. Reliance solely on the computer tends to discourage "browsing" or following leads that arise during manual search.

3. The descendency approach. Some journals (e.g., the July editions of the Journal for Research in Mathematics Education) publish abstracts of research, in which citations are indexed according to subject. These bibliographies may be screened for topical relevance.

4. The ancestry approach. This technique retrieves information by tracking citations from study to study. As a group of studies assembles, their individual listings of references are resources for unearthing earlier research.

Other resources include the bibliographies of previous reviews and references secured from specialized educational agencies.

It must be faced, these avenues fail to guarantee total coverage (Feldman, 1971). A few published studies will no doubt be missed; unpublished research will probably go largely undetected. A representative study collection may only derive from patient, exhaustive, continuing tracks of available sources.

Once potential studies have been identified, their reports must be collected and reviewed. Journal articles are handy on library shelves. Interlibrary services may seek the loan of some dissertations and theses; others may be purchased in hard- or microfilm copy. Unpublished and other fugitive literature may perhaps be gathered from ERIC or the authors. It then remains to determine which studies are welcome in the final body for meta-analysis. Prior to the screening process, the reviewer should establish guidelines by which reports will be rejected. Some criteria will of course be topic-specific. The following general guidelines apply, by reason of the initial assumptions:

1. The meta-analysis depends on a provision of sufficient data for effect size calculations. Where reports omit such data, these may sometimes be secured through correspondence with the authors. Every effort to find missing data is urged. However, if critical data are missing and cannot be otherwise secured, the offending report must be stricken.

2. Studies with polychotomous outcome (say, "improved" or "not improved") should be excluded.

3. Studies with sample sizes of less than 10 subjects should be excluded.

An additional guideline is noted:

4. Results of a primary study may have been reported in more than one location. Since these reports are not independent, only one may be used in the meta-analysis.

Reports not rejected will comprise the terminal body of studies. Most specifically, no report should be excluded on methodological grounds, since the effects of poor research designs will be considered during the data analysis. A qualifying note should be observed. If a series of studies all share the same design flaw, consistent results across studies may be recorded and yet may err as an artifact of the flaw (Hedges, 1982d, p. 66). The body of studies should contain at least a modest number of acceptable research designs (Jackson, 1980). The meta-analysis of studies in which designs are consistently poor should not be attempted.

Coding the Studies

In meta-analysis, reports of primary research supply the data to be synthesized. These studies are similar to the extent that all relate to the topic selected.

However, treatments with similar labels probably will differ in a multitude of ways--research design, sample sizes, study duration, measuring instruments, school grade level, and so on. Variations in such properties will impart a likelihood of conflict in the findings across studies. All observed conflicts should be explained, that is, the meta-analysis must be designed to examine relations between the properties of studies and their outcomes.

Therefore, each study must be coded with respect to its characteristics and findings. Study properties take the role of independent variables (also known as "moderating" or "explanatory"). These apply regardless of the study's nature, whether experimental or correlational. Effect magnitudes derived from study outcomes are the dependent variable in meta-analysis. For experimental studies, an effect size calculates from summary data of experimental and control group means and standard deviations. In correlational studies, the basic metric is Pearson's r . This "effect" is usually available (or can be translated) directly from outcome statistics.

Representing Independent Variables

Identification of independent variables is the initial task for the coding of studies. Each study characteristic is viewed as a candidate variable, and all elections are based on the single criterion that the property may have affected study outcomes. As an example, Hartley's (1977) meta-analysis on the effects of individualized instruction in mathematics included the following independent variables:

Study Variables

- Year of study
- Type of publication
- Involvement of investigator

Student Variable

- Grade level
- Sex
- Ability level
- Socioeconomic status

Treatment Variables

Type of treatment
 Subject matter
 Length of treatment

Design Variables

Number of subjects
 Research design rating

Variables Specific to Tutoring

Age of tutor
 Ability of tutor
 Training of tutor
 Selection of tutor

(p. 11)

Some variables may be identified at the outset, e.g., type and length of treatment, year of study, type of publication, and involvement of investigator (an assessment of experimenter bias). Other variables may be discovered by probing the studies themselves.

For easing the later data analysis, the list of independent variables should be subdivided as categorical and continuous.

Categorical variables are nonordinal in nature; their dimensions are expressed in verbal terms. Their freedom to vary is restricted by: (1) assignments to groups; or, (2) qualitative assessment. In the first style, the assignment groups are clear polychotomies, e.g., for type of publication, the source of a study may be a journal, book, dissertation, or others. Dimensions of the second kind are judged by the reviewer, e.g., high/medium/low for research design quality.

Continuous variables are measured on familiar numerical scales, e.g., duration of study in "days." Values will be specified in the

reports. Thus, they need no further treatment except to assure that, for a particular variable, its scale across studies is common, e.g., all study durations in "days" or in "weeks."

Certain variables in both divisions may be viewed as mandatory, especially those related to research design quality and publication bias. These are the sources which, in past applications of meta-analysis, have been found most likely to influence effect magnitudes. Essential categorical variables include:

- Type of publication
- Type of treatment
- Type of measuring instrument (e.g., standardized)
- Research design rating

(It should be noted that the type of treatment may be constant, rather than variable, across studies.) Essential continuous variables include:

- Year of publication
- Duration of treatment
- Sample sizes
- Instrument reliability

It seems interesting to observe that continuous variables may be made categorical by grouping the data and labeling subdivisions of their scales, e.g., year of publication labeled as "ancient," "middle-aged," and "recent." Upon this two-way street, categoricals may be made quasi-continuous by assigning numerical values to subcategories (e.g., high = 3, medium = 2, low = 1) or by assigning dummy variables to polychotomies (e.g., 1 for western U.S., 0 for central U.S., -1 for

eastern U.S.). Such translations may be required during data analysis.

Once the variables have been identified, each study is coded accordingly. Values for the continuous variables may be recorded directly (or converted as required to assure common metrics). Categoricals whose nature is polychotomous may also be recorded directly. The remaining categoricals will require subjective judgments, and the reviewer should establish a priori guidelines for making those judgments, e.g., experimenter bias could be rated "much" if the researcher actively took part and "little" if he did not.

In particular, the method for rating research designs will need elaboration. Since the inception of meta-analysis, its use of poor as well as good designs has been a hardy bone of contention, reflecting beliefs that flawed designs are likely to yield flawed conclusions. These fears seem valid, and as a result, each meta-analysis should provide for assessing research design quality and its effect on study outcomes. Eight assessment criteria have been proposed (Stock, Okun, Haring, Miller, Kinney, and Ceurvorst, 1982):

1. Adequate problem definition;
2. Adequate description of a relevant population;
3. Appropriate sampling procedures;
4. Control of error sources;
5. Adequate measuring instruments;
6. Appropriate data analysis;
7. Reasonable conclusions;
8. Adequate report.

Yin, Bingham, and Heald (1976) have offered similar assessment criteria.

This model accepts their proposals and provides for a separate rating of each criterion, using scales selected by the reviewer (say, excellent = 3, average = 2, poor = 1). Resulting numbers may be summed and divided by 8 to give an overall rating for research design quality.

It may be (forlornly) hoped, each study used in the meta-analysis will contain information on all independent variables. Research in education seems hardly so mechanized. Gaps will probably exist, making the assessment of relationships between independent and dependent variables tenuous or impossible. The reviewer should make every effort to secure missing data from authors of study reports.

Representing Dependent Variables--Experiments

In experimental meta-analysis, dependent variables are the types of outcome constructs surveyed by the studies. For example, El-Nemr's (1980) comparative study of biology instruction by tradition and by inquiry included the following dependent variables: achievement; science process skills; critical thinking skills; laboratory skills; attitudes toward biology; and, interest in science. Each dependent variable is expressed as an effect size (ES), defined for each study outcome as: $ES = (\bar{Y}^E - \bar{Y}^C) / S^Y$. The superscripts refer to (E)xperimental and (C)ontrol groups, $\bar{Y}^E$ and $\bar{Y}^C$ are sample group means, and S^Y is within-group standard deviation. ES is thus a standardized mean difference, whose import seems conceptually simple--zero

ES

means no effect, a positive ES implies that findings favor the treated group, and negative results suggest the treated group has suffered.

A variation in basic ES should be considered. The ES base is the traditional experiment, where outcomes are derived from end-of-treatment observations for experimental and control groups. However, this base may be extended to include analogous outcomes, e.g., the gain scores commonly used in educational research. Gain scores result when pretest observations are subtracted from the corresponding post-test scores for the experimental and control groups. Effect size is then the difference between average gain scores divided by an estimate of standard deviation (Hartley, 1977, p. 30):

$$ES = \frac{\bar{G}^E - \bar{G}^C}{S^G} .$$

Two variants of the standardizing statistic S^Y or S^G have been proposed, the control group standard deviation S^C or the statistic S^P derived from pooling variances by the equation,

$$(S^P)^2 = \frac{(n^E - 1)(S^E)^2 + (n^C - 1)(S^C)^2}{n^E + n^C - 2} ,$$

where n 's are group sample sizes.

Glass (Glass et al., 1981, p. 107) prefers the use of S^C , based on the following argument:

Within one study (such as an analysis of variance), several treatments may be compared with the same control group. The

associated pooled variances $(S_1^P)^2, (S_2^P)^2, \dots, (S_k^P)^2$ will surely be different. If pooled variances then are used to calculate $ES_1, ES_2, \dots, ES_k$, resulting values will probably differ, even where mean differences are the same. That is, $\bar{Y}_1^E - \bar{Y}^C = \bar{Y}_2^E - \bar{Y}^C$ will not necessarily imply that $ES_1 = ES_2$. Thus, standardizing mean differences by S^C has the advantage of assigning equal ES's to equal mean differences. Furthermore, pooled variances rely on the assumption of homogeneous population variances, which may not be valid.

Hedges and Hunter-Schmidt prefer the use of S^P in finding ES (as does Cohen--1977, p. 20). They argue that there rarely seems a large difference between the sample variances (Hunter et al., 1982, p. 97); indeed, the t and F tests use S^P in assessing significance, under the assumption of homogeneous variances. Hedges (1981) has showed, the pooled S^P is a less biased estimator than S^C because of its smaller sampling error. Moreover, in many studies, a pooled variance is reported, while separate experimental and control group variances are not. Where this is the case, the Glass approach is forced to use S^P , thus mixing standardizing statistics across studies.

All these arguments have merit. However, the assumption of homogeneous variances is so commonplace in educational research, its prevalence seems to soften Glass's position. The Hedges and Hunter-Schmidt stand seems more persuasive.

Hence, effect size in the model will be defined as:

$$ES = \frac{\bar{Y}^E - \bar{Y}^C}{S^P} . \quad (ES-1)$$

For each study used in the meta-analysis, homogeneous population variances will be presumed (but the model will arrange to study the effects of that assumption). The variance $(S^P)^2$ is an unbiased estimate of the population variance σ^2 (Ferguson, 1981, p. 177). Since the sample variances, while assumed homogeneous, will probably differ in magnitude, S^P will not be viewed as representing either S^E or S^C , but rather σ .

Further sanctions for using S^P directly ensue.

1. Where a study's summary data provides the group means $\bar{Y}^E$ and $\bar{Y}^C$ and further provides for finding S^P , equation ES-1 is convenient to use. However, where an outcome's significance level has been determined by $\underline{t}$ or two-group $\underline{F}$ test, alternate methods exist for finding ES, even if no data other than $\underline{t}$ or $\underline{F}$ and sample sizes have been provided. Since $t = (\bar{Y}^E - \bar{Y}^C)/S^P(1/n^E + 1/n^C)^{1/2}$ and $ES = (\bar{Y}^E - \bar{Y}^C)/S^P$, then $t = ES/(1/n^E + 1/n^C)^{1/2}$. Thus:

$$ES = t \sqrt{\frac{1}{n^E} + \frac{1}{n^C}} . \quad (ES-2)$$

For the two-group $\underline{F}$ test, $F = t^2$. Hence:

$$ES = \sqrt{F} \sqrt{\frac{1}{n^E} + \frac{1}{n^C}} . \quad (ES-3)$$

In using $\underline{t}$ or $\underline{F}$ to find ES, it must be recalled that signs apply. ES is "+" if test results have favored the experimental group and "-" for results that favored control.

2. In a one-way analysis of variance (ANOVA) and analysis of covariance (ANCOVA), the within-group variance $(S^W)^2$ estimates the population variance, under the assumption of homogeneous variances across all groups. Since $(S^P)^2$ is defined to estimate the population variance under the same assumption, $(S^P)^2$ is $(S^W)^2$, whose value is normally given in summary tables. For j treatments (not counting control), each mean difference divided by $S^W = S^P$ give j ES values directly, i.e.:

$$ES_i = \frac{\bar{Y}_i^E - \bar{Y}_i^C}{S^W}, \quad i = 1, \dots, j \text{ treatments.} \quad (ES-4)$$

In the ANCOVA, adjusted group means will need to be used.

3. In a factorial ANOVA, within-cell variance $(S^W)^2$ estimates the population variance, under the assumption of equal cell variances and the assumption of the finite fixed ANOVA model normally used in educational research. Hence, $(S^P)^2$ is $(S^W)^2$, whose value is usually given in summary tables. As in ES-4, mean differences for the treatment cells compared with their controls, each divided by $S^W = S^P$, give ES values directly. An additional advantage may be observed. Where significant interactions are found, these may be flagged as potential moderating variables and assessed across studies as part of the data analysis.

These techniques seem straightforward, deceptively so. Problems may hover, confounding the effort for ES derivation.

1. Incomplete reporting. Each study report must contain enough data to calculate effect magnitudes. In studies using more than two groups, their means and a common S^P are required; otherwise, there seems little hope of recovering ES's from the reports (Glass, 1978). ES measures may occasionally be constructed, even where critical data have been omitted. In particular, some reports present no more than sample sizes and the indication that a mean difference was significantly positive or negative at a customary level, say, $p < .05$ or $p < .01$. A conservative approximation of ES may be derived by finding (in t tables) the corresponding critical- t for a two-tailed test and degrees of freedom $n^E + n^C - 2$. This value of t may then be applied in equation ES-2, such that:

$$ES = \text{critical-}t \sqrt{\frac{1}{n^E} + \frac{1}{n^C}} . \quad (ES-5)$$

The estimate may be made more precise if: (1) it is known that the test of significance was two-tailed or one-tailed; and/or, (2) an exact value of p has been provided. In the latter case, critical- t is derived by interpolation within the t table. Again, it must be recalled that ES is positive if test results have favored the experimental group and negative if results have favored control.

2. Data dependence. Meta-analysis is a process of statistically combining independent effect magnitudes. Through necessity or choice, educational researchers sometimes design experiments where outcome data are recorded two or more times from the same group of

subjects. In such designs, the one sample group is both experimental and control; data pairs have not been drawn from independent samples and are thus dependent pairs. Nonetheless, such data may be used in ES calculations, provided that statistical procedures for deriving $\underline{t}$ or $\underline{F}$ values have adjusted for the correlation between paired measurements. These procedures are standard (see Ferguson, 1981, Chapter 19), so it remains to assure that they have been applied. ES's may then be computed by techniques earlier described for $\underline{t}$ tests and the ANOVA. However, three notes of caution should be observed:

- a. Repeated measurements designs will usually depart the finite fixed model assumed for the factorial ANOVA, in favor of random/mixed models. The standardizing statistic S^P must then be derived from variance estimates appropriate for testing row, column, and interaction effects.
- b. In general, ES's acquired from a repeated measurements design will be assigned to different analysis groups, depending on the type of treatment and the type of outcome construct related to each ES. On occasion, the coding process may yield two or more ES's for the same type of treatment and the same construct. Such ES's are not independent, so their dependency must be compensated by methods to be described under "Data Analysis."
- c. Experiments with repeated measures are disadvantaged by the likelihood of carry-over effects upon subjects, e.g., boredom, fatigue, and practice (Ferguson, 1981, p. 318). Such

effects, if present, stand to bias ES values. As a result, each ES derived from repeated measurements designs should be marked as quasi-independent, for later assessment of unexplained variance in data analysis.

3. Lack of a control group. The previous descriptions of techniques for ES calculation presume the existence of a control or "traditional" group for each study. However, the body of studies may contain reports of experiments where treatments have been compared with each other and no control group was involved. An effect size may be computed which relates the means of any two such treatments A and B from:

$$ES_{AB} = \frac{\bar{Y}_A - \bar{Y}_B}{S^P},$$

where S^P is the pooled standard deviation for a t test or the within-group/within-cell standard deviation S^W for an F test. Nonetheless, resulting ES's cannot be used directly, since this model requires a referencing of each treatment mean to a control group mean. A process is needed wherein each ES_{AB} can be split as two effect sizes ES_A and ES_B which reference each treatment to a hypothetical control group. Glass has described such a process, as follows (Glass et al., 1981, p. 124):

- a. Calculate an effect size ES_{AB} relating any two treatments in the study.

- b. From all other studies which have compared Treatment A with a control group, calculate the average, $\bar{\Delta}_A$, of their ES's. Similarly, calculate $\bar{\Delta}_B$.
- c. Break ES_{AB} into two effects ES_A and ES_B which estimate the effect sizes for Treatment A compared with "control" and Treatment B compared with "control." Two reasonable conditions may be imposed on ES_A and ES_B :

$$ES_{AB} = ES_A - ES_B ;$$

$$ES_A - \bar{\Delta}_A = \bar{\Delta}_B - ES_B .$$

Solving these linear equations in two unknowns yields:

$$ES_A = \frac{ES_{AB} + \bar{\Delta}_A + \bar{\Delta}_B}{2} ;$$

$$ES_B = ES_A - ES_{AB} .$$

In studies with multiple treatments, this process may be repeated until each treatment has been compared with the hypothetical control. Resulting ES estimates then may be entered for data analysis.

Representing Dependent Variables--Correlations

A synthesis of correlational studies involves the integration of coefficients descriptive of the relationship between two variables, e.g., achievement and socioeconomic status, or teacher effectiveness and student learning. Since most correlational studies report coefficients routinely, their synthesis delivers less hardship than with experiments.

In correlational meta-analysis, the basic metric (dependent variable) is Pearson's r . However, within the body of studies, a host of other statistics may be encountered: biserial, point biserial, phi, contingency, and rank-order coefficients, chi square, t , F , and more. There usually exists an algebraic path from most of these statistics to Pearson's r , so formulas may be used for these conversions, as given by Glass, McGaw, and Smith (1981, pp. 149, 150).

Organizing the Coding

The representation of variables ends the phase of data collection for meta-analysis. Accumulating data from tens or hundreds of studies may create majestic confusion unless the reviewer has adopted a systematic scheme for indexing and retrieving information.

To ease the coding burden, Figure 3 presents a format which can be used to identify and provide for data recording. Because the data analysis will require the survey of various subsets, a separate coding sheet (Figure 3) should typically be used for each effect size or Pearson's r . Two subsets are stressed: (1) the types of outcome construct (e.g., achievement, attitude, and others); and, (2) the types of treatments surveyed in the studies. These variables are prominently displayed at the top right of the coding sheet. Subsequent major divisions permit the entry of remaining categorical and continuous independent variables. Those considered essential have been specified in Figure 3, and space is provided for entering others whenever they are identified. The rating of each study's design

CONSTRUCT

Author(s)

TREATMENT

CATEGORICAL INDEPENDENT VARIABLES

[illegible]

CONTINUOUS INDEPENDENT VARIABLES

[illegible]

RESEARCH DESIGN RATING: X =

1. Problem definition =	
2. Population description =	
3. Sampling procedures =	
4. Error control =	
5. Instruments =	
6. Data analysis =	
7. Conclusion =	
8. Report =	

ES or r DETERMINATION

[illegible]

FOR USE IN ANALYSIS

r_{yy}	c_m		g or z

***Dependent ES's**

Check here for quasi-independence ☐

Figure 3. Individual coding sheet.

receives special status. The sheet allots space for resolving that rating.

ES calculations may require all or part of the following data: $\bar{Y}^E$, $\bar{Y}^C$, $(S^E)^2$, $(S^C)^2$, S^P , t or F , n^E , n^C , and instrument reliability. Correlations will require reliabilities for both instruments. Where instrument reliability is omitted, its value should be taken as "1." During the phase of data collection, ES is computed or Pearson's r is entered as a raw value. These statistics will be modified for data analysis, and space on the coding sheet is made for entering later results. It was previously mentioned that a separate coding sheet should be used for each effect size. The only exception occurs when two or more ES's result for the same type of construct and treatment in a repeated measurements design. These are dependent ES's. Their raw values should be developed on the same coding sheet where space is provided, for later combination to a single value. Whenever any ES has derived from a repeated measurements design, its value should be flagged as quasi-independent, possibly contaminated by carry-over effects.

Coding the data for meta-analysis may involve a "measurement" error when characteristics and findings are extracted from written reports (Glass et al., 1981, p. 75). The principal source of such error is a variability in coder judgments across studies. To guard against such bias, the rule should be observed that no subjective judgment will be rendered in the absence of a priori guidelines. Where two or more coders are used, the coding process should achieve

acceptable levels of intercoder agreement, and all intercoder reliabilities should be reported.

Paradigm for Data Collection

The opening phase of meta-analysis gathers the data for subsequent synthesis, interpretation, and reporting of results. For either experimental or correlational studies, procedures for data collection flow as shown in Figure 4.

Initial efforts are preparatory. Once the topic area has been defined, the reviewer frames preliminary questions which meta-analysis may be expected to answer. Question formation implies some insight into types of treatments and constructs which have been surveyed in past research, and so these variables, along with other characteristics of studies, are used to construct a coding sheet for data identification and recording. In parallel efforts, through computer searches, abstracting services, and surveys of previous reviews and bibliographies, the reviewer identifies reports potentially useful for the research integration. To minimize the likelihood of overlooking studies, a thorough scouring of resources is essential.

Actual data recording proceeds from a direct review of each report. Some studies may be excluded through a priori guidelines. From each of the studies accepted, the reviewer extracts information on the independent/dependent variables, assesses research design quality, and determines effect magnitudes or values of Pearson's r . Each study is also surveyed for additional research questions, study

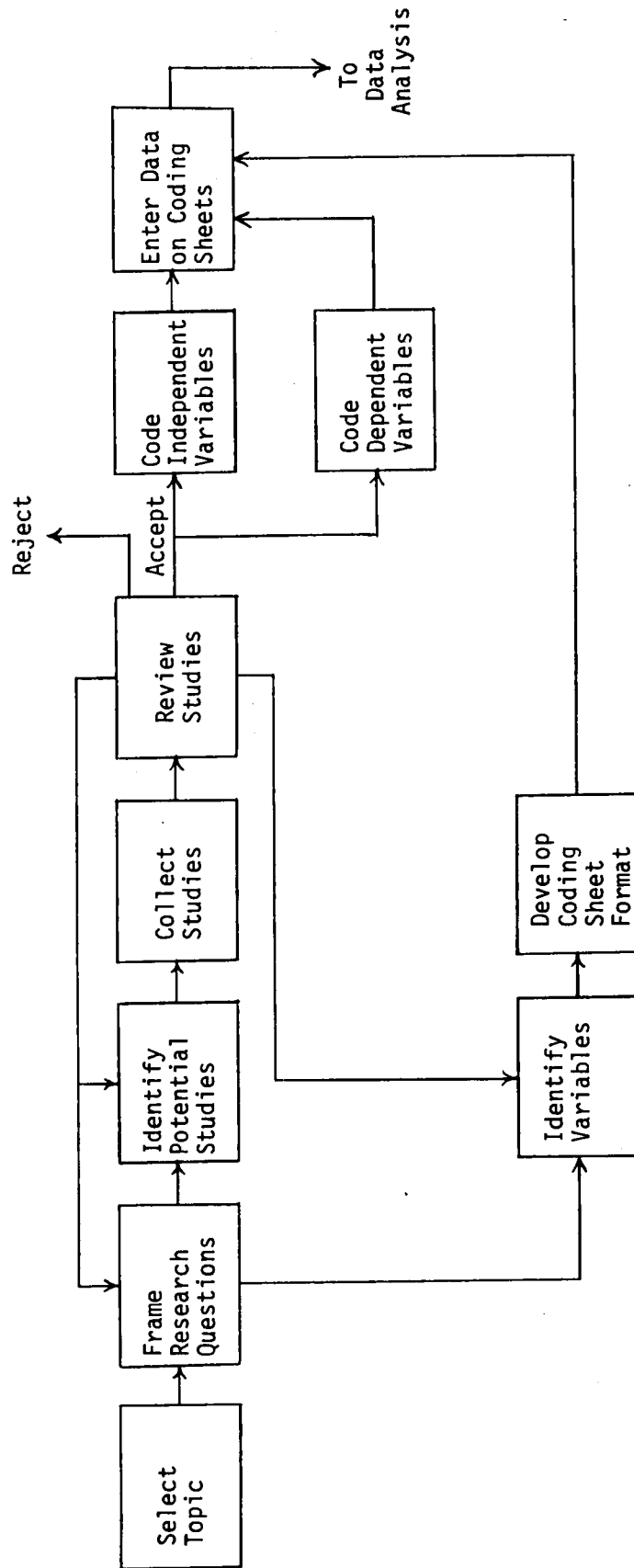


Figure 4. Flow of procedures in data collection.

characteristics, treatments, and outcomes, and for references to studies not already listed. As shown by feedback loops in Figure 4, question formation, identification of variables, and the search for studies are on-going efforts which converge on finalization as the coding process nears conclusion.

A coding sheet for each independent effect size, set of dependent ES's, or Pearson's r becomes the end product of data collection.

Data Analysis--Experiments

The central purpose of data analysis is the estimation of means and confidence intervals for selected groups of effect magnitudes. Each mean is then presumed to express the overarching effect for the construct involved. A synchronal effort assesses relations between study properties and effect size.

Past applications of meta-analysis usually have pooled all study ES's to derive a grand effect size which reflects all features of studies, treatments, and outcomes. Secondary efforts then have calculated ES means for subgroups of interest. These past analyses have in general disregarded fundamental statistical precepts (Hedges, 1982a). Reviewers typically have failed to distinguish sample statistics from population parameters, seeming content to average ES's and declare the sample mean a population effect magnitude. Such an approach may be misleading unless the underlying populations have identical effects across all studies. This may or may not be the case; sometimes the effects are indeed homogeneous, but other times,

heterogeneous (Hedges, 1982a). Moreover, past reviews have treated ES means as purely descriptive, declining to question whether those means departed significantly from zero.

Thus, there appears a valid concern that research integration by meta-analysis has lacked statistical underpinning. The present model addresses that concern. The purposes of data analysis as defined by Glass (1976) are retained, but statistics in the model will be based on sampling theory and the inferential methods developed by Hedges (1981, 1982a-d).

Definitions of Terms

Since the data analysis makes extensive use of estimated values, the following concepts are revisited (Ferguson, 1981, pp. 158-160):

1. Biased estimator--A sample statistic whose expected value systematically departs from the population value.
2. Unbiased estimator--A sample statistic whose expected value approaches the population value as the number of samples increases.
3. Consistent estimator--A sample statistic which approaches the population value as the sample size increases.
4. Relative efficiency of estimators--The estimator with the smallest sampling variance is the more efficient.
5. Sufficiency of an estimator--The condition attained when one estimator is more efficient than any other. A sufficient method of estimation uses all the information in the sample.

Assumptions

Initial assumptions were previously drawn as a general aid to data collection. The present set is more specific, directed toward the statistical treatment of effect size information.

1. It is assumed that all source data for effect size calculations have met the requirements for a valid t test.

Effect size is defined to be the standardized mean difference $(\bar{Y}^E - \bar{Y}^C)/S^P$. The significance of a difference between experimental means most likely will have been assessed by t or F test, under the conditions of normally distributed group observations and a common variance for both groups. These assumptions carry over to ES calculations. If the conditions are satisfied, exact statistical properties of effect size estimators may be derived (Hedges, 1982a). In practice, the assumptions seem reasonably well met, and the robustness of the t test suggests that ES calculations will be similarly robust.

2. It is assumed that instruments which measure the same construct are linearly equatable.

Measurements from study to study will probably have been derived from different measuring instruments, even where the same outcome construct has been surveyed. If these different instruments yield highly correlated test scores, they are viewed as linearly equatable, which implies that their response scales can be made equivalent under linear transformations. Standardized mean differences then are scale-invariant measures of treatment effects, that is, the effect magnitudes are independent of the different measuring

instruments. Under this condition, effect magnitudes may be legitimately compared and combined with other ES's.

Where instruments have measured the same construct, the assumption of linear equatability seems justified, especially when standardized instruments have been employed. (However, the model arranges for a test of the assumption during the process of data analysis.) There seems no reason to believe that effect magnitudes for one construct (say, achievement) are comparable to ES's for another construct (such as attitude). During the data analysis, different constructs should be analyzed separately.

3. It is assumed that the use of large-sample approximations is valid.

The application of large-sample approximations permits a reduction of nonstandard distributions to the normal or chi square. Hedges (1982a) has showed that such approximations yield excellent results when sample sizes are equal to or greater than 10 and effect magnitudes are smaller than 1.5. The initial assumptions constrain sample sizes to 10 or greater. Effect sizes less than 1.5 seem characteristic of educational research (Gage, 1978).

4. It is assumed that the body of studies constitutes a probabilistic sample.

Inferential statistics rely on randomized experiments or surveys with explicit sampling procedures. Meta-analysis seldom meets either condition (Glass et al., 1981, p. 199). The validity of inferential procedures must then depend on how well the available studies represent a theoretical population.

If the reviewer were able to find all existing studies that bear on a topic, the resulting body could perhaps be viewed as representative, all there is. However, total coverage seems too much to be expected, in particular with regard to unpublished research. Reports appearing in print seem likely to present statistically significant findings, while results in unpublished studies are most likely nonsignificant; where such fugitive research is not included in the body for meta-analysis, a bias must accrue from its omission (Glass et al., 1981, p. 64). Rosenthal (1980) has studied the extent of damage caused by such bias (the "file drawer" problem). If an observed ES mean may be drawn near nonsignificance by the addition of only a few null results, that mean is not resistant to the file drawer problem. However, the presence of a large group of studies and/or a large ES mean usually protects against the file drawer threat as a rival hypothesis. Hence, the best assurance for a probabilistic sample is a careful tracking down of many studies.

Nonetheless, the assumption of a probabilistic sample, while essential for the use of inferential statistics, seems less than powerful. Inferential conclusions should be drawn with appropriate caution.

Preparing for Data Analysis

The initial effort of data analysis is correction of the raw ES's afforded by data collection to more precise values. In particular, compensation is provided for distribution bias, errors of measurement, and data dependence.

Correction for distribution bias. Hedges (1981) has showed that sample ES's are biased estimators of underlying population effects. Suppose there exist k effect magnitudes, each of which compares (E)xperimental and (C)ontrol group observations which are normally distributed with population means μ_i^E and μ_i^C and common variance σ_i^2 , $i = 1, \dots, k$. In this notation, the i th population effect size is defined as:

$$\delta_i = \frac{\mu_i^E - \mu_i^C}{\sigma_i} \quad . \quad (DA-1)$$

To estimate δ_i , sample analogues are substituted. Hence:

$$ES_i = \frac{\bar{Y}_i^E - \bar{Y}_i^C}{S_i^P} \quad , \quad (DA-2)$$

this being the equation used for initial effect size calculations. Since k population values exist, it is assumed that measures across studies are linearly equatable, such that δ_i has the same interpretation across studies. Its estimator is ES_i , a sample statistic which therefore possesses a sampling distribution. Where $\tilde{n}_i = n_i^E n_i^C / (n_i^E + n_i^C)$, ES_i is distributed as $(1/\sqrt{\tilde{n}_i})$ times a noncentral t random variable with non-centrality parameter $\delta_i \sqrt{\tilde{n}_i}$ and degrees of freedom $m_i = n_i^E + n_i^C - 2$ (Hedges, 1981). With this distribution, ES_i becomes an inference-sufficient yet a biased estimator of δ_i . Hedges offers the correction factor $c(m_i)$ which yields an (U)nbiased estimator according to:

$$ES_i^U = c(m_i)ES_i = \frac{c(m_i)(\bar{Y}_i^E - \bar{Y}_i^C)}{S_i^P}, \quad (DA-3)$$

where exact values of $c(m_i)$ are given in Hedges' (1981) table. However, an excellent approximation may be derived from:

$$c(m_i) \approx 1 - \frac{3}{4m_i - 1}, \quad (DA-4)$$

with maximum error of .0005 when $m_i \geq 10$. Figure 5 depicts the ratio ES/ES^U as a function of degrees of freedom. As m grows large, the ratio tends to a value of 1, so ES is essentially unbiased in large samples.

Correction for errors of measurement. A use of equation DA-3 corrects ES_i for distribution bias in the absence of measurement error. If the response measure is not perfectly reliable, errors of measurement cause ES_i^U to underestimate δ_i as a function of instrument reliability r_i^{YY} . An error-free estimator, symbolized as g_i , is derived from:

$$g_i = \frac{c(m_i)ES_i}{\sqrt{r_i^{YY}}} = \frac{c(m_i)(\bar{Y}_i^E - \bar{Y}_i^C)}{S_i^P \sqrt{r_i^{YY}}}. \quad (DA-5)$$

Where measurement errors are negligible (or unknown), r_i^{YY} may be taken conservatively as "1," and DA-5 reduces to DA-3. All raw ES values derived during data collection should be corrected accordingly and recorded on their coding sheets.

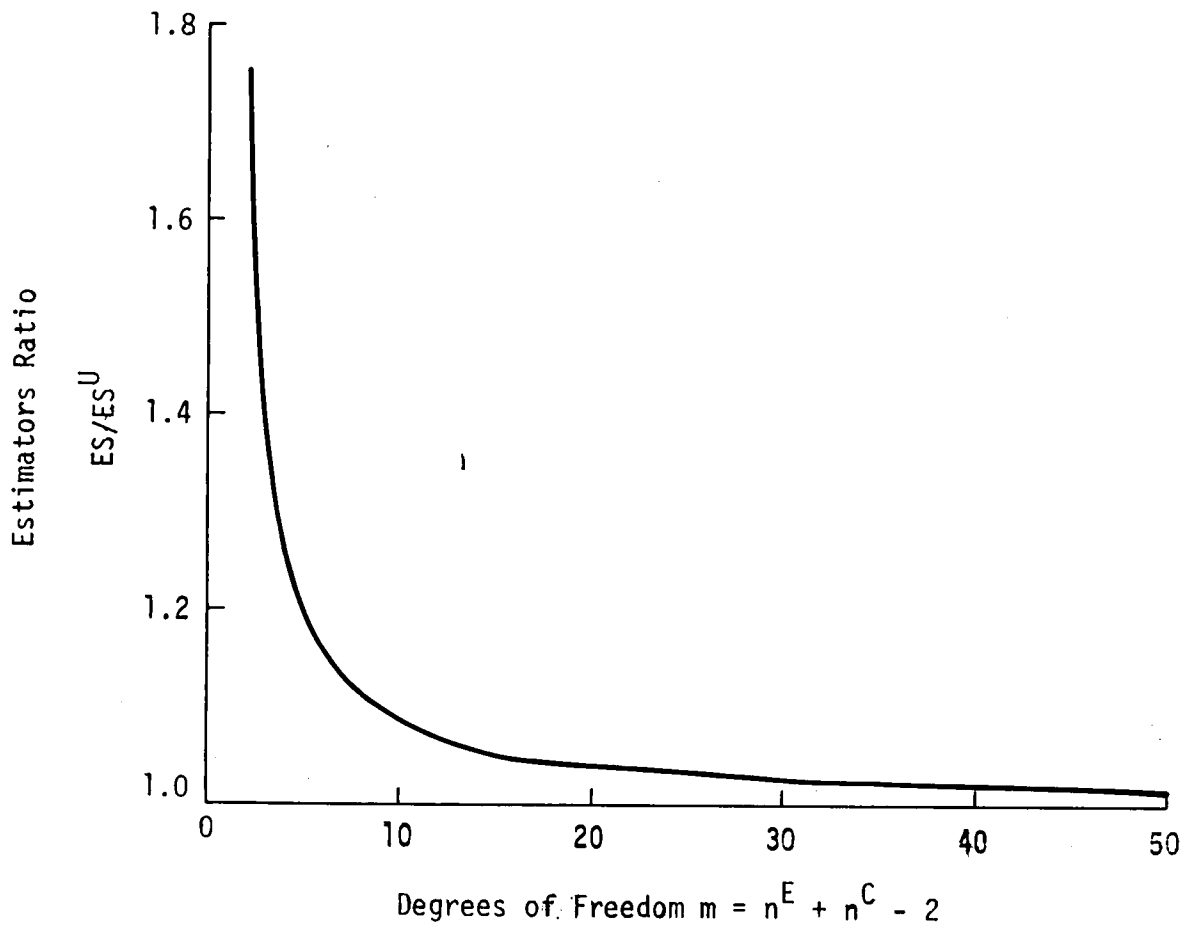


Figure 5. Ratio of biased/unbiased estimators as a function of degrees of freedom.

Compensation for data dependence. The foregoing efforts provide a corrected effect size g_i for each raw ES_i , including those in groups recorded as dependent. Equations used in the data analysis assume that all effect magnitudes are statistically independent, so compensatory treatment for clusters of dependent g 's must be considered. A simple but risky "treatment" is an ignoring of data dependencies. Each g_i would then be used on the basis of independence. Such an assumption, though not the case, is practical, especially if the number of dependencies is small compared to the overall body of g_i 's. Inferential calculations would most likely be conservative, since dependencies tend to increase standard errors of estimate (Glass et al., 1981, p. 200). However, if dependencies in number seem proportionately large, each cluster should be combined to yield an equivalent effect magnitude, perhaps by factor analysis. Giaconia and Hedges (1982) have used a simple average g to represent a dependent cluster, retaining the common sample sizes n_i^E and n_i^C . There is then no violation of the independence assumption, but a different problem perhaps will arise. In some "rare" cases, there may occur a strong moderator variable with effects within as well as across all studies (Hunter et al., 1982, p. 119). In such a case, the average g would be conceptually ambiguous.

Thus, the effects of data dependence may be ignored, perhaps reduced, but not clearly erased. The approach of averaging each cluster of dependent g 's seems preferable, being simple and consistent with the independence assumption. Each set of dependent g 's should be combined accordingly and recorded on its coding sheet.

Simple Description of Effect Magnitudes

After study outcomes have been transformed to error-free independent effect magnitudes, any technique for describing statistics may be applied, e.g., frequency histograms, averages, and measures of variability (Glass et al., 1981, p. 154). To provide a general pictorial of effect size distributions, an informative technique is the schematic "box-and-whiskers" plot, for which an example appears in Figure 6. Effect size distributions for the measured constructs are described by columns of dots or stars. The top and bottom of each central box are hinges located roughly at the first and third quartiles of each set of data. Dotted lines (whiskers) guide the way to inner fences marked "f," with the length of each whisker one-and-a-half times the height of its box. This distance is doubled to locate outer fences marked "F." Data points outside an inner fence are considered outliers, and points beyond an outer fence become far outliers. Any outlier is viewed with suspicion; far outliers are viewed with incredulity. Outliers probably represent statistical artifacts, e.g., typographical, computational, or transcription errors. Thus, the plot identifies points which may become sources of unexplained variance during later data analysis.

Testing the Homogeneity of Effect Size Groups

A primary thrust of data analysis is the calculation of effect size means. The mean of any effect size group then represents the overall effect for characteristics and outcomes of studies involved. Within the present model, the effect size collection is to be partitioned for

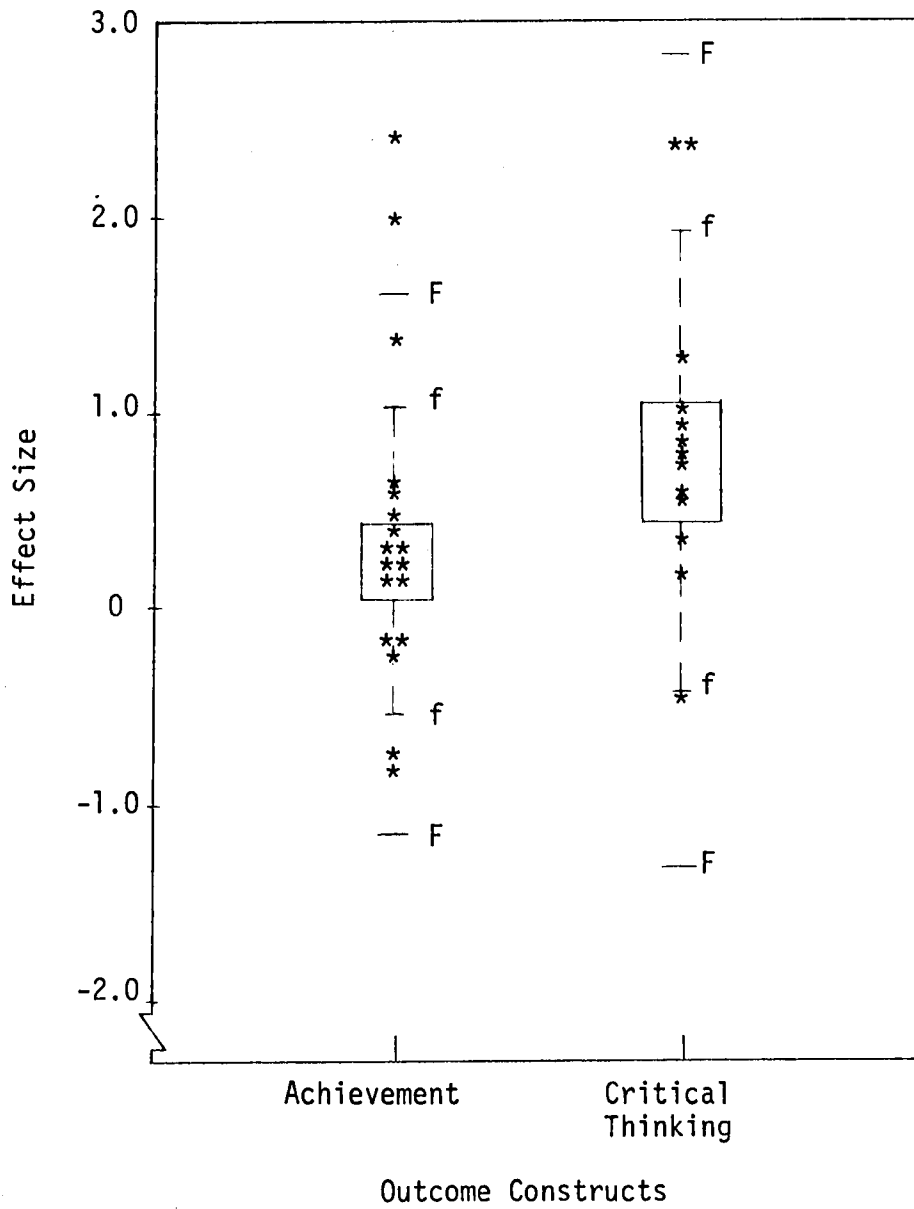


Figure 6. Example box-and-whiskers plot for effect size and various outcome constructs.

separate analysis according to construct (say, achievement, critical thinking skills, attitudes, and process skills), and so a mean computed for each subset then would express the overall structural effect.

However, for any such mean to be valid, each effect size in the subset must share the population effect size δ estimated by the mean. Each g_i then estimates a δ_i , and all δ_i 's are identical, i.e., $\delta_i = \delta$. Under this condition, the effect size group is said to be homogeneous, its mean becomes a proper estimator of the population δ , and its variance is purely a result of sampling error. Where the variance of an effect size group is larger than should be expected from chance, the group is said to be heterogeneous, and its mean no longer represents one underlying population effect. Outliers are a possible cause of variability beyond sampling variance; their deletion may produce a homogeneous effect size group. The other likely cause of heterogeneity is a study characteristic that moderates (is correlated with) effect size. Where such relationships exist, the underlying population effects are also variable, and two or more means are thus embedded in the effect size collection.

Therefore, with regard to a selected effect size group, the initial effort of synthesis is to determine whether the group is homogeneous; do all the g 's share a common effect size δ ? This section provides a statistical test which assesses observed variability against the sampling variance for k effects (Hedges, 1982a).

1. Establish the hypothesis $H_0: \delta_i = \delta, i = 1, \dots, k$, versus the alternative that at least one δ_i is different from the others.

2. Calculate the test statistic H_T . Under a large-sample approximation, $g_i \sim N[\delta_i, \sigma_i^2(\delta_i)]$, where the population variance $\sigma_i^2(\delta_i)$ is estimated by:

$$\sigma_i^2(g_i) = \frac{n_i^E + n_i^C}{n_i^E n_i^C} + \frac{g_i^2}{2(n_i^E + n_i^C)} \quad (\text{DA-6})$$

H_T is then the sum of squares of each g_i about a mean effect size weighted for sample size differences, with the i th such square being weighted by the reciprocal of the variance DA-6. The computational formula is:

$$H_T = \sum_{i=1}^k \frac{g_i^2}{\sigma_i^2(g_i)} - \frac{\left[\sum_{i=1}^k \frac{g_i}{\sigma_i^2(g_i)} \right]^2}{\sum_{i=1}^k \frac{1}{\sigma_i^2(g_i)}} \quad (\text{DA-7})$$

H_T is distributed as chi square with $(k - 1)$ degrees of freedom:

$$H_T \sim \chi^2_{k-1}.$$

3. Do not reject H_0 if H_T is smaller than the critical value of χ^2_{k-1} at a desired significance level. The group of k effects is then judged homogeneous and independent of properties which vary across studies. Thus, the group mean represents one underlying population effect size δ . Calculation of this mean is discussed in the subsequent section.

4. Reject H_0 if H_T exceeds the χ^2_{k-1} critical value. The group of k effects is then judged heterogeneous, which indicates the

presence of variance beyond that explained by sampling error. It ought to be noted that even small variations in values of g may lead to a large test statistic H_T , if sample sizes are consistently "very large" (Hedges, 1982d, p. 21). Where the range of g values seems narrow as well, the reviewer may choose to treat the group as homogeneous, even though population effects appear to differ. A more conservative approach entails a search for causes of the heterogeneity. Procedures for such searches will be described in subsequent sections.

Calculating Means and Confidence Intervals

Where a set of effect magnitudes is found to be homogeneous, it is natural to estimate their common population effect size δ by pooling g_i 's. Since estimates from studies with large sample sizes will be more precise than estimates from smaller samples, more weight should be accorded the more precise estimates. Hedges (1982a) has offered such a weighted mean, defined by:

$$\bar{g}^w = \frac{\sum_{i=1}^k \frac{g_i}{\sigma_i^2(g_i)}}{\sum_{i=1}^k \frac{1}{\sigma_i^2(g_i)}} \quad . \quad (\text{DA-8})$$

The numerator and denominator appear as terms in DA-7, and so the test for homogeneity can also produce $\bar{g}^w$. This mean is a consistent estimator of the population effect size δ .

The normal asymptotic distribution of $\bar{g}^w$ may be used to obtain large-sample confidence intervals for δ (Hedges, 1982a). The

approximate distribution is: $\bar{g}^w \sim N[\delta, \sigma^2(\delta)]$, where the population variance $\sigma^2(\delta)$ is estimated by:

$$\sigma^2(\bar{g}^w) = \frac{1}{\sum_{i=1}^k \frac{1}{\sigma_i^2(g_i)}} \quad . \quad (\text{DA-9})$$

This denominator is a term of DA-7, and so the test for homogeneity can also produce $\sigma^2(\bar{g}^w)$. A $100(1 - \alpha)$ percent confidence interval for δ at the significance level α is derived from:

$$\bar{g}^w - z_{\alpha/2} \sigma(\bar{g}^w) \leq \delta \leq \bar{g}^w + z_{\alpha/2} \sigma(\bar{g}^w), \quad (\text{DA-10})$$

with $z_{\alpha/2}$ obtained from tables of the normal distribution.

Testing Homogeneity Within Effect Size Classes

Since measuring instruments for different outcome constructs cannot be assumed to be linearly equatable, a separate analysis for each construct will be required. Under ideal conditions, a construct's entire set of effect magnitudes will be found homogeneous, revealing the existence of a common population effect across broad ranges of study characteristics. However, a construct's collection will most likely be heterogeneous. In studying the effects of open education, Giaconia and Hedges (1982) found structural heterogeneity for each of their 19 dependent variables.

Where a selected set of g 's seems heterogeneous, searches for homogeneous subsets within the overall collection may be rewarding. These subsets or classes are formed from characteristics of studies

established as categorical independent variables during the process of data collection. For instance, the construct of achievement perhaps has been measured in a number of instructional treatments. The subset of g 's for each treatment thus may be grouped as a class and tested for homogeneity independently. For within-class testing of homogeneity, the following procedures may be employed.

1. Partition the effect size collection into p classes according to the categorical dimension considered most likely to influence the effect magnitudes. For each subgrouping, calculate the within-class test statistic H_{Wj} , $j = 1, \dots, p$, using equations DA-6 and DA-7. Also calculate the means $\bar{g}_j^W$ by DA-8 and standard deviations $\sigma_j(\bar{g}^W)$ by DA-9.

2. Assess each class's homogeneity according to the guidelines described for the general test.

3. If all the classes are homogeneous, each class mean $\bar{g}_j^W$ is a consistent estimator of the population effect size δ_j underlying the subset. Confidence intervals may then be computed by equation DA-10. The cause of heterogeneity for the overall collection was undoubtedly a difference between class means, indicating that effect size is dependent on the study characteristic used as a basis for classification.

4. If most of the classes remain heterogeneous, a division into smaller classes or by way of a different classification may be employed, e.g., by levels of research design quality, type of publication, or type of measuring instrument. Procedures for within-class testing would then be repeated for each new or additional classification.

5. If some of the classes are homogeneous and some still heterogeneous, the reviewer may choose to segregate the homogeneous classes and record their means, confidence intervals, and observed relationships with study characteristics. Further subdivision and testing of the heterogeneous classes would then proceed as described in part 4.

6. The process of classifying and testing should continue until acceptable levels of homogeneity are obtained or until no further classification can be considered. Under the latter condition, efforts must then focus on other techniques for finding a cause of heterogeneity in each offending subset.

Analyzing Heterogeneous Effect Size Groups

A test of homogeneity inquires whether the observed variability of an effect size group is greater than would be expected from chance. Rejection of the null hypothesis implies the existence of effect size variance unexplained by sampling error, and the reviewer is then obliged to seek its cause or causes. Where a large number of effect magnitudes is involved, the heterogeneous group may be reduced to homogeneous classes, as the previous section describes. However, this method will not reveal two other likely sources of unexplained variance:

1. Outlier data points;
2. Significant relationships between continuous independent variables and effect magnitudes.

These potential sources will now be considered in turn.

Data points which seem substantially dispersed from the bulk of effect magnitudes in a group are called outliers. These may spring from a variety of causes, e.g., heterogeneous variances, carry-over effects in repeated measurements designs, or typographical, computational, and transcription errors. Outliers usually may be found by inspection of the group's effect size distribution, say, within a box-and-whiskers plot. If present, such outliers should be deleted and the parent group retested for homogeneity. If H_0 is not rejected, outliers probably caused the unexplained variance, so the effect size mean and confidence interval (excluding outliers) may be recorded. If H_0 is rejected, effect magnitudes are probably related to one or more study characteristics, and a search for moderating variables is in order.

Glass (1978) has recommended the general strategy of coding study characteristics as a vector of predictor variables and using multiple regression to determine correlations between the characteristics and effect magnitude. This method presumes the existence of constant sample sizes across studies. Hence, its results become suspect when synthesized studies exhibit different sample sizes (Hedges, 1982d, p. 41). Alternative methods have been developed for fitting a regression model to effect size data (Hedges, 1982c). These methods provide consistent, asymptotically efficient estimates of regression coefficients, which may then be tested for significant departures from zero.

1. Assume the standardized mean difference g_i , $i = 1, \dots, k$, depends on p independent variables, with $k \geq p$. Define the following matrix for values of the independent variables:

$$X = \begin{bmatrix} x_{11} & \cdot & \cdot & \cdot & x_{1p} \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ x_{k1} & \cdot & \cdot & \cdot & x_{kp} \end{bmatrix}$$

Denote the vector of effect size estimates by $\underline{g} = (g_1, \dots, g_k)'$, and define the matrix $V = \text{diag}[\sigma_1^2(g_1), \dots, \sigma_k^2(g_k)]$, where the variances are given by DA-6. The vector of regression coefficients $\underline{\beta} = (\beta_1, \dots, \beta_p)'$ is estimated by:

$$\hat{\underline{\beta}} = (b_1, \dots, b_p)' = (X'V^{-1}X)^{-1}X'V^{-1}\underline{g} \quad . \quad (\text{DA-11})$$

2. The hypothesis that all elements of $\underline{\beta}$ are simultaneously zero is tested by the statistic:

$$H_1 = \hat{\underline{\beta}}'X'V^{-1}\underline{g} \quad . \quad (\text{DA-12})$$

H_1 is distributed as χ_p^2 . If H_1 is less than the critical value at a significance level α , the hypothesis that $\underline{\beta} = 0$ is supported, and none of the independent variables seems related to effect size. Such a finding will appear contradictory, since effect size data for the group is heterogeneous. Probably the model has examined the wrong variables.

3. If the test that $\underline{\beta} = 0$ is rejected, one or more independent variables are significantly related to effect size. A confidence interval for elements β_j of $\underline{\beta}$ will reveal nonzero coefficients and identify significant independent variables. Define the diagonal matrix:

$$V_j = (X'V^{-1}X)^{-1} \quad . \quad (DA-13)$$

With $j = 1, \dots, p$, a $100(1 - \alpha)$ percent confidence interval for β_j is given by:

$$b_j - z_{\alpha/2} \sqrt{v_j} \leq \beta_j \leq b_j + z_{\alpha/2} \sqrt{v_j} \quad , \quad (DA-14)$$

where v_j is the j th diagonal element of V_j and $z_{\alpha/2}$ is obtained from tables of the normal distribution.

It needs to be noted that all attempts to find a cause of heterogeneity may end in frustration. The cause may be a variable not included for data analysis. Missing data may strike with a vengeance; where gaps exist, the reviewer must make do with what he has, rendering the assessment of some variables impossible at worst and suspect at best. If the heterogeneity of a group remains unexplained, its mean becomes a descriptive statistic from which no inferences can be drawn.

Comparing Means of Effect Size Groups

Occasions may arise when the reviewer wishes to assess the differences among means of various effect size groups, e.g., as related to different instructional treatments or to different school

grade levels. In particular, when partitioning a heterogeneous group produces homogeneous classes, significant differences will exist among class means, and yet which pair or pairs of means have unequal values?

The situation is similar to the observation of a significant F test in analysis of variance with more than two treatments. Hence, for comparing k effect size means, the corresponding subsets may be processed by analysis of variance to yield the within-group variance estimate $(S^W)^2$. This statistic may then be used in multiple comparison procedures, say, by the method of Duncan or Scheffé.

This approach compares the ordinary means $\bar{g}_i$, $i = 1, \dots, k$, and seems appropriate for roughly assessing mean differences. However, the weighted means $\bar{g}_i^W$ are more precise estimates of population values and should be used in describing results.

Data Analysis--Correlations

Procedures for analyzing correlation coefficients are entirely analogous to techniques presented for effect size integration. Efforts involve an adjustment of r values to Fisher's z , tests of homogeneity, and searches for causes of heterogeneity. Statistical procedures for combining correlations have been developed by Hedges and Olkin (1983). Again, the central thrust is estimating means, each of which is then presumed to express the overall correlation for studies involved.

Suppose k correlations exist, $r_1, \dots, r_k$, with sample sizes $n_1, \dots, n_k$. Analysis is simplified if each correlation r_i is transformed to Fisher's z by:

$$z_i = \frac{1}{2} \log \left[\frac{1 + r_i}{1 - r_i} \right] . \quad (\text{DA-15})$$

For any group of correlations, the initial task is to determine whether all the z_i 's could reasonably have come from populations which share a common correlation. A test of homogeneity may employ the test statistic:

$$H = \sum_{i=1}^k (n_i - 3) z_i^2 - \frac{\left[\sum_{i=1}^k (n_i - 3) z_i \right]^2}{\sum_{i=1}^k (n_i - 3)} , \quad (\text{DA-16})$$

which is assessed by comparison with the critical value of χ^2_{k-1} at a desired significance level. If H is less than the critical value, the hypothesis of homogeneity is upheld.

For homogeneous correlations, a natural estimate of the common z -transform ζ is the weighted mean:

$$\bar{z}^w = \frac{\sum_{i=1}^k (n_i - 3) z_i}{\sum_{i=1}^k (n_i - 3)} , \quad (\text{DA-17})$$

whose associated variance is given by:

$$\sigma^2(\bar{z}^w) = \frac{1}{\sum_{i=1}^k (n_i - 3)} \quad . \quad (\text{DA-18})$$

The standard deviation implied by DA-18 is used to establish a confidence interval for ζ as follows:

$$\zeta_1 = \bar{z}^w - z_{\alpha/2} \sigma(\bar{z}^w) \leq \zeta \leq \bar{z}^w + z_{\alpha/2} \sigma(\bar{z}^w) = \zeta_2 \quad . \quad (\text{DA-19})$$

Here, $z_{\alpha/2}$ is the 100(1 - α) percent critical value of the standard normal table. Conversions from z to r may be derived from:

$$\bar{r}^w = \hat{\rho} = \tanh(\bar{z}^w) = \frac{e^{2\bar{z}^w} - 1}{e^{2\bar{z}^w} + 1} \quad ; \quad (\text{DA-20})$$

$$\rho_1 = \tanh(\zeta_1) \leq \rho \leq \tanh(\zeta_2) = \rho_2 \quad . \quad (\text{DA-21})$$

Where a group of correlations is found heterogeneous, a search for homogeneous classes may proceed in the manner described for effect size partitions. Similarly, outliers may be identified as causes of unexplained variance. If the reviewer then must search for moderating variables, regression should again be used.

Assume k correlations and p independent variables. Define the following matrices:

$$X = \begin{bmatrix} x_{11} & \cdot & \cdot & \cdot & x_{1p} \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ x_{k1} & \cdot & \cdot & \cdot & x_{kp} \end{bmatrix} \quad ;$$

$$\underline{z} = (z_1, \dots, z_k)' \quad ;$$

$$V = \text{diag} [(n_1 - 3), \dots, (n_k - 3)] ;$$

$$\underline{\beta} = (\beta_1, \dots, \beta_p)' .$$

Then the vector of regression coefficients is estimated by:

$$\hat{\underline{\beta}} = (X'VX)^{-1}X'V\underline{z} . \quad (\text{DA-22})$$

A test that $\underline{\beta} = 0$ uses the test statistic:

$$H_1 = \hat{\underline{\beta}}'(X'VX)\hat{\underline{\beta}} , \quad (\text{DA-23})$$

which has a chi square distribution with p degrees of freedom. If H_1 exceeds the critical value at a desired significance level, the hypothesis that $\underline{\beta} = 0$ is rejected, and the reviewer will need to find significant regression coefficients by way of a confidence interval for each β_j in $\underline{\beta}$. A $100(1 - \alpha)$ confidence interval is given by:

$$\hat{\beta}_j - z_{\alpha/2}v_j \leq \beta_j \leq \hat{\beta}_j + z_{\alpha/2}v_j , \quad (\text{DA-24})$$

where v_j is the j th diagonal element of the matrix $(X'VX)^{-1}$, and $z_{\alpha/2}$ is obtained from tables of the normal distribution.

Summary of Calculations

The process of data analysis involves the correction or transformation of raw values, tests of homogeneity, calculation of means and confidence intervals, and regression of outcome data onto study characteristics. A summary of calculations appears in Table 1,

with references to appropriate equations. Except for regression procedures, the calculations may be done manually or by microcomputers. (A program for testing effect size homogeneity and calculating means/confidence intervals on the Intertec SuperBrain microcomputer is presented in Appendix C.) Regression procedures may use any standard computer program which manipulates matrices, e.g., SAS Proc Matrix.

Table 1
Summary of Basic Calculations

Operation	Reference Equations	
	Experiments	Correlations
Adjusting raw values for:		
Distribution bias	DA-3, DA-4	
Errors of measurement	DA-5	
Simplifying analysis		DA-15
Testing homogeneity	DA-6, DA-7	DA-16
Finding means	DA-8	DA-17
Finding confidence intervals	DA-9, DA-10	DA-18, DA-19
Converting z to r		DA-20, DA-21
Finding regression coefficients	DA-11	DA-22
Test of $\beta = 0$	DA-12	DA-23
Finding β_j confidence intervals	DA-13, DA-14	DA-24

To organize the process of analysis, it may be helpful to employ a Master Coding Sheet for data recording. Figure 7 presents a format oriented toward effect size data, but its modification for correlations

may be easily accomplished. For a selected effect size group, the sheet permits the entry of independent variables, sample sizes, effect magnitudes, data from the test for homogeneity, the weighted mean, and its confidence interval.

Paradigm for Data Analysis

Following data collection, efforts turn to data synthesis, where effect magnitudes (or correlations) are combined into means with their confidence intervals or used to define relationships with characteristics of studies. Procedures for data analysis flow as shown in Figure 8.

Initial efforts are preparatory. Raw effect magnitudes are corrected for bias, errors of measurement, and data dependence (or Pearson r -values are transformed to Fisher's z). The total set of data is partitioned according to construct, and a group is then selected for statistical analysis.

Formal synthesis begins with a test of the data's homogeneity at the structural level. (Note: All tests of homogeneity can yield a weighted group mean and the standard deviation used for constructing a confidence interval.) If the null hypothesis is not rejected, the data are deemed estimates of a common population effect and thus are independent of study characteristics. The overall mean and its confidence interval then become the sole products of data analysis, and the reviewer is free to work the next construct. However, data at this level will more likely be heterogeneous, and measures must be taken to remove or explain the cause(s) of heterogeneity.

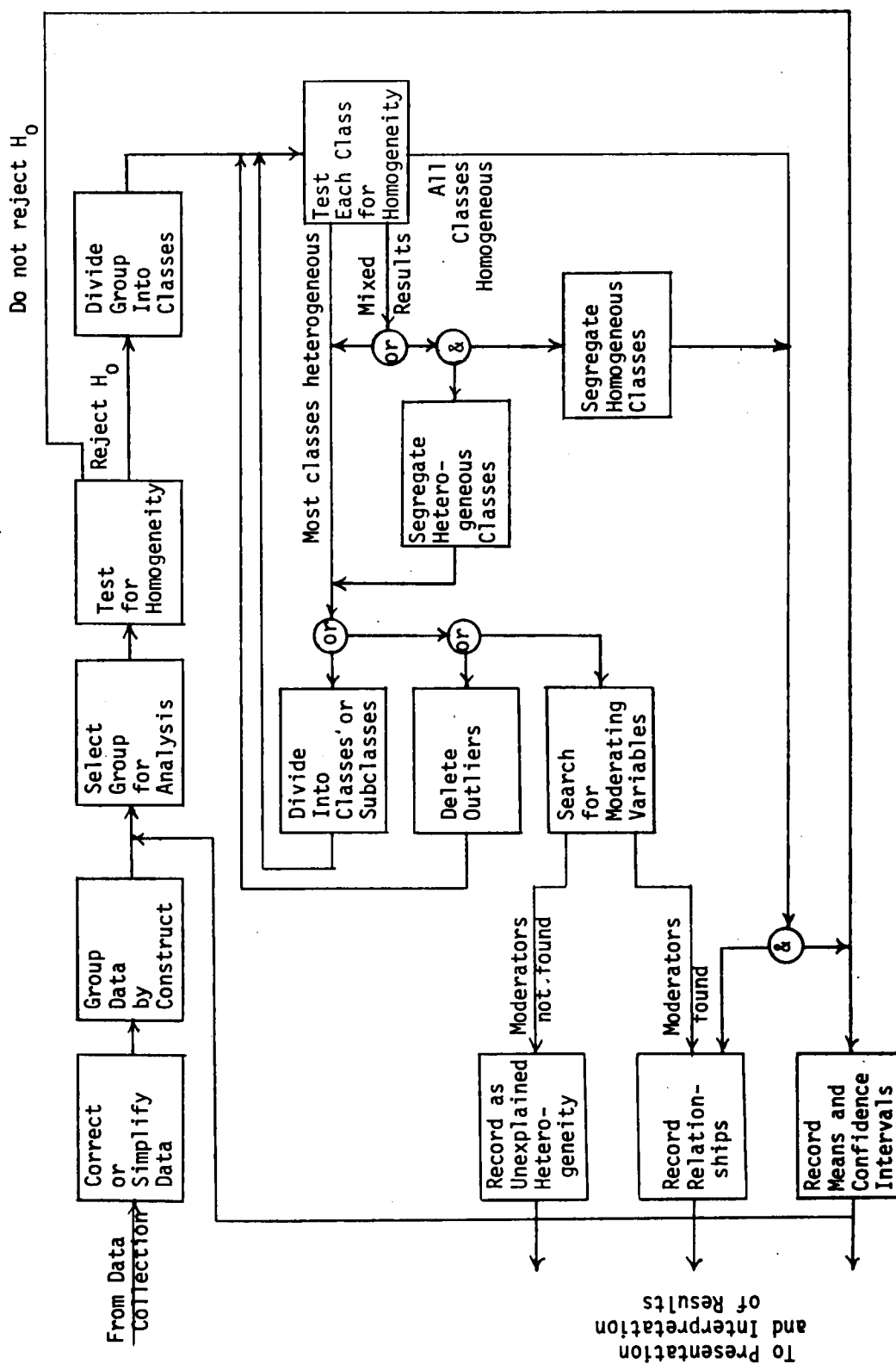


Figure 8. Flow of procedures in data analysis.

A possible cause is the existence of different population effects within the overall collection. The dataset thus is divided into classes (say, according to experimental treatments) for a test of the hypothesis that a different effect size exists for each class. One of three outcomes may ensue:

1. All classes may be homogeneous, indicating that effect size depends on the classification variable. Means, confidence intervals, and observed relationships are recorded as products of data analysis.

2. Most classes may be heterogeneous. The reviewer must then exercise options. The data may be further subdivided to classes within classes or completely repartitioned according to other classification variables. Within-class tests of homogeneity would then be repeated. Other options include: (a) identification of outlier data points, whose deletion requires a retest of classes for homogeneity; and, (b) a search by regression for study characteristics which moderate effect size. Where moderators are found, their relationships with effect size are recorded. If no moderating variable is identified, the cause of heterogeneity remains unexplained, and this result should be noted accordingly.

3. Results may be mixed, with some classes homogeneous and some heterogeneous. Again the reviewer must exercise options. All classes may be treated as heterogeneous; then part 2 would be repeated. Alternatively, the homogeneous groups may be described by means, confidence intervals, and observed relationships with study characteristics. The heterogeneous classes would be treated in the manner of part 2.

The search for causes of heterogeneity continues until the con-structural dataset has been reduced to homogeneous classes, heterogeneous classes where significant relationships exist between study characteristics and effect size, and classes whose heterogeneity cannot be explained.

The overall process then is repeated until all constructs have been analyzed.

Presenting and Interpreting Results

The meta-analysis of research is directed toward a quantitative synthesis of findings and delineation of relationships among those findings and characteristics of studies (Glass, 1978).

Toward these ends, the data analysis yields group means and confidence intervals for various crosscuts of the total dataset. Variables which significantly influence study outcomes are also identified. Results of analysis then should be presented in forms which demonstrate their relevance with regard to the research questions established in data collection. Univariate description (of which the box-and-whiskers plot is an example), tables, percentiles, frequency distributions, ordinary graphs, fitted curves--any technique that summarizes data and provides for data comparison has potential value as a descriptor of meta-analytic results.

The products of data analysis will in general have been tested for their significance by way of confidence intervals for group means and regression coefficients. What their significance actually implies must

be interpreted by the reviewer. Since problem statements in research integration usually are cast as questions rather than hypotheses, the effort of interpretation seeks to draw conclusions that answer the questions, e.g., the effect of inquiry teaching on achievement in biology is (say) .7, statistically significant for $p < .05$.

From a descriptive standpoint, the message of a positive, negative, or zero result seems clear, whether for an effect magnitude, correlation, or relationship between a study characteristic and the dependent variable. Regarding inferential statistics, the issue seems not so clearcut. Where a confidence interval contains zero, the observed result may safely be claimed nonsignificant. Confidence intervals near zero should be accorded grains of suspicion, to shore up the assumption of a probabilistic sample of studies. Even where results appear statistically significant, any such outcome is suspect if it was produced by studies having poor research designs. Moreover, statistical significance fails to guarantee educational worth. Where data points are many, an apparently small effect may be significant but of little practical value to curriculum revision and/or policy decisions (Gay, 1981, p. 350). This raises the question, what is "small"? According to Glass, there seems little wisdom in attaching qualitative adjectives to effect magnitudes and correlations (Glass et al., 1981, p. 104). Depending on what benefits may be obtained at what cost, an effect size of 2.0 may be "poor" and one of .2 may be "good." Cohen (1977, p. 25) also has suggested that "small" and "large" are relative terms, not only to each other, but to the specific content area

of the topic under review. [Nonetheless, Cohen (1977, pp. 25-27) projects a small effect as being near .2 and a large effect as being about .8]. It seems the best assurance for a clearly significant result is an observed consistency in outcomes which have produced that result.

The purpose of meta-analysis is to answer research questions. It also serves to highlight issues that research has left unresolved. Jackson (1980) has observed that it is virtually impossible to perform an integrative review without encountering questions which demand further future inquiry. A good integration may sum up but does not end. In this sense, "it is only a beginning" (Feldman, 1971, p. 100).

Preparing the Report

The translation of reviewer's notes, coding sheets, computer printouts, and remembrances into a public document is a task with "profound implications" for accumulating knowledge (Cooper, 1982, p. 299). Other researchers, second-guessing inferences of the published report, should be able to applaud or condemn its integrity as advancing or impeding the progress of knowledge. Areas needing further inquiry also ought to be conveniently identifiable.

While the importance of publicly presenting research results is readily acknowledged, there seems a dearth of specific advice on how and what results should be presented. In general, researchers are accorded roomy latitude in document preparation, and especially this is true with regard to meta-analysis, where there exist no systematic

guidelines for evaluating the validity of results and no formal guidelines describing the structure of final reports. A widely held precept in all research is that reports should include enough information such that a reader may critically examine the evidence (Jackson, 1980). Within this generality, the reviewer probably will follow informal guidelines provided by reviews on the same/similar topics or perhaps will choose a format convenient for the particular problem.

In light of missing guidelines, the conclusions of meta-analysis stand to become less and less reliable unless measures are taken to make the process and its reporting more rigorous (Cooper, 1982). To bridge the void, a set of guidelines toward what constitutes an adequate report has been constructed, embracing criteria considered essential by developers and critics of meta-analysis. Figure 9 displays these guidelines as a checklist for report evaluation, with questions marked "yes" implying strength and other responses suggesting deficiencies.

A use of Figure 9 appears compatible with the presentation format already familiar to behavioral researchers: Introduction/Methods/Results/Conclusions and Discussion. This correspondence with primary research has prompted an inclusion of the format in the present model.

Model Specification

This chapter has developed a model for meta-analysis oriented toward research in education, with procedures synthesized from methodological literature.

Intro- duction	1. Is the general topic fully described?	
	2. Is prior evidence (e.g., reviews) critically examined?	
	3. Are research questions clearly stated? Comprehensive?	
	4. Are dependent variables operationally defined?	
Method	5. Are the target populations specified?	
	6. Is the search for available studies fully reported?	
	Are guidelines specified for excluding studies?	
	Do the studies represent the full body of research?	
	Is a bibliography of studies made available?	
	7. Is the list of study characteristics comprehensive?	
	Are study characteristics clearly coded?	
	Is research design quality rated for each study?	
	Are criteria reported for such ratings? Appropriate?	
	Are there at least a few good designs?	
Results	Are forms of publication entered as an independent variable?	
	Are types of measuring instruments recorded?	
	8. Are procedures described for deriving effect magnitudes or correlations?	
	Are all effect magnitudes or correlations listed in both size and direction?	
	9. Were strong attempts made to locate missing data?	
	10. Does analysis account for sample size variations?	
	Does analysis account for data dependence?	
	11. Is separate analysis performed for different constructs?	
	12. Are methods for data crosscutting clearly described?	
	13. Are appropriate techniques described for deriving group means and/or relationships?	
Conclusions /Discussion	14. Are inferential statistics applied to the data?	
	Are inferential procedures described? Appropriate?	
	15. Are the effects of missing data, poor research designs, or other biasing effects described and compensated?	
	16. Are results of analysis clearly and fully presented?	
	17. Are the accessible populations described?	
	Are inferences made to those populations?	
	18. Are conclusions clearly related to the research questions?	
	19. Does the meta-analysis determine what is known/not known about the topic?	
	20. Does the report suggest areas in need of further inquiry?	
	Does the report provide enough information such that a reader may critically examine the evidence?	

Figure 9. Checklist for evaluating reports of meta-analysis.

As shown by Figure 10, procedures in the model flow according to three phases:

1. Data collection, where research questions are formed, the body of studies is assembled, study characteristics are presented as independent variables, and study outcomes are presented as dependent variables;

2. Data analysis and interpretation, where homogeneous groups of effects or correlations are described by means and confidence intervals, heterogeneous groups are described by significant relationships between study characteristics and study outcomes, and these results are interpreted in terms of the research questions;

3. Report preparation, whose content rests upon prescribed evaluative guidelines.

As shown by dotted lines in Figure 10, procedures are consistent with the conceptual framework developed by Jackson (1980). Figure 10 and the associated text of Chapter III thus specify the model.

CHAPTER IV

REPORT OF META-ANALYSIS: EFFECTS OF HAND-HELD
CALCULATORS IN MATHEMATICS EDUCATION

The foregoing chapter develops a model for meta-analysis of research in education. To illuminate the process, an investigation of research within mathematics education was performed, as herein reported.

Introduction

In 1974, the National Council of Teachers of Mathematics issued a portentous statement urging calculator usage in the schools (NCTM, p. 3):

With the decrease in the cost of the minicalculator, its accessibility to students at all levels is increasing rapidly. Mathematics teachers should recognize the potential contribution of this calculator as a valuable instructional aid. In the classroom, the minicalculator should be used in imaginative ways to reinforce learning and to motivate the learner as he becomes proficient in mathematics.

Such a pronouncement may well have been viewed as a bit premature. The cost of a simple four-function device had only recently dropped below \$50; thus, at the time, maybe one in ten families and essentially no schools possessed the calculator (Schafer, Bell, and Crown, 1975). Further price declines could be envisioned, but an order-of-magnitude shrinkage (which occurred within the decade) must have seemed a far-off sparkle in the smoke of crystal balls.

Nonetheless, this small device appeared too powerful and potentially omnipresent to ignore. Computation was and is the central thrust of grade school mathematics, and of course such computation is the calculator's first purpose, lessening drudgery, saving time, and encouraging precision. Aside from the pure computational benefit, advocates of calculators found a host of other reasons for their use in schools (as summarized by Suydam, 1976, p. 12):

1. They aid in evaluating, understanding, and learning algorithmic processes;
2. They facilitate concept development;
3. They lessen the need for memorization, reinforcing basic facts with immediate feedback;
4. They enlarge the scope of problem solving by allowing realistic problems;
5. They motivate; they encourage discovery, exploration, and creativity;
6. They exist.

At the college level, calculators found willing acceptance (Suydam, 1978, p. 1). Secondary schools as well seemed warm toward a device which promised to ease computational effort, thus permitting the students more time to ponder mathematical ideas. However, their projected usage downward from the junior high school years provoked a barrage of skeptical comment (Suydam, 1978, pp. 5, 6):

1. They are not available to all students;
2. They could be used as substitutes for paper-and-pencil skills;

3. They encourage the false impression that mathematics is largely mechanical and involves no more than computation;
4. Not enough research exists on their effects.

Argument one would lose its edge when advances in microcircuitry did indeed cause prices to tumble. Points two and three perhaps built a straw man; few educators had promoted calculator use at the expense of basic skills or knowledge of the nature of mathematics (Suydam, 1978, p. 6). Still, a fear continued that the use of calculators was likely to render students mere button-pushers. Argument four was clearly valid and, from 1976 onward, induced a deluge of study reports, "one of the largest bodies of research on any topic or material in mathematics education" (Suydam, 1982a, p. 1). Many studies had one purpose, to establish whether calculators stand to damage basic skills. The answer consistently seemed to be "no," provided that the fundamentals had first been abstracted with paper and pencil (Suydam, 1979a, p. 3).

This seeming lack of harmful effects perhaps helped dissipate a dread of student dependence upon the devices. Resistance to their use has appeared to diminish (Suydam, 1982a, p. 1). By 1978, some 80 percent of precollege students owned or had access to the devices. By 1979, most high and grade schools had procured a set of calculators (East and Moursund, p. 16). In 1980, the NCTM reaffirmed its earlier position, urging calculator use at all grade levels (p. 8). Recent surveys of teachers and parents show positive attitudes toward the device when used to supplement paper-and-pencil commitments (Suydam, 1982a, p. 10).

Why is it, then, that most mathematics teachers seldom use the calculator in their classes?

The data on usage indicate that less than 20 percent of elementary teachers and less than 36 percent of secondary teachers have employed the calculator (Suydam, 1982a, p. 3).

A number of reasons perhaps have combined to produce this condition of apathy. Some educators may not be aware of the research; or, if aware, they may have ignored it, relying instead on their own insights into the issue (East and Moursund, 1979, p. 16). Confusion about the research may be a factor. Mixed results of student effects have been observed, and many studies have displayed ambiguous findings (Parkhurst, 1979). It also is the case that major studies, while recording no harmful effects of calculator use, have failed to show significant improvement either in student achievement or in attitude toward mathematics (e.g., Wheatley and Shumway, 1979). Teachers may thus be prone to ask, why bother with the calculator? They also may be at sea on how best to use it. Much research has focused on the likelihood of harming basic skills. Short effort has been given to exploring instructional methods for enhancing student achievement through a systematic use of calculators.

Moreover, such research is not apparently forthcoming. The bulk of calculator studies appeared from 1976 through 1980, at which time preoccupations were snared by the microcomputer (Suydam, 1981a, p. 1). Coincidentally or not, research on its precursor then displayed a sharp decline.

There seems a need to redirect attention toward the calculator and what Begle (1979) called its "crucial implications" for mathematics education (p. 21). A well-directed future path will require an accurate sense of present location, based on integrated facts from past research involving calculators. Hence, a study was conceived to yield such knowledge. Prior research has been conducted on a multitude of topics, e.g., attitudes of parents, teacher effects, types and levels of use, availability, and implications for teacher preparation. However, effects on pupils and students are education's prime concerns, and under the steorage of those concerns, the present study found its locus.

Problem Statement

The purpose of the study was to synthesize the findings of all existing research about the calculator's effects upon students in mathematics education. The problem thus was one of research integration. In light of the mass of existing reports, the process of meta-analysis was selected as best suited for a synthesis of results, to express what is known about student effects in terms of size as well as significance and to orient the path of future studies. Accomplishing those ends would constitute the problem's resolution.

Organization

This report adopts a style consistent with the standard format for behavioral research: introduction; method; results; and, conclusions/recommendations.

The background of the problem and its statement have already been provided. A review of literature, operational definitions, delimitations, limitations, and specific research questions will conclude the introduction.

In the method section, procedures and outcomes of data collection will be presented.

Procedures and findings of data analysis and discussions of those findings will appear in the section for results.

The final section draws conclusions and offers recommendations with regard to future research and to calculator usage in the classroom.

Review of Related Reviews

Within conventional research, the literature review involves a critical audit of pertinent records, mostly reports of kindred endeavors. However, in meta-analysis, such reports are the sources of data, tapped for their properties during the process of data collection. Consequently, this review does not survey individual studies but will concentrate instead on related reviews.

Since the calculator's advent in the mid-1970s, Suydam has become the leading chronicler of its research (1976; 1978; 1979a; 1980a; 1981a; 1982a). Her initial work provided to the National Science Foundation a status report of calculators in precollege education (1976). From questionnaire surveys and literature searches, arguments on usage pro and con were tabulated; data were presented on the calculator's prevalence and on parent/teacher attitudes toward

its use in the classroom. The literature searches also revealed the status of existing research on student effects. Not many useful reports were found to exist. Some two dozen studies with desk calculators had appeared across the years, and it was hoped that their findings would apply to the hand-held model. However, in view of their small sample sizes, short lengths of treatment, and patent confounding of variables, few results seemed to transfer, except for those couched in the broadest of statements, e.g., "children can learn to use calculators" and "children generally enjoy using calculators" (Suydam, 1976, p. 32). Eight studies using the hand-held device were identified and abstracted. Whether these were considered "research" appeared in doubt; investigators described their efforts as "exploration," "inquiry," and "preliminary study" (Suydam, 1976, p. 28). No attempt was made to integrate their findings. That calculators could be used to teach certain topics seemed clear, but it was certainly not clear that gains in achievement would be a result (Suydam, 1976, p. 27).

On a yearly basis from 1978 through 1982, the Calculator Information Center at Ohio State University issued a state-of-the-art review with Suydam as author. The first four reports were mainly concerned with progress toward installing calculators in the schools and with ways the devices were being employed (Suydam, 1978; 1979a; 1980a; 1981a). Increasingly, precollege students found the calculator available, with the trend inversely proportional to resistance regarding their usage. In the elementary grades, five uses seemed predominant:

(1) drill of basic facts; (2) checking work with paper and pencil; (3) games; (4) direct calculation in problems; and, (5) exploring mathematical ideas. Usage for drill was seen to decrease with increasing grade level. Above first grade, the most frequent usage was checking. At the secondary level, emphasis varied among: (1) direct calculation in problems; (2) recreation and games; (3) exploration; and, (4) use of textbook material that was calculator specific.

The final review of the Suydam series was explicitly devoted to a summary of research (1982a). Some 150 documents had amassed on various topics. Those concerning student effects addressed: (1) achievement within a traditional format; (2) attitude toward mathematics; and, (3) achievement within a special curriculum.

1. Achievement with traditional instruction. In 75 studies, comparisons were drawn between the achievements of experimental and control groups using or not using calculators as an aid within traditional instruction. These studies spanned all precollege grades, with distribution peaking at Grades 6 and 9. Some of the studies had more than one finding, such that 95 comparisons were made.

- a. 43 showed higher scores for (C)alculator than for (N)on-(C)alculator groups.
- b. 47 showed no apparent difference.
- c. 5 displayed advantage to the NC groups.

These data implied that calculator usage yields achievement as high or higher than when calculators are not used. However, it needs to be noted that sometimes the C groups were permitted calculators in

posttesting and sometimes they were not. A survey of this factor showed the preordained conclusion that the calculator's use on tests will build higher scores. Hence, the foregoing summary data were confounded by the mode of testing. With regard to problem solving, Suydam observed that calculators seem to benefit the process by removing the focus from computation and allowing concentration on approaches to solving the problem. No figures were quoted to substantiate the observation.

2. Attitude toward mathematics. It had been hypothesized that calculator usage will improve student attitudes toward mathematics. In 36 studies, the attitudes of C and NC groups were compared in post-treatment surveys. No differences were found in 30 comparisons; 6 reported improved attitudes from calculator usage.

3. Achievement in a special curriculum. Thirty-three studies investigated methods and materials which had been especially developed for instruction with the calculator. Experimental groups showed higher scores in about half the cases, and for the remainder there was no apparent difference between C and NC groups. Again, confounded variables call these findings into question. Were the cases of better achievement a result of developed instruction or the use of calculators or the two in combination?

The Suydam reviews are historical documents charting the thrust of calculators into the classroom. Other related reviews have appeared periodically, their concerns more narrow than the broad brush of the Suydam chronicles. It could be surmised that particular studies would

be surveyed by more than one reviewer, and this was the case. Hence, findings across reviews are not independent.

1. Parkhurst (1979) reviewed nine studies of junior high school students grouped for C and NC instruction. Most of the studies measured no significant differences in achievement. Where differences were present, the advantage went to calculator treatments. This positive trend was deemed to be an optimistic signal for the use of calculators in mathematics instruction. However, Parkhurst noted that the studies usually were based on vague or untold calculator uses during instruction. He further observed potential uncontrolled biases in results of all the studies, arising from teacher variability and the calculator's novelty.

2. Roberts (1980) examined 37 studies segregated by research at the elementary, secondary, and college levels. For elementary grades, 6 of 11 studies showed a computational advantage from the use of calculators, even though the device had been barred from posttests. Corresponding results were observed for 6 of 11 studies at the secondary level and for 7 of 8 reports with college bases. In all the studies remaining, no significant differences were found between the C and NC groups regarding their achievement. Across all grades, concept acquisition from the use of calculators appeared to be minimal, as did attitudinal changes. Roberts thought interpretation of the many study findings to be hampered by defective research designs, especially with regard to nonrandom assignment of students to treatments, uncontrolled teacher variables, the calculator's absence on the

posttests, and contamination due to NC students using calculators out of class.

3. Rabe (1981) assessed the work of 18 researchers who had conducted 26 studies. Fourteen calculator treatments yielded achievement gains of significant value, ten studies showed no significant differences between the C and NC groups, and two revealed higher achievement for the groups not using calculators.

4. From a general overview of seven studies, Neubauer (1982) concluded that a use of calculators prior to the junior high school years seems ill-advised, since students then are still learning the basics. He drew a similar conclusion with regard to low-achieving students, based on the "apparent ineffectiveness" of calculators as a viable learning aid (Neubauer, 1981, p. 7).

5. Sigg (1982) reviewed 22 studies sorted by effects on computation, attitude, and problem solving. In all these reports, computational scores appeared as high or higher for the C groups as the NC groups. Attitudes did not seem to change from use of the calculator. With regard to problem solving, study results appeared inconclusive. Sigg observed that benefits there are likely to be nominal unless strategies are taught along with calculator use for computation.

Eleven previous reviews have now been examined. Suydam's concluding review (1982a) appraised all calculator studies found to exist at the time; hence, her report may be considered a representative statement. None of the other reviewers disclosed how studies were selected, so the representativeness of their samples may be

questioned. All reviews, including Suydam's, proceeded by narration plus occasional vote-counting for a quantitative synthesis of findings. Thus, all reviews were subject to the faults of these approaches to research integration, i.e., subjective determinations; disregard of sample sizes, research design quality, and other properties which varied across studies; and, the absence of statistical analysis. Results were further confounded by the use and nonuse of the calculator in achievement testing. Except for Roberts (1980), no reviewer differentiated findings by this factor.

Their limitations notwithstanding, most of the reviews agreed that calculators need not damage basic skills. Consensus also existed that the attitudes of students seem unchanged by calculators. Still in controversy was their use at lower grade levels, and their status with regard to problem solving seemed less than determined. Studies with special methods and/or materials implied that the devices probably can serve pedagogical functions in addition to their use in computation. However, due to confounded criterion variables, results of these studies may well be viewed with suspicion.

The judgment seems fair that past reviews have produced few firm conclusions on effects of calculators. This lack of definitive statements inhibits an "accurate sense of present location" and thus helps justify a meta-analysis of the research.

Operational Definitions

As indicated by prior reviews, past studies of student effects have focused on the constructs achievement and attitude. In the

general case, a sample of students was given calculator treatment, while a second group received equivalent instruction but no calculators. Effects were then expressed in comparative terms. To highlight this base, the following definitions were constructed:

Achievement effect--The degree to which posttreatment outcomes for a calculator sample differed from the outcomes for a non-calculator group, where outcomes were measured by common achievement test instruments.

Attitude effect--The degree to which posttreatment outcomes for a calculator sample differed from the outcomes for a noncalculator group, where outcomes were measured by common survey instruments.

These effects should be construed as overarching; each may be described on multiple dimensions. Attitude was taken to include the six affective concerns of the Mathematics Attitude Inventory, Minnesota Research and Evaluation Project, 1972: attitude toward (or enjoyment of) mathematics; anxiety toward mathematics; self-concept in mathematics; motivation to increase mathematical knowledge; attitude toward mathematics teachers; and, perception of the value of mathematics. Achievement is related to skills demonstration. Thus, it was considered that its dominant concerns would consist of: skills acquisition (as measured at the end of treatment); skills retention (as measured after a period following treatment); and, transfer of skills to related mathematics. However, the construct must subdivide further, in light of: (1) distinctions among the particular skills that posttests try to

measure; and, (2) the use or nonuse of the calculator during post-testing.

1. Distinctions among mathematical skills regard the places where skills are applied, in working exercises (where operations are specified a priori) or in problem solving (where solvers must select the process). Definitions were constructed to recognize these categories:

Operational skills--Those involved with the solution of exercises.

Such skills may be considered as composite (total skill) or individual with regard to computation and concepts.

Problem solving skills--Those involved with the solution of verbal problems. Such skills may be considered as composite (total skill) or individual with regard to productivity (number of problems attempted) and selectivity (number of proper strategies).

2. The calculator's use or nonuse in testing presents two other dimensions for the achievement construct. Where calculators were barred from tests, the achievement effect depicts how well basic skills were maintained within the treatment period. With calculators used for both instruction and posttesting, the achievement effect depicts how well the calculator extended student skills. The following definitions emphasize these distinctions:

Maintenance effect--The degree to which a calculator sample maintained its paper-and-pencil skills, as compared to the similar skills of a noncalculator sample. A maintenance effect implies that calculators were not permitted during posttesting.

Extension effect--The degree to which a calculator sample extended its skills through unrestricted calculator use, as compared to the paper-and-pencil skills of a noncalculator sample. An extension effect implies that calculators were used in post-testing.

It hardly may escape notice that the construct of achievement is extremely comprehensive. For purposes of this study, variations were defined in terms of skills applied in basic operations and problem solving, with further distinctions derived from maintenance and extension concerns. Combining these factors generated Table 2, each of whose elements represents an achievement effect. This table was intended first of all to illuminate subtleties. It also found utility in the formulation of questions whose answers would be sought by the meta-analysis.

Research Questions

The purpose of meta-analysis is a production of statements (outcomes) which express what research findings, when statistically combined, reveal. Primary research predicts its outcomes by way of hypotheses. In contradistinction, the meta-analysis creates no outcome expectations; it asks questions and pursues collective answers in the body of all research that exists on a topic. This section develops (R)esearch (Q)uestions of calculator effects on student achievement and attitude.

Concerning achievement (see Table 2):

Table 2
Distinctions Among Achievement Effects

Skills	Basic Operations			Problem Solving		
	Composite	Computation	Concepts	Composite	Productivity	Selectivity
Acquisition	Maint.* Ext.**	Maint. Ext.	Maint. Ext.	Maint. Ext.	Maint. Ext.	Maint. Ext.
Retention	Maint. Ext.	Maint. Ext.	Maint. Ext.	Maint. Ext.	Maint. Ext.	Maint. Ext.
Transfer	Maint. Ext.	Maint. Ext.	Maint. Ext.	Maint. Ext.	Maint. Ext.	Maint. Ext.

*Maint. abbreviates maintenance and implies no calculator use during posttesting.

**Ext. abbreviates extension and implies a use of calculators during posttesting.

- RQ-1. What are the calculator effects regarding acquisition of composite operational skills in the maintenance sense? In the sense of extension?
- a. What are the effects regarding acquisition of computational skills in the maintenance sense? In the sense of extension?
 - b. What are the effects regarding acquisition of conceptual skills in the maintenance sense? In the sense of extension?
- RQ-2. What are the calculator effects regarding acquisition of composite problem solving skills in the maintenance sense? In the sense of extension?
- a. What are the effects regarding acquisition of problem solving productivity skills in the maintenance sense? In the sense of extension?
 - b. What are the effects regarding acquisition of problem solving selectivity skills in the maintenance sense? In the sense of extension?
- RQ-3. What are the analogous effects regarding retention of operational skills in the maintenance sense? In the sense of extension?
- RQ-4. What are the analogous effects regarding retention of problem solving skills in the maintenance sense? In the sense of extension?

RQ-5. What are the analogous effects regarding transfer of operational skills in the maintenance sense? In the sense of extension?

RQ-6. What are the analogous effects regarding transfer of problem solving skills in the maintenance sense? In the sense of extension?

These questions were drawn from distinctions observed in Table 2. In addition:

RQ-7. What are the calculator effects on student estimation skills?

Concerning attitude:

RQ-8. What are the calculator effects on attitude toward (or enjoyment of) mathematics?

RQ-9. What are the effects upon anxiety toward mathematics?

RQ-10. What are the effects upon self-concept in mathematics?

RQ-11. What are the effects on motivation to increase mathematical knowledge?

RQ-12. What are the effects on attitude toward mathematics teachers?

RQ-13. What are the effects on student perceptions of the value of mathematics in society?

All of the previous questions were intended to be tested by inferential statistics. Concerns also exist in areas less amenable to statistical inquiry: possible sex-related differences in calculator effects; and, effects of a special calculator curriculum. In the first case, groups for female calculator versus male noncalculator and groups for male calculator versus female noncalculator may not have been

statistically comparable. In the second case, the outcomes of studies with special curricula were confounded by the presence of both developed materials and the calculator treatment. Rather than address these concerns on a statistical basis, observational questions were posted:

RQ-14. Do effects on achievement and attitude seem to differ for male and female students?

RQ-15. What effects on achievement and attitude were observed as a result of special calculator instruction?

The foregoing itemization contains all the questions examined by meta-analysis. Implicit in each question are inquiries into relationships between calculator effects and properties which vary across studies, e.g., school grade level, student ability level, and length of calculator treatment. In forming the list of questions, little heed was paid to whether existing research would provide full responses. Unanswered questions help guide the perception of needs in future studies.

Method

Similar to conventional research, the method of meta-analysis is the process of data collection, but reports instead of subjects furnish the evidence. Finding and reviewing reports are thus the central aims of the method's procedures, which specifically include: delimiting the scope of the meta-analysis; identifying studies for the target population; selecting the body of useful reports; and, coding their properties and findings into a suitable format for statistical analysis. The resulting coded data then are the method's end products.

Delimitations

The topic for meta-analysis was initially formed in general terms, "student effects of calculators in mathematics education." This structure then was delimited with respect to school grade levels and to types of calculators.

1. The target population was established as mathematics pupils and students in regular classes of Grades K through 12. These are the grades where calculator use has been controversial or condemned.
2. The calculator type considered basic was the hand-held electronic device most likely to be found in present-day and future classrooms. However, a number of studies have employed the desk varieties, whose exclusion would have discarded potentially useful data. Hence, the desk styles were included, provided that they corresponded both in appearance and function to hand-held devices. Mechanical computing machines were not considered. Special-purpose models such as those for recreation and for preprogrammed feedback were also excluded.

Limitations

Since meta-analysis depends for its data on prior research, its findings may only reflect what that research has provided. Hence, the target population may not coincide precisely with the population accessible, and conclusions of meta-analysis may only pertain to the latter.

A second limitation results from the problem of finding all research that exists on a topic. The likelihood seems slender of success at locating all studies, and especially those unpublished. Hence, the validity of meta-analysis is dependent upon the degree to which its body of located studies represents the total research.

Collecting the Body of Studies

The best assurance for acquiring a representative sample is a careful tracking down of study reports. Collecting this body for meta-analysis advances in two stages: (1) finding studies with potential value; and, (2) selecting those whose data can be used.

The task of tracking studies for the present meta-analysis began with computer searches of the ERIC and Dissertation Abstracts International databases, under the descriptors "mathematics education," "mathematics instruction," "research," "research reports," "bibliographies," "literature reviews," "calculators," "hand-held calculators," "pocket calculators," and "minicalculators." Resulting abstracts yielded an initial reference listing of research on calculators, including journal articles, project reports, papers from scholarly meetings, reports to foundations, other studies recorded by ERIC, and dissertations. The printouts also cited reviews of research (already noted) and extensive bibliographies of reports (Machlowitz, 1976; Driscoll, 1979; Higgins, 1979; Suydam, 1979b, 1979c, 1980b, 1981b, 1982b, 1982c, 1983). These documents thus became sources for expanding the list of references. In addition, the following means were applied to finding studies:

1. Searching volumes manually of the Current Index to Journals in Education, Dissertation Abstracts International, and the annual bibliographies of the Journal for Research in Mathematics Education, from 1972 to date;
2. Requesting references directly from the Calculator Information Center at Ohio State University;
3. Tracking citations from study to study, during the task of surveying reports for their value to meta-analysis.

The foregoing efforts identified studies with possible merit. To derive the final body, each report was screened by way of predetermined guidelines for acceptance or rejection. Two guidelines referred to scope conditions:

1. The study must have been concerned with mainstream students in Grades K-12;
2. The study must have employed electronic hand-held or desk calculators.

Other guidelines were dictated by assumptions for the meta-analytic model (see Chapter III, "Collecting the Body of Studies"):

3. The study must have reported sufficient data for the calculations required in meta-analysis (e.g., group means and standard deviations);
4. The study must have provided continuous (rather than polychotomous) outcome data;
5. Each sample size of the study must have contained at least 10 subjects, except for those studies where classes were used as the units of analysis;

6. Where results of a primary study had appeared in more than one location, the more complete report would be used in the meta-analysis;

7. No report would be rejected on the grounds of flawed design. Some reports could be excluded by perusal of their abstracts. All other documents were reviewed hands-on. Libraries at the University of Tennessee, Knoxville, furnished copies of journal articles; ERIC reports were also there on microfiche copy. Other reports were gathered directly from authors and publishers, insofar as their addresses could be obtained. Dissertations were taken on loan from institutions which subscribe to free information exchange, and for those which do not, their products were purchased and reviewed on microfilm copy.

Reports accepted by way of the guidelines are listed in Appendix A. No report was found to contain discontinuous outcome data. No report was rejected solely on the basis of missing data; critical information was solicited from the authors. Three reports could not be located (Capoferi and Winowski, 1975; Nebel, 1978; Rudnick, 1978). The final group of documents included 12 journal articles, 12 ERIC reports, a project report, an unpublished report, and 53 dissertations, for a total of 79 studies.

Coding the Studies

The end collection of reports supplied all data to be used for this investigation. In one respect, the studies were similar; all related to effects of calculators. However, the studies differed in particular

characteristics, e.g., school grade levels, measuring instruments, sample sizes, and lengths of treatment. Hence, the studies were not replications; indeed, their findings were likely to vary as a function of the variable study properties. A singular strength of meta-analysis draws from its capacity to investigate relationships among the significant features of studies and their outcomes. To utilize this advantage, information for each study must be coded into a suitable format for statistical analysis. Study properties take the role of independent variables. Effect magnitudes related to outcome constructs (in the present case, achievement and attitude) are assigned as the dependent variables.

1. Representing independent variables. The initial coding effort concerned the identification of properties which conceivably affected study outcomes, i.e., identifying independent variables. For easing the forthcoming data analysis, each variable was to be defined as "categorical" or "continuous," with the former verbally described and the latter described by numbers. Inclusion of certain variables was mandated by the model (see Chapter III, "Representing Independent Variables"). Intuition yielded others, and the list was then extended by review of the studies themselves. These considerations yielded categoricals as follows, with the parenthetical letters used as coding symbols:

- a. Form of publication--(J)ournal; ERIC; (D)issertation; (O)thers.

- b. Type of measuring instrument--(S)tandardized; (X_r), experimenter designed with reliability quoted; ad hoc, experimenter or teacher designed without reliability assessment.
- c. School grade level--K through 12.
- d. Student ability level--Mixed; (L)ow; (A)verage; (H)igh.
- e. Student socioeconomic status, SES--Mixed; (L)ow; (M)iddle; (U)pper.
- f. Subject matter--Particular areas studied during the treatment.
- g. Experimenter bias, a judgment of the researcher's direct involvement in the treatment--(M)uch; (L)ittle.
- h. Curriculum--(T)raditional, except for the calculator; (Sp)ecial, including developed materials or methods.
- i. Calculator usage--(F)unctional, essentially for computation; (P)edagogical, an aid in teaching/learning concepts.
- j. Calculator availability--ratio of one device to the number of students who used it (e.g., 1/2).
- k. Ethnicity--predominant ethnic/cultural group within the experiment.
- l. Sex-related differences in calculator effects--none; results (F)avored (F)emales; results (F)avored (M)ales.

Continuous variables were considered as follows, with related metrics shown in parentheses:

- m. Year of report publication.
- n. Length of calculator treatment (school days).
- o. Amount of calculator drill and practice prior to treatment (hours).

- p. Time of calculator use per treatment session (minutes).
- q. Teacher training with the calculator (hours).
- r. Teacher experience (years).
- s. Retention period (weeks), for studies which examined retention.
- t. Research design rating (1 to 3).

Each study was coded with respect to the foregoing variables. It ought to be noted that only one treatment was involved, the use of calculators; hence, the type of treatment was constant rather than variable. Results of the coding are tabulated in Part 1 of Appendix B. Some data were missing in most of the studies: ethnicity; amount of drill and practice; time of calculator use per treatment session; teacher training; and, teacher experience. Thus, these variables were excluded from appended tabulations. The method for rating research design quality needs elaboration. Each study was assessed with concern for eight criteria: (1) problem definition; (2) population description; (3) sampling procedures; (4) error control; (5) instruments; (6) data analysis; (7) conclusions; and, (8) the overall report. Under predetermined guidelines (which Appendix B describes), each criterion was evaluated on a scale of 1 to 3, except that sampling and error control were accorded double value and were rated 1 to 6. Resulting numbers were summed and divided by 10 to yield a numerical rating.

2. Representing dependent variables. Focal areas of this investigation were the constructs achievement and attitude. Since both

were present in the calculator studies, they prescribed dependent variables as follows:

- a. Achievement effects at all the sublevels attending skills acquisition, retention, and transfer with regard to basic operations and problem solving in the senses of maintenance and extension;
- b. Attitude effects concerning attitude toward mathematics, anxiety toward mathematics, self-concept in mathematics, motivation to increase mathematical knowledge, attitude toward mathematics teachers, and perception of the value of mathematics in society.

For purposes of data analysis, quantification of all these variables to a common metric would be required. It may be recalled that effect was defined as the extent to which posttreatment outcomes for a calculator sample differed from the outcomes for a noncalculator sample. Thus, each effect could be expressed in terms of its size, i.e., an effect size ES determined as the difference between the means of (E)xperimental (calculator) and (C)ontrol (noncalculator) outcome data, with this difference standardized to a scale-invariant metric through its division by within-group standard deviation S^P . Hence:

$$ES = \frac{\bar{Y}^E - \bar{Y}^C}{S^P} . \quad (R-1)$$

This is the basic effect size equation defined for a general meta-analysis (see Chapter III, "Representing Dependent Variables").

Thus, it became the core equation used to express calculator effects, with $\bar{Y}$ referred to a raw-score mean, an adjusted group mean, or a gain-score mean regarding either achievement or attitude. Under equation R-1, ES may be observed as zero, positive, or negative. Zero ES suggests equality of E and C means and thus no effect. With one exception, a positive value shows favor to the calculator-treated group, and negative results imply the treated group has suffered. The exception regards anxiety toward mathematics; there the signs of favor are reversed, i.e., a negative ES implies a lower anxiety level for the calculator sample.

Where studies had tested outcome data in terms of t or two-group F , the following variations of equation R-1 could be employed:

$$ES = t \sqrt{\frac{1}{n^E} + \frac{1}{n^C}} = \sqrt{F} \sqrt{\frac{1}{n^E} + \frac{1}{n^C}} \quad , \quad (R-2)$$

with n 's referred to sample sizes. However, the use of equation R-1 was always required where studies had employed analysis of variance (ANOVA) or covariance (ANCOVA) with more than two groups. In such cases, $\bar{Y}^E$, $\bar{Y}^C$, and S^P were data demanded. Where E and C group standard deviations were provided, S^P was computed as the familiar pooled value. Where E and C group standard deviations were omitted, S^P was computed as the square root of within-cell variance or error mean square (whichever was provided in the summary tables), both of which are estimates of the population variance.

All calculator effects were coded by way of the foregoing methods. Summary tables in each report provided the data, illustrated by the sample coding sheet of Figure 11. A sheet such as this was developed for each calculator effect size, recording not only ES data but also information for the independent variables related to that effect. The process of data collection technically was ended when all such coding sheets had been accumulated.

However, equations R-1 and R-2 produced ES's which were biased and needed correction to more precise values (see Chapter III, "Preparing for Data Analysis"). Compensation was provided for distribution bias, errors of measurement, and data dependence.

a. Distribution bias and errors of measurement. Each raw ES is a sample statistic which estimates the underlying population effect size δ . Thus, ES possesses sampling distribution, biased as a function of degrees of freedom $m = n^E + n^C - 2$. This bias may be corrected by a factor closely estimated as $c(m) \approx 1 - 3/(4m - 1)$. Measurement error is related to instrument reliability r^{YY} , which error may be corrected by a square root factor. For each raw ES computed by equations R-1 or R-2, an error-free value g was derived as:

$$g = \frac{c(m)ES}{\sqrt{r^{YY}}} \quad . \quad (R-3)$$

All the coding sheets (including that of Figure 11) reflected these corrections. Frequently, studies failed to include reliability figures for their test and survey instruments. In such cases, the conservative value of $r^{YY} = 1$ was applied.

Author(s) DoeCONSTRUCT Acq. comp. P.S. skills
(maintenance)Study ID No. 00

CATEGORICAL INDEPENDENT VARIABLES

Form of publication	D
Type of instrument	Std
Grade level	6
Ability level	Mix
SES	Mix
Subject matter	Decimals
Experimenter bias	L
Curriculum	T
Calculator usage	F
Calc. availability	1/1
Ethnicity	Mix
Sex-related differences	None

CONTINUOUS INDEPENDENT VARIABLES

Year of publication	1978
Length of treatment	30 days
Drill and practice	2 hrs
Calc. use per session	
Teacher calc. training	
Teacher experience	5-8 yrs
Retention period	

RESEARCH DESIGN RATING: 2.2

1. Problem definition = 3 2. Population description = 2 3. Sampling procedures = 2 4. Error control = 5 5. Instruments = 3 6. Data analysis = 3 7. Conclusions = 2 8. Report = 2	$\frac{22}{10}$
---	-----------------

$\bar{Y}^E$	$\bar{Y}^C$	S_E^2	S_C^2	S^P	t	n^E	n^C	Raw ES	rYY	c(m)	g
21.79	19.38			5.55		48	50	.434	.83	.992	.473

Average g for
dependent data

Figure 11. Example coding sheet for meta-analysis.

b. Data dependence. Most of the studies yielded more than one effect size g computed for the same experimental and control groups. Resulting values were thus not independent, except as they related to different outcome (dependent) variables. Contamination by data dependence could in large measure be avoided by separate analyses of the various outcome constructs, but on occasion the E and C groups in one study yielded more than one effect size for the same outcome variable. When such a case occurred, the cluster of dependent g 's was averaged to provide a quasi-equivalent g . This value was then used in the data analysis.

Part 2 of Appendix B displays all effect sizes g for each report in the final body of studies. Values ranged from -1.065 to $+1.966$. Each study typically provided more than one calculator effect; the high was 35; the low was 1. The 79 reports thus yielded 529 effect magnitudes, the basic data for statistical analysis.

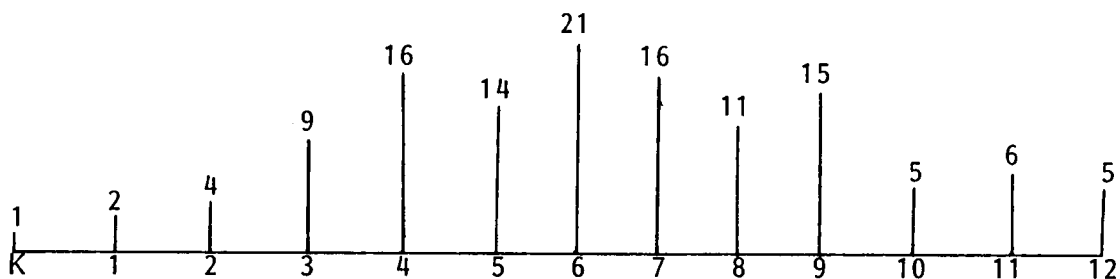
All coding of independent variables and effect size calculations were performed by the investigator. Intercoder reliabilities thus did not apply.

Features of the Data

Previously, it was mentioned that the findings of meta-analysis are constrained by the data existing, e.g., a target population may be projected, but the actual population is determined by prior research. To derive a sense of such constraints, the information of Appendix B was studied with regard to salient features, especially those concerning data distributions, calculator usage during the treatments, and

ranges of continuous independent variables. A survey of these features will aid the interpretation of findings in subsequent data analysis.

1. Distribution of studies by school grade level. The target population consisted of mainstream mathematics pupils and students in Grades K-12. Within the body of 79 studies, the number of experiments which examined each grade level was distributed as follows:



(A number of studies had attended more than one grade level, and so the composite of experiments exceeds 79.) The early grades K, 1, and 2 were not well-represented by the data, and the later grades 10, 11, and 12 were only moderately represented. Inferences drawn from data analysis thus will achieve their best application in the intermediate and junior high school grades.

2. Effect size distributions by dependent variable. A total of 529 effects would be available for use in the data analysis. Fifty-three effects referred to special calculator instruction compared with traditional instruction without calculators; these were the data referred to special curricula. From studies where experimental and control groups received equivalent instruction except for calculators, 476

effects were derived, including 405 for the achievement construct and 71 for the attitude construct. Table 3 displays the quantities of effect magnitudes contained in subsets for the various achievement dependent variables. (An additional five effects regarded student estimation skills.) Effects referring to skills acquisition seem thoroughly researched, thus giving a solid foundation for data analysis. Much less data existed with regard to skills retention and transfer.

With reference to the attitude construct, the following effect size quantities existed:

Attitude toward mathematics--56;

Anxiety toward mathematics--4;

Self-concept in mathematics--5;

Motivation to increase mathematical knowledge--2;

Attitude toward mathematics teachers--2;

Perception of the value of mathematics in society--2.

3. Calculator usage in the treatments. Etlinger (1974) has described two roles in education for the calculator: functional, for computation; and, pedagogical, to facilitate teaching and learning. Within the body of studies for the present meta-analysis, the functional role by far outstripped pedagogical usage. Except for the studies with special curricula (and their confounded outcome variables), the calculator seldom was used in concept development or investigation of problem solving strategies. Predominant uses were drill and practice, checking work done with paper and pencil, and direct computation in exercises and problems.

Table 3
Numbers of Effect Sizes in Subsets for Achievement Dependent Variables

Skills	Basic Operations						Problem Solving					
	Composite			Concepts			Composite			Productivity		
	Maint.	Ext.	Maint.	Ext.	Maint.	Ext.	Maint.	Ext.	Maint.	Ext.	Maint.	Ext.
Acquisition	87	19	58	31	39	1	67	31	1	7	6	13
Retention	13	5	5	2	3	-	1	3	-	-	1	2
Transfer	4	-	-	-	-	-	-	1	-	-	-	-

4. Ranges of continuous independent variables. The 79 reports were generated from 1969 through 1982, with most of the studies appearing from 1976 through 1980.

Lengths of calculator treatment ranged from zero to a full school year. (Zero treatment was applied in the occasional study which simply compared the test scores of a group that used calculators and a group that worked with paper and pencil.) Fifty-seven studies endured for four weeks or longer.

In studies of skills retention, the retention period ranged from 2 to 12 weeks, with a median value of 4.

Research design ratings could range from 1 to 3. Actual ratings ranged from 1.2 to 2.7. Scores from 1.8 to 2.2 were considered average ratings. Under this criterion, 60 of the 79 studies showed adequate or better designs.

Results

The foregoing "Method" provided a set of error-free effect magnitudes and related study properties as the end products of data collection. It remained to analyze these data, responding as fully as possible to Research Questions specified within the "Introduction." The present section describes the procedures and findings of data analysis.

Overview of Procedures

The central outcomes of meta-analysis are means and confidence intervals for the groups of effect magnitudes related to the questions

of research. To permit the derivation of such inferential statistics, assumptions were applied as follows (see Chapter III, "Data Analysis--Experiments"):

1. All source data for effect size calculations were assumed to have met the requirements for a valid t test;
2. Instruments which measured the same construct were assumed to be linearly equatable;
3. Large-sample approximations were assumed to be valid;
4. The body of studies was assumed to constitute a probabilistic sample.

Under these assumptions, the mean of an effect size group represents an overarching effect that integrates all study findings from which the effect magnitudes were derived. A confidence interval (CI) about this mean provides a test of the null hypothesis that the mean is not significantly different from zero. Where the CI does not contain zero, the null hypothesis is rejected.

However, for any such mean to be valid, each effect size in the group must be an estimator of the population effect size δ . Where this condition exists, the effect size group is said to be homogeneous, and any observed effect size variance is considered a result of sampling error (chance). The mean is then a proper estimator of the population δ , and confidence intervals for the mean are also valid.

It seems likely that the variance of some effect size groups will exceed the value expected from sampling error. Such a group is then deemed heterogeneous, and its mean cannot be said to represent a

single underlying population effect. Outlier data points (see Chapter III, "Simple Description of Effect Magnitudes") are possible causes of variability beyond sampling error, and the deletion of such outliers may produce a homogeneous effect size group whose mean and CI's are valid. The other likely cause of heterogeneity is a study characteristic with which effect size is related. Where such an independent variable moderates the dependent variable, the underlying population effects are also variable, and two or more means are embedded within the effect size collection. Further exploration to explain the heterogeneity then is required.

Hence, for any effect size group, analysis begins with determining whether or not the group is homogeneous. This determination proceeds by way of the statistical test described for a group of k effects (see Chapter III, "Testing the Homogeneity of Effect Size Groups"):

1. Establish the hypothesis $H_0: \delta_i = \delta, i = 1, \dots, k$, versus the alternative that at least one δ_i is different from the others.

2. Calculate the test statistic H_T . Where g_i , n_i^E , and n_i^C are respectively an effect size and its related experimental and control group sample sizes, the population variance $\sigma_i^2(\delta_i)$ is estimated by:

$$\sigma_i^2(g_i) = \frac{n_i^E + n_i^C}{n_i^E n_i^C} + \frac{g_i^2}{2(n_i^E + n_i^C)} \quad (R-4)$$

Then the test statistic H_T is calculated as:

$$H_T = \sum_{i=1}^k \frac{g_i^2}{\sigma_i^2(g_i)} - \frac{\left[\sum_{i=1}^k \frac{g_i}{\sigma_i^2(g_i)} \right]^2}{\sum_{i=1}^k \frac{1}{\sigma_i^2(g_i)}} . \quad (R-5)$$

H_T is distributed as χ^2 with $(k - 1)$ degrees of freedom.

3. Do not reject H_0 if the calculated H_T is smaller than the critical χ^2_{k-1} at a desired significance level. The group of k effects is then judged homogeneous, thus revealing the existence of a common population effect size δ across all ranges of study characteristics. It is natural to estimate the population δ by pooling g_i 's. A mean effect size made more precise by weighting with regard to sample sizes is derived as:

$$\bar{g}^w = \frac{\sum_{i=1}^k \frac{g_i}{\sigma_i^2(g_i)}}{\sum_{i=1}^k \frac{1}{\sigma_i^2(g_i)}} . \quad (R-6)$$

Where the population variance $\sigma^2(\delta)$ is estimated by:

$$\sigma^2(\bar{g}^w) = \frac{1}{\sum_{i=1}^k \frac{1}{\sigma_i^2(g_i)}} , \quad (R-7)$$

a $100(1 - \alpha)$ percent confidence interval at the significance level α is derived from:

$$\bar{g}^w - z_{\alpha/2} \sigma(\bar{g}^w) \leq \delta \leq \bar{g}^w + z_{\alpha/2} \sigma(\bar{g}^w) \quad , \quad (R-8)$$

with $z_{\alpha/2}$ obtained from tables of the normal distribution.

4. Reject H_0 if the calculated H_T is larger than the critical χ^2_{k-1} . The group of k effects is then judged heterogeneous, indicating the presence of effect size variance unexplained by sampling error. This variance should (if possible) be removed in one or both of two ways: deleting outlier data points; and, finding homogeneous subsets in the overall effect size collection (see Chapter III, "Simple Description of Effect Magnitudes" and "Testing Homogeneity Within Effect Size Classes"). Means and confidence intervals for the resulting effect size groups may then be derived by equations R-6 through R-8. Where heterogeneity cannot be removed, its cause should be investigated in a search for significant relationships between study properties and effect magnitudes (see Chapter III, "Analyzing Heterogeneous Effect Size Groups").

It ought to be observed that all attempts to find a cause of heterogeneity may be frustrated. The cause would then most likely be a variable not included for data analysis. If the heterogeneity of a group remains unexplained, this outcome should be clearly stated. The group's effect size mean may then be offered as a descriptive statistic from which no statistical inferences may be drawn.

Presentation of Findings

To explore calculator effects on student achievement and attitude, 529 effect sizes g were available for analysis. These data were first

partitioned with respect to the Research Questions. Each resulting effect size subset was then treated in accordance with the previous procedures. All calculations were done on the Intertec SuperBrain microcomputer by way of the specially written program of Appendix C. Consistent with the levels typically used in primary research, homogeneity was tested at a .05 level of significance, and a confidence level of 95 percent was applied to express the intervals about effect size means.

Results of analysis are provided in the present section. Each Research Question will be restated and will usually be considered with reference to a data summary figure which displays the procedural sequence for effects determinations. Versions of this figure will appear throughout the analysis, so its format bears attention to project its basic contents. As shown by the example Figure 12, the format offers space for recording results of analysis at one, two, or three levels:

Stage 1. Analysis at the top level will have considered the Research Question directly, disregarding properties that varied across studies. All related effects will be listed in sequence from smallest to largest, thereby defining the numerical distribution of the data. To illustrate this distribution and to identify outlier data points, these effects will be plotted in a box-and-whiskers diagram, whose central box contains the middle 50 percent of effect size data. (Where there exists too little data for the box-and-whiskers to be useful, it will be omitted.) Fences "f" are marked on the "whiskers" at a distance one

Top Level Effect Sizes g	(k effect sizes listed from smallest to largest)		
Box-and-Whiskers			
Results of Test for Homogeneity	Homogeneous if $H_T < \text{critical-}\chi^2_{k-1}$ $\bar{g}^W$ = weighted effect size mean () = 95 percent confidence interval Heterogeneous if $H_T \geq \text{critical-}\chi^2_{k-1}$		
2nd Stage g's are top level subsets	Subset 1	Subset 2	Subset 3
	(effect size listing)	(effect size listing)	(effect size listing)
Results of Test for Homogeneity	Suppose homogeneous $\bar{g}^W =$ ()*	Suppose heterogeneous	Suppose homogeneous $\bar{g}^W =$ ()*
3rd Stage g's are subsets of a 2nd Stage subset	Subset 2-1	Subset 2-2	Subset 2-3
	(effect size listing)	(effect size listing)	(effect size listing)
Results of Test for Homogeneity	Suppose heterogeneous $\bar{g}^W =$	Suppose homogeneous $\bar{g}^W =$ ()*	Suppose homogeneous $\bar{g}^W =$ ()*

° Indicates outlier excluded
d Indicates previously dependent data

*Indicates rejection of null hypothesis

Figure 12. Format of data analysis summary figures.

and a half times the width of the box, as measured from its sides. Data points outside these fences will be considered outliers and will be identified with a superscript "o" (e.g., 1.749^o). Usually, such outliers will be excluded from analysis, to avoid their large contributions to the variance of the effect size group. If the group was found homogeneous, its mean $\bar{g}^w$ is the effect. The mean and its CI will be presented, and analysis is done. A star "*" on the CI indicates a mean effect significantly different from zero. In general, positive effects have favored calculator groups, while negative effects imply the calculator groups have suffered.

Stage 2. Whenever a top level g-group was found to be heterogeneous, lower level analysis has been required. This second effort has concerned a search for homogeneous subsets in the top collection, subsets related to a study characteristic (e.g., student ability level) found to influence effect sizes in the top level grouping. Where such a relationship has existed, the summary figure provides information from 2nd Stage analysis, starting with lists of effect sizes grouped according to the moderating variable (e.g., low ability, general or mixed abilities, and high ability). It needs to be noted that grouping the top level data for analysis of its subsets may permit the addition of effect size data which had been excluded at the top level on the basis of dependence. (For example, a number of studies presented findings for a total group of students and also gave findings for low and high ability students within the same group. The latter groups were thus dependent in relation to the former, and all could not be

used within the same effect size group.) Each data point thus added for the 2nd Stage analysis is to be identified by a superscript "d" (e.g., $-.229^d$). Further, it should be noted that additional outliers may be identified and deleted from one or more subsets. Each of these will then have been tested for homogeneity. Where homogeneous subsets were discovered, their means and CI's will be presented, and analysis in their regard is done.

Stage 3. If any 2nd Stage subset was heterogeneous, analysis at a still lower level has been required. Again, the added effort has concerned a search for homogeneous subsets formed according to another moderating variable (e.g., school grade level). Hence, 3rd Stage analysis has proceeded along the guidelines of Stage 2. Wherever homogeneity was found, the mean and CI for that group will be presented. If any 3rd Stage subset was heterogeneous, its number of effects was probably too small for further subdivision and test of homogeneity. The cause of heterogeneity was then investigated by correlational analysis of effect magnitudes and other independent variables. Whenever the cause could not be explained (the usual case), the mean of the heterogeneous group is offered as a descriptive statistic.

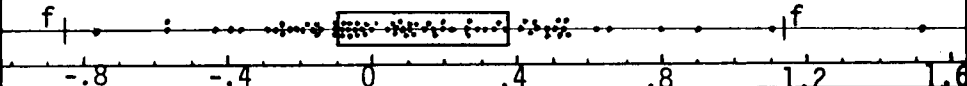
Analysis results are now presented, question by question. A summary and discussion of findings will then be provided.

RQ-1. What are the calculator effects regarding acquisition of composite operational skills in the maintenance sense? In the sense of extension?

It may be recalled that the distinction between maintenance and extension effects concerns the calculator's status during posttesting. A maintenance effect implies that all tests were worked with paper and pencil, while an extension effect implies unlimited use of calculators by experimental groups. Hence, the maintenance part of RQ-1 could well have been stated as follows:

At the end of calculator treatment, how well have calculator groups maintained their paper-and-pencil skills in basic operations, as compared with the control groups?

Eighty effect magnitudes initially were available to address this question; all of these are listed at the top of Figure 13, the analysis summary. The box-and-whiskers plot for these effects reveals a reasonably normal distribution, that is, no tendencies to cluster are apparent. Nonetheless, the group with one outlier excluded was heterogeneous, since the test statistic $H_T = 134.65$ exceeds the critical χ^2_{78} value of 99.61 at a .05 significance level. This result led to 2nd Stage analysis, which revealed a significant relationship between effect size and student ability level. Effect sizes thus were grouped as "low ability," "general ability" (the mixture of abilities observed in a regular classroom), and "high ability." These groups included seven bits of previously dependent data. The sets for low and high abilities each were found homogeneous, with respective mean effects of $-.107$ and $-.031$. Neither value was significantly different from zero at the .05 level, since the 95 percent CI's contained the null. The grouping for general student ability was heterogeneous. In 3rd Stage

Top Level Effect Sizes g	-.766 -.253 -.152 -.078 -.018 .075 .173 .291 .440 .516 -.572 -.240 -.138 -.067 .004 .088 .182 .302 .444 .532 -.571 -.226 -.107 -.060 .009 .100 .198 .342 .478 .624 -.430 -.211 -.106 -.055 .045 .102 .203 .349 .494 .645 -.393 -.189 -.100 -.055 .061 .111 .220 .378 .502 .797 -.373 -.182 -.096 -.044 .068 .120 .258 .414 .508 .900 -.291 -.164 -.092 -.042 .068 .141 .261 .423 .509 1.102 -.266 -.153 -.090 -.019 .068 .153 .270 .438 .509 1.510°									
Box-and-Whiskers										
Results of Test for Homogeneity	$H_T = 134.65$ $\text{Chi}^2_{78} = 99.61$ Heterogeneous									
2nd Stage g's by Ability Level	Low		General						High	
	-.766 .388 ^d		-.571 -.138 -.055 .102 .270 .494						-.393	
	-.572 .645		-.373 -.126 ^d -.018 .111 .291 .502						-.085 ^d	
	-.430 1.102°		-.266 -.107 .004 .141 .302 .508						-.078	
	-.291		-.253 -.106 .009 .153 .342 .509						-.019	
	-.169 ^d		-.240 -.100 .045 .173 .349 .509						.019 ^d	
	-.164		-.226 -.096 .068 .182 .378 .516						.120	
	-.044		-.211 -.092 .068 .198 .414 .532						.478	
	-.042		-.189 -.090 .068 .205 ^d .423 .624							
	-.042 ^d		-.182 -.067 .075 .220 .438 .797							
	.061		-.153 -.060 .088 .258 .440 .900							
	.203		-.152 -.055 .100 .261 .444							
Results of Test for Homogeneity	$H_T = 14.84$ $\text{Chi}^2_{12} = 21.03$ Homogeneous $\bar{g}^W = -.107$ (-.26, .04)		$H_T = 96.93$ $\text{Chi}^2_{64} = 83.66$ Heterogeneous						$H_T = 3.86$ $\text{Chi}^2_6 = 12.59$ Homogeneous $\bar{g}^W = -.031$ (-.24, .17)	
3rd Stage g's, General Ability by School Grade Level	Grade 4 -.571 .068 -.373 .100 -.226 .153 -.055		Grades K-12, except Grade 4 All g's listed for 2nd Stage General Ability, except for Grade 4 g's and the outlier .900							
Results of Test for Homogeneity	$H_T = 4.58$ $\text{Chi}^2_6 = 12.59$ Homogeneous $\bar{g}^W = -.152$ (-.30, -.004)*		$H_T = 72.54$ $\text{Chi}^2_{56} = 74.45$ Homogeneous $\bar{g}^W = .137$ (.08, .19)*							

°Indicates outlier excluded

^dIndicates previously dependent data

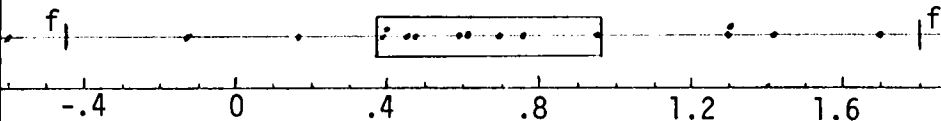
*Indicates rejection of null hypothesis

Figure 13. Analysis summary: Maintenance effects regarding achievement in acquiring composite operational skills.

analysis, the effect size group was found related to school grade level; the Grade 4 subset of effects had caused the heterogeneity. When the group for Grade 4 was separated from its parent subset, both resulting groups were homogeneous. For Grade 4, the mean effect was $-.152$, significantly different from zero. The group reflecting all grades but Grade 4 possessed a mean effect of $.137$, also a significant value. With these homogeneous findings, the analysis was concluded.

The extension part of RQ-1 concerns the degree to which the calculator groups extend their operational skills as a result of treatment and the calculator's usage during posttesting. Seventeen effect magnitudes initially were available to address the question regarding extension. These effects are listed at the top of Figure 14, the analysis summary. Even with two outliers excluded, the data were heterogeneous ($H_T = 37.85 > 23.68$). This result led to 2nd Stage analysis, which revealed a significant relationship between effect size and student ability level. Hence, the low ability data were grouped separately and were found to be homogenous with a mean effect of $.325$, significantly different from zero. Only two data points existed for the high ability students; thus, a group was not determined. For general students, the subset remained heterogeneous, and no significant moderating variable could be discovered. Its mean of $.737$ is quoted as a descriptive statistic.

The basic RQ-1 has concerned composite operational skills, considered to subsume computation and concepts. To deal with these sublevel interests, RQ-1 contains subquestions; these will now be regarded in turn.

Top Level Effect Sizes g	$-.603^{\circ}$.393 .473 .690 1.294 1.687 $-.124$.399 .587 .739 1.295 1.966 $^{\circ}$.165 .452 .609 .947 1.419					
Box-and-Whiskers						
Results of Test for Homogeneity	$H_T = 37.85$ $\chi^2_{14} = 23.68$ Heterogeneous					
2nd Stage g's by Ability Level	Low		General			
	$-.124$.393 .399 .587		$.197^d$.690 1.295 $.426^d$.739 1.419 .452 .947 1.687 .609 1.294			
Results of Test for Homogeneity	$H_T = 2.28$ $\chi^2_3 = 7.82$ Homogeneous $\bar{g}^W = .325$ $(.04, .61)^*$		$H_T = 32.99$ $\chi^2_{10} = 18.31$ Heterogeneous $\bar{g}^W = .737$			

$^{\circ}$ Indicates outlier excluded
 d Indicates previously dependent
data

*Indicates rejection of null
hypothesis

Figure 14. Analysis summary: Extension effects regarding achievement in acquiring composite operational skills.

a. What are the effects regarding acquisition of computational skills in the maintenance sense? In the sense of extension?

Fifty-three effect magnitudes initially were available to address the maintenance part of this question. These effects are listed at the top of Figure 15, the analysis summary. With two outliers excluded, the group remained heterogeneous ($H_T = 125.38 > 67.50$). This result led to 2nd Stage analysis, which revealed a significant relationship between effect size and student ability level. Effect sizes thus were grouped as "low," "general," and "high," including five bits of previously dependent data. The subsets for low and high abilities each were found homogeneous, with nonsignificant mean effects of $-.009$ and $.024$. Even disregarding two outliers, the grouping for general student ability stayed heterogeneous. In 3rd Stage analysis, this effect size group was found related to school grade level; Grade 4 effects apparently had created the heterogeneity. When these eight effects were separated from their parent subset, the resulting group reflecting all school grades but Grade 4 was found homogeneous, with a significant mean effect of $.124$. The Grade 4 effects were heterogeneous, and no significant moderating variable could be determined. The subset's mean of $-.181$ is quoted as a descriptive statistic.

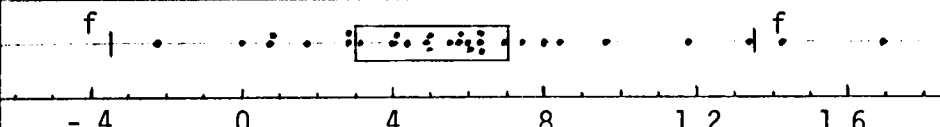
Regarding extension effects for computational skills, 31 effect magnitudes were used in addressing the question. These effects are listed at the top of Figure 16, the analysis summary. With the two outliers excluded, the group was found homogeneous ($H_T = 37.06 < 41.34$), with a significant mean effect of $.636$.

Top Level Effect Sizes g	-1.065° -.693 -.680 -.547 -.506 -.479	-.448 -.404 -.386 -.357 -.349 -.322	-.232 -.156 -.146 -.113 -.095 -.080	-.078 -.062 -.001 .010 .022 .048	.052 .055 .085 .088 .090 .091	.099 .113 .127 .134 .141 .150	.176 .195 .212 .251 .263 .299	.346 .357 .364 .500 .604 .650	.780 .785 .794 .906 1.564°
Box-and-Whiskers									
Results of Test for Homogeneity	$H_T = 125.38$ $\chi^2_{50} = 67.50$ Heterogeneous								
2nd Level g's by Ability Level	Low	General						High	
	-.680 -.357 -.156 -.130 ^d -.062 -.001 .328 ^d .780 .906°	-.693 -.547 -.506 -.479 -.448 -.404 -.386 -.349 -.322	-.232 -.146 -.113 -.095 -.080 -.078 -.010 .048 .052	.055 .085 .088 .090 .091 .099 .113 .127 .134	.141 .150 .176 .212 .251 .263 .299 ^d .319 ^d .346	.357 .364 .500 .604 .650 .785° .794°	.003 ^d .020 ^d .022 .195		
Results of Test for Homogeneity	$H_T = 13.91$ $\chi^2_7 = 14.07$ Homogeneous $\bar{g}^W = -.009$ (-.20, .18)	$H_T = 83.64$ $\chi^2_{40} = 55.76$ Heterogeneous						$H_T = .19$ $\chi^2_3 = 7.82$ Homogeneous $\bar{g}^W = .024$ (-.21, .25)	
3rd Level g's, General Ability by School Grade Level	Grade 4	Grades K-12, except Grade 4							
	-.693 -.479 -.322 -.078 .010 .052	.113 .127	-.547 -.506 -.448 -.404 -.386 -.349	-.232 -.146 -.113 -.095 -.080 .048	.055 .085 .088 .090 .091 .099	.134 .141 .150 .176 .212 .251	.263 .299 .319 .346 .357 .364	.500 .604 .650	
Results of Test for Homogeneity	$H_T = 18.48$ $\chi^2_7 = 14.07$ Heterogeneous $\bar{g}^W = -.181$	$H_T = 45.40$ $\chi^2_{32} = 46.18$ Homogeneous $\bar{g}^W = .124$ (.06, .19)*							

°Indicates outlier excluded

^dIndicates previously dependent data*Indicates rejection of
null hypothesis

Figure 15. Analysis summary: Maintenance effects regarding achievement in acquiring computational skills.

Top Level Effect Sizes g	-.227 .219 .400 .485 .558 .640 .795 1.337 0 .288 .415 .498 .581 .651 .827 1.411 ⁰ .069 .295 .443 .535 .600 .709 .956 1.675 ⁰ .081 .312 .477 .557 .629 .734 1.170
Box-and-Whiskers	
Results of Test for Homogeneity	$H_T = 37.06$ $\chi^2_{28} = 41.34$ Homogeneous $\bar{g}^W = .636$ (.55, .73)*

*Indicates rejection of null hypothesis

Figure 16. Analysis summary: Extension effects regarding achievement in acquiring computational skills.

b. What are the effects regarding acquisition of conceptual skills in the maintenance sense? In the sense of extension?

With regard to maintenance, 39 effect magnitudes were available to address the question; these are listed at the top of Figure 17, the analysis summary. With the single outlier removed, the group was found homogeneous ($H_T = 51.02 < 52.18$), with a nonsignificant mean effect of .018.

Regarding extension effects for conceptual skills, only one effect size was available; thus, the question could not be addressed.

RQ-2. What are the calculator effects regarding acquisition of composite problem solving skills in the maintenance sense? In the sense of extension?

Sixty-two effect magnitudes initially were available to address the maintenance part of this question. These effects are listed at the top of Figure 18, the analysis summary. The box-and-whiskers plot reveals a fairly normal distribution, but with the single outlier removed, the group remained heterogeneous ($H_T = 100.74 > 79.08$). In 2nd Stage analysis, a significant relationship was found between effect size and student ability levels "low," "general," and "high." The groups for low and high abilities each were found homogeneous, with respective mean effects of .005 and -.118. Neither value was significantly different from zero, since the 95 percent CI's contained the null. The grouping for general student ability was heterogeneous. In 3rd Stage analysis, this effect size group was found related to school grade level. When effects for Grades 4 and 7 were combined,

Top Level Effect Sizes g	-.705 -.301 -.094 -.022 .092 .193 .278 .409 -.631 -.233 -.090 -.010 .137 .193 .281 .475 -.499 -.218 -.066 0 .141 .226 .282 .488 -.465 -.213 -.058 .009 .156 .237 .299 .948° -.361 -.196 -.056 .009 .182 .264 .341
Box-and-Whiskers	
Results of Test for Homogeneity	$H_T = 51.02$ $\chi^2_{37} = 52.18$ Homogeneous $\bar{g}^W = .018$ (-.05, .09)

°Indicates outlier excluded

Figure 17. Analysis summary: Maintenance effects regarding achievement in acquiring conceptual skills.

Top Level Effect Sizes g	-1.028°	-.282	-.088	0	.054	.121	.185	.364	.558
	-.476	-.230	-.058	0	.074	.135	.229	.364	.679
	-.415	-.226	-.055	0	.086	.140	.248	.395	.849
	-.408	-.179	-.045	.007	.098	.160	.314	.411	.904
	-.328	-.128	-.043	.034	.106	.162	.337	.423	.948
	-.307	-.125	-.040	.052	.110	.173	.341	.490	.974
	-.285	-.108	-.037	.053	.115	.180	.359	.523	
Box-and-Whiskers									
Results of Test for Homogeneity	$H_T = 100.74$ $\chi^2_{60} = 79.08$ Heterogeneous								
2nd Stage g's by Ability Level	Low		General						High
	-.328		-.476 -.058 .034 .129 ^d .341 .523						-.408
	-.179		-.415 -.055 .052 .140 .359 .558						-.226
	-.128		-.307 -.043 .074 .160 .364 .679						-.108
	-.045		-.285 -.040 .086 .162 .364 .849						-.037
	.053		-.282 -.023 ^d .098 .173 .386 ^d .904						.007
	.054		-.267 ^d 0 .106 .229 .395 .948						.185
	.135		-.230 0 .110 .248 .411						
	.180		-.125 0 .115 .314 .423						
	.974°		-.088 .025 ^d .121 .337 .490						
Results of Test for Homogeneity	$H_T = 2.05$ $\chi^2_7 = 14.07$ Homogeneous $\bar{g}^W = .005$ (-.20, .21)		$H_T = 82.58$ $\chi^2_{50} = 67.50$ Heterogeneous				$H_T = 1.86$ $\chi^2_5 = 11.07$ Homogeneous $\bar{g}^W = -.118$ (-.33, .09)		
3rd Stage g's, General Ability by School Grade Level	Grades 4 and 7		Grades K-12, except Grades 4 and 7						
	-.476 0 .160		-.307 -.058 .074 .229 .386 .558						
	-.415 .025 .162		-.285 -.040 .106 .248 .395 .849						
	-.125 .086 .341		-.282 0 .110 .314 .411 .904						
	-.055 .098 .363		-.267 0 .129 .337 .423 .948°						
	-.043 .115 .679		-.230 .034 .140 .359 .490						
	-.023 .121		-.088 .052 .173 .364 .523						
Results of Test for Homogeneity	$H_T = 25.65$ $\chi^2_{16} = 26.30$ Homogeneous $\bar{g}^W = .075$ (.008, .14)*		$H_T = 40.79$ $\chi^2_{32} = 46.18$ Homogeneous $\bar{g}^W = .203$ (.13, .28)*						

°Indicates outlier excluded

^dIndicates previously dependent data

*Indicates rejection of null hypothesis

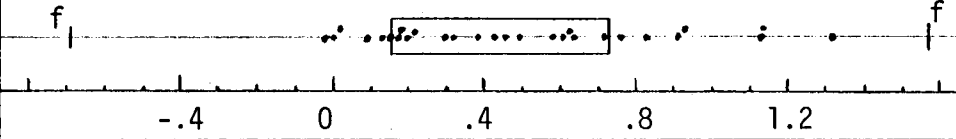
Figure 18. Analysis summary: Maintenance effects regarding achievement in acquiring composite problem solving skills.

the resulting subset was homogeneous, with a significant mean effect of .075. The data set for remaining grades was also homogeneous, with a significant mean effect of .203.

Regarding extension effects for composite problem solving skills, 28 effect magnitudes initially were available to address the question. These effects are listed at the top of Figure 19, the analysis summary. This group was determined as heterogeneous ($H_T = 56.96 > 40.11$), and the 2nd Stage analysis showed a significant relationship between effect size and student ability level. Effect size groups for low and high abilities each were found homogeneous, with respective means of .436 and .458, both of which are significantly different from zero. For general students, the effect size group was heterogeneous, and no significant moderating variable could be determined. Its mean of .271 is quoted as a descriptive statistic.

The basic RQ-2 has considered calculator effects upon composite problem solving skills. There exist two related areas which the calculator may potentially impact: problem solving productivity, defined as the relative numbers of problems attempted by calculator and non-calculator groups within a testing period; and, problem solving selectivity, defined as the relative numbers of proper strategies and/or processes used by C and NC groups for a selected set of problems. To deal with these interests, RQ-2 contains subquestions which will now be regarded in turn.

- a. What are the effects regarding acquisition of problem solving productivity skills in the maintenance sense? In the sense of extension?

Top Level Effect	-.016	.128	.190	.379	.581	.701	.915
Sizes g	0	.137	.218	.424	.603	.758	1.124
	.022	.156	.297	.451	.622	.828	1.135
	.093	.168	.317	.499	.644	.906	1.313
Box-and-Whiskers							
Results of Test for Homogeneity	$H_T = 56.96$ $\text{Chi}^2_{27} = 40.11$ Heterogeneous						
2nd Stage g's by Ability Level	Low	General					High
	.022	-.016	.218	.581	1.124	.190	
	.168	0	.263 ^d	.603 ^d	1.313	.317	
	.499	.093	.297	.803 ^d		.622	
	.701	.128	.379	.828 ^d		.644	
	.758	.137	.424	.878 ^d		.906	
	1.135	.156	.451	.915			
Results of Test for Homogeneity	$H_T = 10.02$ $\text{Chi}^2_5 = 11.07$ Homogeneous $\bar{g}^W = .436$ (.21, .66)*	$H_T = 44.94$ $\text{Chi}^2_{19} = 30.14$ Heterogeneous $\bar{g}^W = .271$			$H_T = 2.66$ $\text{Chi}^2_4 = 9.49$ Homogeneous $\bar{g}^W = .458$ (.17, .75)*		

°Indicates outlier excluded
^dIndicates previously dependent data

*Indicates rejection of null hypothesis

Figure 19. Analysis summary: Extension effects regarding achievement in acquiring composite problem solving skills.

The maintenance part of the question could not be addressed, since only one effect magnitude existed in its regard. Concerning the extension effect, seven effect magnitudes were available; these are listed in Figure 20, the analysis summary. The group was heterogeneous ($H_T = 17.41 > 12.59$), and no moderating variable to explain the heterogeneity could be discovered. The mean of .139 is offered as a descriptive statistic.

- b. What are the effects regarding acquisition of problem solving selectivity skills in the maintenance sense? In the sense of extension?

Six effect magnitudes regarding maintenance effects are listed in Figure 21, the analysis summary. With so little data, a box-and-whiskers plot would not be definitive; hence, it was omitted. The effect size group was found homogeneous ($H_T = 3.11 < 11.07$), but its mean effect of $-.209$ was not significant, since the 95 percent CI contained zero.

With regard to extension effects, 13 effect magnitudes were available to address the question. These effects are listed in Figure 22, the analysis summary. The group was found homogeneous ($H_T = 20.22 < 21.03$), with a significant mean effect of .328.

The foregoing RQ-1 and RQ-2 have considered achievement in acquiring operational and problem solving skills. Because of the relative abundance of data related to those questions, each of their sub-levels could be separately addressed. Forthcoming questions regard achievement in retaining and transferring analogous skills. With much less data existing in those areas, each question is addressed in total.

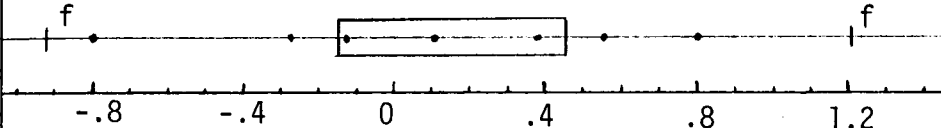
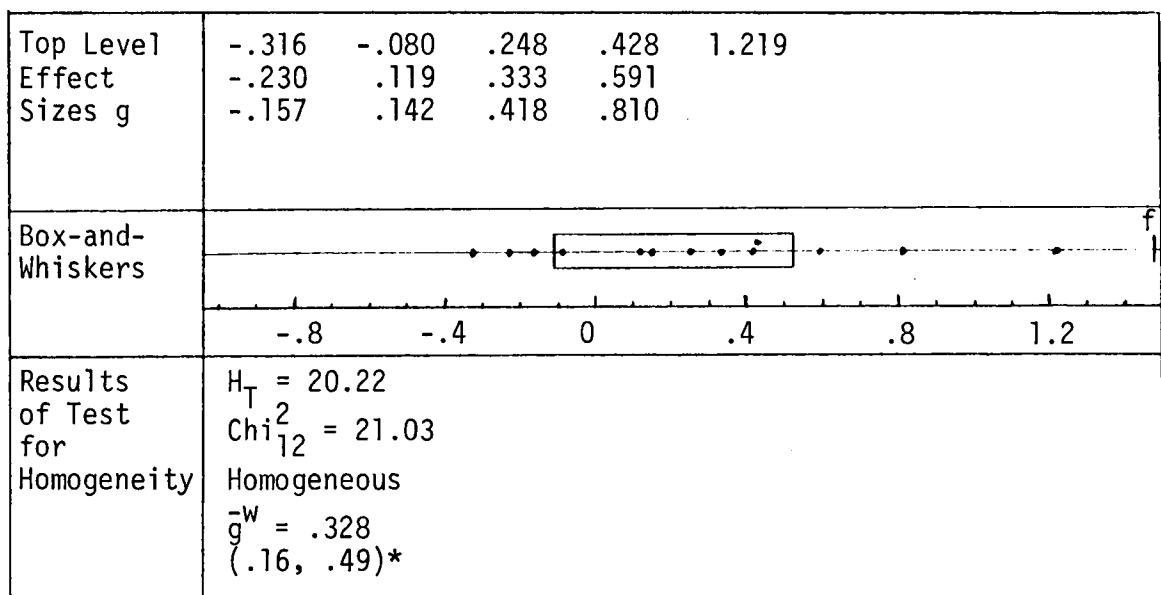
Top Level Effect Sizes	-.801 .102 .811 -.276 .394 -.124 .560
Box-and- Whiskers	
Results of Test for Homogeneity	$H_T = 17.41$ $\chi^2_6 = 12.59$ Heterogeneous $\bar{g}^W = .139$

Figure 20. Analysis summary: Extension effects regarding achievement in acquiring problem solving productivity.

Top Level Effect Sizes g	-.446 -.134 -.415 .183 -.221 .183
Results of Test for Homogeneity	$H_T = 3.11$ $\chi^2_5 = 11.07$ Homogeneous $\bar{g}^W = -.209$ $(-.49, .07)$

Figure 21. Analysis summary: Maintenance effects regarding achievement in acquiring problem solving selectivity.



*Indicates rejection of
null hypothesis

Figure 22. Analysis summary: Extension effects regarding achievement in acquiring problem solving selectivity.

RQ-3. What are the analogous effects regarding retention of operational skills in the maintenance sense? In the sense of extension?

Figure 23 compiles analyses related to these questions. With regard to retaining composite operational skills in the maintenance sense, 12 effects (excluding one outlier) were found homogeneous, with a nonsignificant mean effect of $-.016$. In the sense of extension, five effect magnitudes related to composite operational skills were heterogeneous, with a mean of $.479$ provided as a descriptive statistic. Group effects regarding the retention of computational and conceptual skills in the maintenance sense were each found heterogeneous, with respective means of $-.073$ and $.196$. No moderating variables to explain the heterogeneity could be discovered for any of these heterogeneous groups. Only two effects regarding computation and no effects regarding concepts were available in the sense of extension; hence, those parts of RQ-3 could not be addressed.

RQ-4. What are the analogous effects regarding retention of problem solving skills in the maintenance sense? In the sense of extension?

Because of insufficient data, only one effect could be addressed. Regarding the retention of composite problem solving skills in the sense of extension, three effect magnitudes were available; these are shown in Figure 24, the analysis summary. This small group was homogeneous, with a significant mean effect of $.371$.

Composite Operational Skills (Maintenance)

Effect	-.545	-.448	.073	.149	.963°
Sizes g	-.535	-.165	.107	.180	
	-.530	-.124	.115	.432	
Box-and-Whiskers					
Results of Test for Homogeneity	$H_T = 16.39$ $\chi^2_{11} = 19.68$ Homogeneous $\bar{g}^W = -.016$ $(-.20, .17)$				

Composite Operational Skills (Extension)

Effect	-.141	.905
Sizes g	0	1.905
	.128	
Results of Test for Homogeneity	$H_T = 27.20$ $\chi^2_4 = 9.49$ Heterogeneous $\bar{g}^W = .479$	

Computational Skills (Maintenance)

Conceptual Skills (Maintenance)

Effect	-.698	.274	-.268
Sizes g	-.549	.718	.118
	-.234		.861
Results of Test for Homogeneity	$H_T = 22.65$ $\chi^2_4 = 9.49$ Heterogeneous $\bar{g}^W = -.073$		$H_T = 8.99$ $\chi^2_2 = 5.99$ Heterogeneous $\bar{g}^W = .196$

°Indicates outlier excluded

Figure 23. Analysis summaries: Effects regarding achievement in retaining operational skills.

Top Level Effect Sizes g	.220 .230 .622
Results of Test for Homogeneity	$H_T = 1.98$ $\chi^2_2 = 5.99$ Homogeneous $\bar{g}^W = .371$ $(.11, .64)^*$

*Indicates rejection of
null hypothesis

Figure 24. Analysis summary: Extension effects regarding achievement in retaining composite problem solving skills.

RQ-5. What are the analogous effects regarding transfer of operational skills in the maintenance sense? In the sense of extension?

Because of insufficient data, only one effect could be addressed. Regarding the transfer of composite operational skills in the maintenance sense, four effect magnitudes were available; these are shown in Figure 25, the analysis summary. This group was found homogeneous, with a nonsignificant mean effect of .100.

RQ-6. What are the analogous effects regarding transfer of problem solving skills in the maintenance sense? In the sense of extension?

Because of insufficient data, no effects related to this question could be addressed.

RQ-7. What are the calculator effects on student estimation skills?

As shown in Figure 26, five effect magnitudes were available to address this question. The group was found homogeneous, with a nonsignificant mean effect of .082.

RQ-8. What are the calculator effects on attitude toward (or enjoyment of) mathematics?

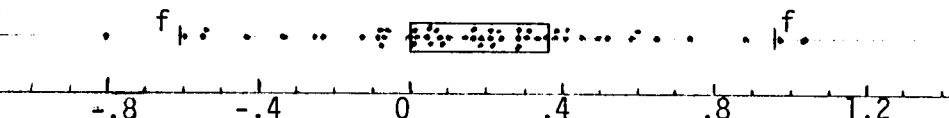
Fifty-six effect magnitudes were available to address this question; these are listed in Figure 27, the analysis summary. After exclusion of three outliers, the remaining group was found homogeneous ($H_T = 69.03 < 69.82$), with a significant mean effect of .190.

Top Level Effect Sizes g	$-.315$ $.530$ $-.127$ $.239$
Results of Test for Homogeneity	$H_T = 3.61$ $\chi^2_3 = 7.82$ Homogeneous $\bar{g}^w = .100$ $(-.24, .44)$

Figure 25. Analysis summary: Maintenance effects regarding achievement in transferring composite operational skills.

Top Level Effect Sizes g	$-.046$ $.562$ $-.027$ $.609$ $.031$
Results of Test for Homogeneity	$H_T = 6.83$ $\chi^2_4 = 9.49$ Homogeneous $\bar{g}^w = .082$ $(-.10, .26)$

Figure 26. Analysis summary: Effects regarding achievement in acquiring estimation skills.

Top Level Effect Sizes g	-.802°	-.229	-.020	.043	.184	.276	.383	.574
	-.605	-.116	-.014	.068	.186	.279	.392	.590
	-.555	-.077	-.002	.082	.205	.284	.411	.637
	-.539	-.071	0	.095	.214	.299	.421	.732
	-.430	-.067	.038	.144	.226	.306	.447	.879
	-.330	-.062	.039	.163	.227	.333	.490	.963°
	-.259	-.055	.042	.165	.239	.346	.514	1.033°
Box-and-Whiskers								
Results of Test for Homogeneity	$H_T = 69.03$ $\chi^2_{52} = 69.82$ Homogeneous $\bar{g}^w = .190$ $(.13, .25)^*$							

°Indicates outlier excluded

*Indicates rejection of
null hypothesis

Figure 27. Analysis summary: Effects regarding attitude toward mathematics.

RQ-9. What are the effects upon anxiety toward mathematics?

Only four effect magnitudes were available to address this question. However, as shown in Figure 28, the group was found homogeneous, with a nonsignificant mean of $-.037$.

RQ-10. What are the effects upon self-concept in mathematics?

As shown in Figure 29, this question was addressed by five effects found homogeneous, with a significant mean effect of $.305$.

RQ-11, RQ-12, and RQ-13 inquire about effects on motivation to increase mathematical knowledge, attitude toward mathematics teachers, and perception of the value of mathematics in society. Because of insufficient data, none of these questions could be addressed.

All of the foregoing questions were examined by inferential statistics. The following questions are nonstatistical and will be provided descriptive answers based on observation of the data.

RQ-14. Do effects on achievement and attitude seem to differ for male and female students?

Thirteen studies investigated sex-related differences in calculator effects (see Part 1 of Appendix B). Eight studies found no significant differences, two had results that favored males, and three had results that favored females. On balance, these findings suggest that calculator effects do not significantly differ for male and female students.

RQ-15. What effects on achievement and attitude were observed as a result of special calculator instruction?

Figure 30 lists effect sizes g and corresponding mean effects related to dependent variables for which effect size data were

Top Level Effect Sizes g	$-.339$ $.234$ $-.087$ $.268$
Results of Test for Homogeneity	$H_T = 5.36$ $\chi^2_3 = 7.82$ Homogeneous $\bar{g}^W = -.037$ $(-.25, .18)$

Figure 28. Analysis summary: Effect regarding anxiety toward mathematics.

Top Level Effect Sizes g	$.040$ $.328$ $.152$ $.568$ $.152$
Results of Test for Homogeneity	$H_T = 3.57$ $\chi^2_4 = 9.49$ Homogeneous $\bar{g}^W = .305$ $(.09, .52)^*$

*Indicates rejection of
null hypothesis

Figure 29. Analysis summary: Effect regarding self-concept in mathematics.

Dependent Variable	Effect Sizes g				Mean Effect $\bar{g}_w$
Maintenance of Composite Operational Skills	-.284 -.174 -.144	.090 .831 .952	.987 1.246 1.280		.798
Extension of Composite Operational Skills	.083 .323	1.085 1.677			.747
Maintenance of Computational Skills	-.227 .009	.255 .274	.290 1.061	1.297	.564
Maintenance of Conceptual Skills	-.356 -.341	-.095			-.268
Maintenance of Problem Solving Skills	-.324 -.177 0	0 .267 .420	.477 .633 .770	.991	.534
Attitude toward Mathematics	-.365 0 .016 .035 .045	.174 .233 .237 .263 .398	.528 .782 .815 .953 1.121	1.230 1.260 1.298	.435

Figure 30. Effects with special calculator instruction: Achievement in acquiring skills; and, attitude toward mathematics.

available. Primary attention had been paid to maintenance of paper-and-pencil skills in composite operations, computation, concepts, and problem solving. All these effects seem large (respectively, .798, .564, $-.268$, and .534), except regarding concepts where the finding may be suspect because of limited data. One extension effect was recorded, whose value of .747 appears sufficiently large to be significant. With regard to attitude toward mathematics, a substantial effect of .435 was observed.

Summary and Discussion

This section of "Results" has described analytic procedures and findings in response to 15 questions of research. Findings now are summarized and discussed in preparation for drawing conclusions.

The first six questions inquired of achievement in skills acquisition, retention, and transfer with regard to basic operations and problem solving in the senses of maintenance and extension. Results concerning these questions are collected into Figure 31, whose elements are directly the observed effects of calculators. Effects determined as significantly different from zero are identified by stars (e.g., .325*). Unstarred effects are considered nonsignificant. Effects related to heterogeneous data have been labeled descriptive statistics by way of a superscript "ds" (e.g., .479^{ds}).

1. Effects regarding skills acquisition. RQ-1 and RQ-2 concerned the acquisition of skills in basic operations and problem solving.

Effects in Basic Operations

Skills	Composite		Computation		Concepts	
	Maint.	Ext.	Maint.	Ext.	Maint.	Ext.
Acquisition (RQ-1)	Ability Low: -.107 High: -.031 General (by grade): 4: -.152* Other: .137*	Ability Low: .325* General: .737 ^{ds}	Ability Low: -.009 High: .024 General (by grade): 4: -.181 ^{ds} Other: .124*	.636*	.018	-
Retention (RQ-3)	-.016	.479 ^{ds}	-.073 ^{ds}	-	.196 ^{ds}	-
Transfer (RQ-5)	.100	-	-	-	-	-

Effects in Problem Solving

Skills	Composite		Productivity		Selectivity	
	Maint.	Ext.	Maint.	Ext.	Maint.	Ext.
Acquisition (RQ-2)	Ability Low: .005 High: -.118 General (by grade): 4 & 7: .075* Other: .203*	Ability Low: .436* High: .458* General: .271 ^{ds}	-	.139 ^{ds}	-.209	.328*
Retention (RQ-4)	-	.371*	-	-	-	-
Transfer (RQ-6)	-	-	-	-	-	-

*Indicates a value significantly different from zero

^{ds} Indicates a descriptive statistic

Figure 31. Summary of achievement effects regarding skills acquisition, retention, and transfer.

a. Maintenance effects. Regarding composite operational skills, nonsignificant effects existed for low and high ability students, while general students (in regular classes) produced significant effects related to school grade level, i.e., $-.152$ for Grade 4 and $.137$ for the other grades combined. Hence, paper-and-pencil skills of low and high ability students in the calculator groups remained at par with basic skills of corresponding students in the control groups. In mixed-ability classes, paper-and-pencil skills significantly improved from calculator treatment, except in Grade 4 where basic skills significantly suffered.

Parallel results were observed with reference to the maintenance of computational skills. Again, the paper-and-pencil achievement of low and high ability students did not change as a result of calculator treatment, but basic skills improved in general classes (effect = $.124^*$), except in Grade 4 (descriptive effect = $-.181$).

With regard to maintenance of concepts, a nonsignificant effect of $.018$ was observed across abilities and grades. Since most calculator treatment had been functional rather than pedagogical, this result may well have been foreseen. Indeed, the absence of effect explains the parallel results for composite operational and computational skills.

The foregoing effects are consistent with the observations of previous reviews, e.g., calculator usage yielded achievement as high or higher than when calculators were not used (Suydam, 1982a, p. 3); and, concept acquisition from the use of calculators appeared to be minimal (Roberts, 1980).

Effects regarding the maintenance of composite problem solving skills continued the pattern of basic operations. Low and high ability students neither profited nor suffered from the calculator treatment; paper-and-pencil skills of general students were significantly improved, although considerably less in Grades 4 and 7 than other grade levels (.075 versus .203). Nonsignificant results attended problem solving selectivity. When experimental and control groups had each worked with paper and pencil, no difference was observed in the selection of correct processes.

The foregoing positive effects appear surprising. Since problem solving strategies had in general not been taught within the calculator treatment, no reason seems apparent why the treatment should have improved problem solving skills with paper and pencil. Perhaps the effects were more affective than cognitive, say by Hawthorne or by novelty.

b. Extension effects. Confirming the expectation that a use of calculators on tests will improve student scores, effects for basic operations and composite problem solving were consistently large and positive across grade levels. Again, relationships existed between effect size and student ability level. In basic operations, the scores of low ability students seemed not so sharply increased as scores for general students (.325* versus .737^{ds}). Throughout the analysis, it has appeared that calculator use has served the low and high ability students less well than the general student. However, for problem solving with the calculator permitted, the situation seems reversed.

Low and high ability effects (.436 and .458) appeared significantly higher than the descriptive effect (.271) for general students.

With regard to problem solving productivity and selectivity, theory had suggested that the use of calculators in testing would speed the effort of computation, thus allowing the students time to work more problems and/or choose more proper processes. Analysis results for productivity were not conclusive (though a trend perhaps existed toward solution of more problems by the calculator groups). The extension effect for selectivity was fairly large and positive (.328*). Hence, the calculator's use in problem solving created not only a computational advantage but also a benefit (probably time) in choosing proper approaches to solutions.

2. Effects regarding skills retention. RQ-3 and RQ-4 concerned the retention of skills in basic operations and problem solving. In the maintenance sense, only nonsignificant or small descriptive effects existed for composite operations, computation, and concepts. Hence, the retention of basic skills seemed neither damaged nor improved by calculator treatment. In the sense of extension, predictably large effects existed for composite operations (.479^{ds}) and for problem solving (.371*). Thus, in retention posttesting, students markedly improved achievement scores through calculator usage.

3. Effects regarding skills transfer. RQ-5 and RQ-6 concerned the transfer of skills in basic operations and problem solving to kindred mathematics. Only one effect was determined, the maintenance effect on basic skills within composite operations; its value of .100 was

not significant. Thus, no difference existed between calculator and noncalculator groups with regard to operational skills transfer.

RQ-7 was also achievement-related, inquiring of effects on student estimation skills. A nonsignificant effect of .082 was observed, so calculator treatment neither damaged nor improved such skills.

RQ-8 and RQ-13 regarded effects on student attitudes. Values were observed as follows:

Attitude toward mathematics: .190*;

Anxiety toward mathematics: -.037 (not significant);

Self-concept in mathematics: .305*.

Calculator treatments thus significantly improved student attitudes toward mathematics and self-concept in mathematics across all ranges of study characteristics.

RQ-14 and RQ-15 were nonstatistical questions regarding: sex-related differences in calculator effects; and, effects resulting from special calculator instruction. Effects did not appear to differ for male and female students. In studies with materials/methods especially developed for instruction with the calculator, descriptive effects in general exceeded corresponding effects within traditional instruction, e.g., .798 versus .137 for maintenance of composite operational skills, .534 versus .203 for maintenance of problem solving skills, and .435 versus .190 for attitude toward mathematics (see Figures 30 and 31, pp. 170 and 172). Hence, special calculator curricula enhanced both achievement and attitude beyond the gains obtained within a traditional format.

Conclusions and Recommendations

This investigation has determined student effects of calculator usage by statistically averaging subsets of the 529 effect magnitudes derived from 79 studies. Because this body of studies was exhaustively collected, it was considered a representative and probabilistic sample for meta-analysis. Hence, conclusions from the findings have been generalized as follows:

1. In Grades K-12 (except Grade 4), students who use calculators in concert with traditional instruction will maintain their paper-and-pencil skills without apparent damage. Indeed, a use of calculators can improve the average student's basic skills with paper and pencil, both in basic operations and in problem solving.

2. Sustained calculator usage in Grade 4 will be counter-productive with regard to students' basic skills.

3. The use of calculators in testing will produce much higher achievement scores than paper-and-pencil efforts, both in basic operations and in problem solving. This statement applies across all grades and ability levels. In particular, it applies for low and high ability students in problem solving. The overall better performance in solving word problems is a result of improved computation and process selection.

4. Students using calculators will possess a better attitude toward mathematics than noncalculator students. This statement applies across all grades and ability levels.

5. Students using calculators will possess an especially better self-concept in mathematics than noncalculator students. This statement applies across all grades and ability levels.

6. Studies with special curricula have conclusively demonstrated that materials/methods can be developed for enhancing student achievement through teaching and learning concepts with the calculator. However, instruction oriented toward the calculator has been relatively unexamined by research.

That calculators stand to benefit student achievement and attitude seems abundantly clear. It no longer seems a question if or should they not be used along with basic skills instruction, but how. In this regard, proposals for action are offered in the areas of classroom usage, curriculum development, and future research.

Recommendations for Classroom Usage

1. Calculators should be used in all mathematics classes of Grades K-12, with levels of usage increasing as grade level increases. Teachers should prepare themselves for calculator instruction through self-training and inservice programs.

2. Due to limited research for grades K-3 and to the apparent negative effect of calculator treatments in Grade 4, the use of calculators in those grades should be restricted to familiarization, recreation, and perhaps occasional drill and problem solving.

3. Students in Grade 5 and beyond should be permitted calculator use in all problem solving activities, including testing situations. This recommendation is based on four observations:

- a. Calculators greatly benefit student achievement in problem solving, especially for low and high ability students;
- b. Positive attitudes related to the use of calculators may help to relieve students' dread of word problems;
- c. Developing computational skill is not the central purpose of problem solving;
- d. More instructional time could be put to problem solving strategies, and the problems could be more relevant and realistic.

Whether to allow calculators for problem solving portions of tests will no doubt remain controversial. However, their use for instruction but exclusion during tests presents what Roberts (1980) has called a "negative orientation" to the use of calculators (p. 94). Hence, their use in tests of problem solving is recommended.

Recommendations for Curriculum Development

It hardly needs to be noted that the current mathematics curriculum is not designed for calculators. Textbooks have provided calculator exercises, thus acknowledging its existence, but teachers by and large have had to break new ground when interacting calculators with students learning mathematics. This requirement helps explain why the calculator has seen so little service in the schools since its advent during the mid-1970s. That calculators may exert significant impact upon the curriculum seems well accepted. However, the direction and extent of that influence has remained essentially unexplored.

Analysis of curriculum revision and redesign appears essential. Specifically:

1. Curriculum developers should determine how the calculator can be optimally absorbed within the existing curriculum. Teachers should strongly assert their contributions to this effort.

2. Curriculum developers should also determine where the existing curriculum will most likely need revision to accommodate optimal calculator usage. These findings will need to be verified or rejected by research; hence, they should be communicated in concert with a plan for systematically performing that research. As Suydam (1976) has noted, curriculum restructure and research "are integrally interwoven, and one cannot be effective without the other" (p. 38).

Recommendations for Future Research

Comparisons of calculator treatment and traditional instruction without calculators seem sufficiently investigated. Future research should be directed toward pedagogical uses and toward comparisons of methods where calculators are included in all treatments. [For example, Turinese (1982) has compared the achievement of second year algebra students studying sequences and series in a calculator oriented curriculum and a textbook oriented program, where both groups used calculators in posttesting.] In her report of 1976, Suydam suggested examination of the following research topics. They still apply.

- a. Effective procedures for calculator use in learning basic facts, computational skills, and problem solving strategies;

- b. Effects of calculator use with specific mathematical subject matters in grades K-12;
- c. Effects of curriculum emphasis changes;
- d. Effects of curriculum sequence changes;
- e. Teacher education to incorporate the calculator;
- f. Optimal calculator designs for use in the classroom.

As calculator specifics evolve, resulting methods and materials should be expedited into the mathematics curriculum.

Grade 4 was the only grade level where negative effects of calculators were observed. This phenomenon was not foreseen. Its cause may only be surmised. Perhaps the subject matter in Grade 4 is not amenable to calculator treatment. Grade 4 marks the beginning of the intermediate levels, and pupils may have just proceeded from the oldest in their environment to the youngest in new surroundings; calculator treatment may thus have provoked additional trauma. The apparent decline of basic skills perhaps is related to brain growth "periodization." At the ages of most Grade 4 pupils, the human brain is said to experience a plateau period of relatively little growth (Toepfer, 1979); hence, in the cognitive domain, calculator procedures may have subverted basic skills. Whatever the cause, the negative effects observed in Grade 4 should be accorded research attention, that such effects may be avoided or redressed.

These recommendations conclude the report of hand-held calculator effects within mathematics education. Clearly, their implementation regards no short term, straight line endeavor, but benefits to student

achievement and attitude surely must render the game worth the candle. The calculator's societal impact may then be felt within the classroom, as well as without.

Evaluation

The model for meta-analysis as developed in Chapter III provides a set of guidelines for report evaluation (Figure 9, p. 101). As shown by Figure 32, the present report complies with all criteria specified for this evaluation.

Intro- duction	1. Is the general topic fully described?	✓
	2. Is prior evidence (e.g., reviews) critically examined?	✓
	3. Are research questions clearly stated? Comprehensive?	✓
	4. Are dependent variables operationally defined?	✓
Method	5. Are the target populations specified?	✓
	6. Is the search for available studies fully reported?	✓
	Are guidelines specified for excluding studies?	✓
	Do the studies represent the full body of research?	✓
	Is a bibliography of studies made available?	✓
	7. Is the list of study characteristics comprehensive?	✓
	Are study characteristics clearly coded?	✓
	Is research design quality rated for each study?	✓
	Are criteria reported for such ratings? Appropriate?	✓
	Are there at least a few good designs?	✓
	Are forms of publication entered as an independent variable?	✓
	Are types of measuring instruments recorded?	✓
	8. Are procedures described for deriving effect magnitudes or correlations?	✓
	Are all effect magnitudes or correlations listed in both size and direction?	✓
Results	9. Were strong attempts made to locate missing data?	✓
	10. Does analysis account for sample size variations?	✓
	Does analysis account for data dependence?	✓
	11. Is separate analysis performed for different constructs?	✓
	12. Are methods for data crosscutting clearly described?	✓
	13. Are appropriate techniques described for deriving group means and/or relationships?	✓
	14. Are inferential statistics applied to the data?	✓
	Are inferential procedures described? Appropriate?	✓
Conclusions /Discussion	15. Are the effects of missing data, poor research designs, or other biasing effects described and compensated?	✓
	16. Are results of analysis clearly and fully presented?	✓
	17. Are the accessible populations described?	✓
	Are inferences made to those populations?	✓
	18. Are conclusions clearly related to the research questions?	✓
	19. Does the meta-analysis determine what is known/not known about the topic?	✓
	20. Does the report suggest areas in need of further inquiry?	✓
	Does the report provide enough information such that a reader may critically examine the evidence?	✓

Figure 32. Checklist for evaluating report of meta-analysis on effects of hand-held calculators.

CHAPTER V

LOOKING BACK AND LOOKING FORWARD

For decades, scientists in education have examined classroom behaviors. The growth of this research appears remarkable in two regards, exponential rather than linear, but lateral rather than vertical. Though abundant, such sprawling research may offer confusion, not clarification; findings that conflict become harder to reconcile, and the research's sheer mass may block the synthesis of those results by traditional methods. The observation has substance that, within the social sciences, the most important past problem has been the failure to produce cumulative knowledge (Schmidt, 1980).

The present study was designed to help relieve the problem by achieving the following purposes:

- Develop a model for meta-analysis of research in education;
- Derive explicit guidelines for reporting and evaluating reports of meta-analysis;
- Demonstrate the model by performing and reporting the meta-analysis of significant research within mathematics education.

These purposes have been accomplished. Hence, the effectiveness of the model in at least the single study can be judged.

1. The model appears methodologically sound and capable of generating valid inferential statistics. However, results may be threatened by data dependences, lack of a probabilistic sample of studies, and failure to resolve heterogeneous findings. Pains to guard

against such threats will counter their impact. External problems attend the process through data omitted from reports, hardship in obtaining data from authors, and difficulty in securing fugitive research. Authors (and editors) can make their reports more useable in research integration by assuring, at the least, that means and standard deviations for all variables and sample sizes for all groups are carefully reported.

2. Requirements for the body of studies are minimal and consistent with research in education, i.e., provision of sufficient data for effect size calculations, study outcomes based on continuous data, and sample sizes of 10 or larger.

3. The model provides for appraisal of the integrative effort by way of the evaluative guidelines.

4. Procedures appear sufficiently flexible for broad application in educational research.

The model thus seems a tool of value in the synthesis of studies, an area of "major growth" in future research (Pillemer and Light, 1980, p. 194). If through its service researchers are tempted more to the task of research integration, the model will have satisfied its purpose: to help advance the science of discovering what we know.

REFERENCES

REFERENCES

- Abrami, P. G., Leventhal, L., and Perry, R. P. (1982). Educational seduction. Review of Educational Research, 52, 446-464.
- Athappilly, K. K. (1978). A meta-analysis of the effects of modern mathematics in comparison with traditional mathematics in the American educational system. Dissertation Abstracts International, 39, 6002A.
- Baker, P. C. (1952). Combining tests of significance in cross validation. Educational and Psychological Measurement, 12, 300-306.
- Bandura, A. (1978). On paradigms and recycled ideologies. Cognitive Therapy and Research, 2, 79-103.
- Begle, E. G. (1979). Critical variables in mathematics education. Washington, D.C.: Mathematical Association of America and the National Council of Teachers of Mathematics.
- Burns, P. K., and Bozeman, W. C. (1981, October). Computer-assisted instruction and mathematics achievement: Is there a relationship? Educational Technology, pp. 32-39.
- Cahen, L. S. (1980). Meta-analysis: A technique with promise and problems. In H. J. Walberg and E. H. Haertel (Eds.), Evaluation in Education (Vol. 4, pp. 37-39). Oxford, Great Britain: Pergamon Press.
- Capoferi, A., and Winowski, E. (1975). Evaluation of pilot studies of classroom use of the hand calculator in Grades 4-6. Mt. Clemens, MI: Fitzgerald and Van Dyke Public Schools Cooperative Project.
- Cohen, J. (1977). Statistical power analysis for the behavioral sciences (revised edition). New York: Academic Press.
- Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. Review of Educational Research, 51, 281-309.
- 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. (1982). Scientific guidelines for conducting integrative reviews. Review of Educational Research, 52, 291-302.

- Cooper, H. M., and Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. Psychological Bulletin, 87, 442-449.
- Cronbach, L. J. (1980). Toward reform of program evaluation. San Francisco: Jossey-Bass.
- DerSimonian, R., and Laird, N. M. (1983). Evaluating the effect of coaching on SAT scores: A meta-analysis. Harvard Educational Review, 53, 1-15.
- Dessart, D. J., and Suydam, M. N. (1983). Classroom ideas from research on secondary school mathematics. Reston, VA: National Council of Teachers of Mathematics.
- Driscoll, M. J. (1979). Elementary school mathematics annotated bibliography. St. Louis, MO: Central Midwestern Regional Educational Laboratory, Inc. (ERIC Document Reproduction Service No. ED 183 378)
- East, P., and Moursund, D. (1979). Calculators and computers in the classroom. Salem, OR: Oregon State Department of Education. (ERIC Document Reproduction Service No. ED 174 456)
- El-Nemr, M. A. (1980). Meta-analysis of the outcomes of teaching biology as inquiry (Doctoral dissertation, University of Colorado, 1979). Dissertation Abstracts International, 40, 5813A.
- Etlinger, L. (1974). The electronic calculator: A new trend in school mathematics. Educational Technology, 14, 43-45.
- Eysenck, H. J. (1978). An exercise in mega-silliness. American Psychologist, 33, 317.
- Feldman, K. A. (1971). Using the work of others: Some observations on reviewing and integrating. Sociology of Education, 44, 86-102.
- Ferguson, G. A. (1981). Statistical analysis in psychology and education (5th ed.). New York: McGraw-Hill.
- Gage, N. L. (1978). The scientific basis of the art of teaching. New York: Teachers College Press.
- Gay, L. R. (1981). Educational research: Competencies for analysis and application. Columbus, OH: Charles E. Merrill Publishing Company.
- Giaconia, R. M., and Hedges, L. V. (1982). Identifying features of effective open education. Review of Educational research, 52, 579-602.

- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5, 3-8.
- Glass, G. V. (1978). Integrating findings: The meta-analysis of research. Review of Research in Education, 5, 351-379.
- Glass, G. V., McGaw, B., and Smith, M. L. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage.
- Glass, G. V., and Smith, M. L. (1979). Meta-analysis of research on the relationship of class-size and achievement. Evaluation and Policy Analysis, 1, 2-16.
- Hartley, S. S. (1977). Meta-analysis of the effects of individually paced instruction in mathematics (Doctoral dissertation, University of Colorado). Dissertation Abstracts International, 38, 4003A.
- Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 6, 107-128.
- Hedges, L. V. (1982a). Estimation of effect size from a series of independent experiments. Psychological Bulletin, 92, 490-499.
- Hedges, L. V. (1982b). Fitting categorical models to effect sizes from a series of experiments. Journal of Educational Statistics, 7, 119-137.
- Hedges, L. V. (1982c). Fitting continuous models to effect size data. Journal of Educational Statistics, 7, 245-270.
- Hedges, L. V. (1982d). Statistical methodology in meta-analysis. (ERIC Document Reproduction Service No. ED 227 133).
- Hedges, L. V. (1983). A random effects model for effect sizes. Psychological Bulletin, 93, 388-395.
- Hedges, L. V., and Olkin, I. (1980). Vote-counting methods in research synthesis. Psychological Bulletin, 88, 359-369.
- Hedges, L. V., and Olkin, I. (1983). Regression models in research synthesis. American Statistician, 37, 137-140.
- Higgins, J. L. (1979). Research reporting sections, annual meeting of the National Council of Teachers of Mathematics (57th, Boston, Massachusetts, April 18-21, 1979). Columbus, OH: ERIC Information Analysis Center for Science, Mathematics, and Environmental Education. (ERIC Document Reproduction Service No. ED 167 383)

- Horak, V. M. (1981). A meta-analysis of research findings on individualized instruction in mathematics. Journal of Educational Research, 74, 249-253.
- Hunter, J. E., Schmidt, F. L., and Jackson, G. B. (1982). Meta-analysis: Cumulating findings across studies. Beverly Hills, CA: Sage.
- Jackson, G. B. (1980). Methods for integrative reviews. Review of Educational Research, 50, 438-460.
- Johnson, D. W., Johnson, R. T., and Merayuma, G. (1983). Interdependence and interpersonal attraction among heterogeneous and homogeneous individuals: A theoretical formulation and a meta-analysis of the research. Review of Educational Research, 53, 5-54.
- Jones, L. V., and Fiske, D. W. (1953). Models for testing the significance of combined results. Psychological Bulletin, 50, 375-382.
- Kleinbaum, D. G., and Kupper, L. L. (1978). Applied regression analysis and other multivariate methods. North Scituate, MA: Duxbury Press.
- Kulik, J. A., Kulik, C. C., and Cohen, P. A. (1980). Effectiveness of computer-based college teaching: A meta-analysis of findings. Review of Educational Research, 4, 525-544.
- Leviton, L. C., and Cook, T. D. (1981). What differentiates meta-analysis from other forms of review. Journal of Personality, 49, 231-235.
- Light, R. J. (1980). Synthesizing methods: Some judgment calls that must be made. In H. J. Walberg and E. H. Haertel (Eds.), Evaluation in Education (Vol. 4, pp. 13-17). Oxford, Great Britain: Pergamon Press.
- Light, R. J., and Smith, P. V. (1971). Accumulating evidence: Procedures for resolving contradictions among different research studies. Harvard Educational Review, 41, 429-471.
- Luiten, J., Ames, W., and Ackerson, G. (1980). A meta-analysis of the effects of advance organizers on learning and retention. American Educational Research Journal, 17, 211-218.
- Maccoby, E. E., and Jacklin, C. N. (1974). The psychology of sex differences. Stanford, CA: Stanford University Press.
- Machlowitz, E. (1976). Electronic calculators--friend or foe of instruction? Mathematics Teacher, 69, 104-106.

- Mansfield, R. S., and Bassey, T. V. (1977). Meta-analysis of research: A rejoinder to Glass. Educational Researcher, 6, 3.
- Marcucci, R. G. (1980). A meta-analysis of research on methods of teaching problem solving. Dissertation Abstracts International, 41, 2485A.
- National Council of Teachers of Mathematics. (1974). NCTM Board approves policy statement on the use of minicalculators in the mathematics classroom. NCTM Newsletter, 11, 3.
- National Council of Teachers of Mathematics. (1980). An agenda for action: Recommendations for school mathematics of the 1980s. Reston, VA: National Council of Teachers of Mathematics.
- Nebel, K. (1978). Using the calculator as a teaching aid in the classroom: A summary of findings of the study. Western Springs, IL: Western Springs Public School District No. 101.
- Neubauer, S. G. (1982). The use of hand-held calculators in schools: A review. (ERIC Document Reproduction Service No. ED 220 272)
- Parkhurst, S. (1979). Hand-held calculators in the classroom: A review of the research. (ERIC Document Reproduction Service No. ED 200 416)
- Pillemer, D. B., and Light, R. J. (1980). Synthesizing outcomes: How to use research from many studies. Harvard Educational Review, 50, 176-195.
- Rabe, R. M. (1981). Calculators in the mathematics curriculum: Effects and changes. South Bend, IN: Indiana University. (ERIC Document Reproduction Service No. ED 204 178)
- Roberts, D. M. (1980). The impact of electronic calculators on educational performance. Review of Educational Research, 50, 71-88.
- Rosenthal, R. (1976). Experimenter effects in behavioral research. New York: Irvington.
- Rosenthal, R. (1978). Combining results of independent studies. Psychological Bulletin, 85, 185-193.
- Rosenthal, R. (1980). Combining probabilities and the file drawer problem. In H. J. Walberg and E. H. Haertel (Eds.), Evaluation in Education (Vol. 4, pp. 18-21). Oxford, Great Britain: Pergamon Press.

- Rosenthal, R., and Rubin, D. B. (1979). Comparing significance levels of independent studies. Psychological Bulletin, 86, 1165-1168.
- Rosenthal, R., and Rubin, D. B. (1982). Comparing effect sizes of independent studies. Psychological Bulletin, 92, 500-504.
- Rudnick, J. (1978). Pocket sized calculators versus seventh grade math students. Philadelphia: Temple University.
- Schafer, P., Bell, M. S., and Crown, W. D. (1975). Calculators in some fifth-grade classrooms: A preliminary look. Elementary School Journal, 76, 27-31.
- Schmidt, F. L. (1980, September). The research task of the 1980s: Integrating research findings across studies to produce cumulative knowledge. Paper presented at the meeting of the American Psychological Association, Montreal.
- Schmidt, F. L., and Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.
- Schmidt, F. L., Hunter, J. E., and Caplan, J. R. (1981). Validity generalization results for two occupations in the petroleum industry. Journal of Applied Psychology, 66, 261-273.
- Schmidt, F. L., Hunter, J. E., Pearlman, K., and Shane, G. S. (1979). Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. Personnel Psychology, 32, 257-281.
- Shah, I. (1968). Caravan of dreams. Baltimore: Penguin.
- Sigg, P. O. (1982). The hand-held calculator: Effects on mathematical abilities and implications for curriculum change. South Bend, IN: Indiana University School of Education. (ERIC Document Reproduction Service No. ED 218 147)
- Smith, M. L. (1980). Publication bias and meta-analysis. In H. J. Walberg and E. H. Haertel (Eds.), Evaluation in Education (Vol. 4, pp. 22-24). Oxford, Great Britain: Pergamon Press.
- Smith, M. L., and Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. American Psychologist, 32, 752-760.
- Stock, W., Okun, M., Haring, M., Miller, W., Kinney, C., and Ceurvorst, R. (1982). Rigor in data synthesis: A case study of reliability in data analysis. Educational Researcher, 11, 10-14, 20.

- Suydam, M. N. (1976). Electronic hand calculators: The implications for pre-college education. Final report. Abbreviated version. Washington, D.C.: National Science Foundation. (ERIC Document Reproduction Service No. ED 127 206)
- Suydam, M. N. (1978). State-of-the-art review on calculators: Their use in education. Columbus, OH: Calculator Information Center. (ERIC Document Reproduction Service No. ED 167 426)
- Suydam, M. N. (1979a). The use of calculators in pre-college education: A state-of-the-art review. Columbus, OH: Calculator Information Center. (ERIC Document Reproduction Service No. ED 171 573)
- Suydam, M. N. (1979b). Calculators: A categorized compilation of references. Columbus, OH: Calculator Information Center. (ERIC Document Reproduction Service No. ED 171 572)
- Suydam, M. N. (1979c). Reference bulletins from the Calculator Information Center. Columbus, OH: Calculator Information Center. (ERIC Document Reproduction Service No. ED 167 426)
- Suydam, M. N. (1980a). Using calculators in pre-college education; Third annual state-of-the-art review. Columbus, OH: Calculator Information Center. (ERIC Document Reproduction Service No. ED 206 454)
- Suydam, M. N. (1980b). Calculators: A categorized compilation of references. Supplement 1. Columbus, OH: Calculator Information Center. (ERIC Document Reproduction Service No. ED 199 087)
- Suydam, M. N. (1981a). The use of calculators in pre-college education: Fourth annual state-of-the-art review. Columbus, OH: Calculator Information Center. (ERIC Document Reproduction Service No. ED 206 454)
- Suydam, M. N. (1981b). Reference bulletins from the Calculator Information Center. Bulletins 25-28. Columbus, OH: Calculator Information Center. (ERIC Document Reproduction Service No. ED 206 452)
- Suydam, M. N. (1982a). The use of calculators in pre-college education: Fifth annual state-of-the-art review. Columbus, OH: Calculator Information Center. (ERIC Document Reproduction Service No. ED 220 273)
- Suydam, M. N. (1982b). Calculators: A categorized compilation of references. Supplement 2. Columbus, OH: Calculator Information Center. (ERIC Document Reproduction Service No. ED 213 598)

- Suydam, M. N. (1982c). References to calculator uses in education. Reference Bulletins Nos. 29 and 30. Columbus, OH: Calculator Information Center. (ERIC Document Reproduction Service No. ED 225 886)
- Suydam, M. N. (1983). Calculators: A categorized compilation of references. Supplement 3. Columbus, OH: Calculator Information Center. (ERIC Document Reproduction Service No. ED 225 855)
- Toepfer, C. F., Jr. (1979). Brain growth periodization--A new dogma for education. Middle School Journal, 10, 3, 18-20.
- Turinese, D. M. (1982). A use of the hand calculator in the second year algebra curriculum. Dissertation Abstracts International, 43, 102A. (University Microfilms No. 82-13466)
- Walberg, H. J., and Haertel, E. H. (1980). Research integration: An introduction and overview. In H. J. Walberg and E. H. Haertel (Eds.), Evaluation in Education (Vol. 4, pp. 5-10). Oxford, Great Britain: Pergamon Press.
- Wheatley, G. H., and Shumway, R. J. (1979). Impact of calculators in elementary school mathematics. Washington, D.C.: National Science Foundation. (ERIC Document Reproduction Service No. 175 720)
- Willson, V. L., and Putnam, R. R. (1982). A meta-analysis of pretest sensitization effects in experimental design. American Educational Research Journal, 19, 249-258.
- Wolins, L. (1962). Responsibility for raw data. American Psychologist, 17, 657-658.
- Yin, R. K., Bingham, E., and Heald, K. A. (1976). The difference that quality makes: The case of literature reviews. Sociological Methods and Research, 5, 139-156.

APPENDICES

APPENDIX A

BIBLIOGRAPHY OF STUDIES USED IN THE META-ANALYSIS

1. Abo-Elkhair, M. (1981). An investigation of the effectiveness of using minicalculators to teach the basic concepts of average in the upper elementary grades (Doctoral dissertation, Florida State University, 1980). Dissertation Abstracts International, 41, 2980A.
2. Aldridge, W. S. (1977). Effects of electronic calculators on achievement of middle school remedial mathematics students (Doctoral dissertation, University of Georgia, 1976). Dissertation Abstracts International, 37, 4078A.
3. Allen, M. B. (1976). Effectiveness of using hand-held calculators for learning decimal quantities and the metric system (Doctoral dissertation, Virginia Polytechnic Institute). Dissertation Abstracts International, 37, 850A.
4. Andersen, L. E. (1977). The effects of using restricted and unrestricted modes of presentation with electronic calculators on the achievement and attitude of seventh grade pupils. Dissertation Abstracts International, 37, 6321A. (University Microfilms No. 77-07399)
5. Boling, M. A. (1977). Some cognitive and affective aspects of the use of hand-held calculators in high school consumer mathematics classes (Doctoral dissertation, Louisiana State University). Dissertation Abstracts International, 38, 2623A.
6. Borden, V. L. (1976). Teaching decimal concepts to sixth grade students using the hand-held calculator (Doctoral dissertation, University of Northern Colorado). Dissertation Abstracts International, 37, 4192A.
7. Brey, R. K. (1980). Effects of problem-solving activities and calculators on problem solving and computation in grade four (Doctoral dissertation, University of Michigan). Dissertation Abstracts International, 41, 1914A.
8. Campbell, P., and Virgin, A. E. (1976). An evaluation of elementary school mathematics programs utilizing the mini-calculator. Toronto, Canada: Ontario Department of Education. (ERIC Document Reproduction Service No. ED 137 120)
9. Cech, J. P. (1972). The effect of the use of desk calculators on attitude and achievement with low-achieving ninth graders. Mathematics Teacher, 65, 183-186.

10. Chang, L. L. (1979). An examination into the effects of calculator-assisted instruction on the mathematics achievement and attitude of seventh and eighth grade disadvantaged students. Dissertation Abstracts International, 40, 1323A. (University Microfilms No. 79-20254)
11. Connor, P. J. (1981). A calculator dependent trigonometry program and its effect on achievement and attitude of eleventh and twelfth grade college bound students (Doctoral dissertation, Temple University). Dissertation Abstracts International, 42, 2545A.
12. Connor, T. J. (1979). An investigation of the use of hand-held calculators by students in elementary schools. Gainesville, FL: University of Florida. (ERIC Document Reproduction Service No. ED 190 355)
13. Dean, D. K. (1981). The effectiveness of using a hand-held calculator as an instructional aid in teaching the basic multiplication facts to fourth graders (Doctoral dissertation, Michigan State University, 1980). Dissertation Abstracts International, 41, 3829A.
14. Duffy, J. A., and Thompson, G. E. (1980). Improving mathematics achievement of elementary students through the use of electronic hand calculators. Columbus, OH: Columbus, Ohio Public Schools, Department of Management Planning and Information. (ERIC Document Reproduction Service No. ED 209 263)
15. Dye, D. L. (1981). The use and non-use of calculators on assessment testing. St. Paul, MN: Minnesota State Department of Education. (ERIC Document Reproduction Service No. ED 211 359)
16. Eckmier, J. L. (1978). An investigation of the use of calculators with low-achieving fourth-grade students in mathematics achievement and attitude (Doctoral dissertation, University of Southern California). Dissertation Abstracts International, 38, 7109A.
17. Edens, H. S. (1983). Effects of the use of calculators on mathematics achievement of first grade students (Doctoral dissertation, University of Virginia, 1981). Dissertation Abstracts International, 43, 3248A.
18. Elliott, J. W. (1981). The effect of using hand-held calculators on verbal problem solving ability of sixth grade students (Doctoral dissertation, University of Oregon, 1980). Dissertation Abstracts International, 41, 3464A.

19. Engelmeyer, W. J. (1979). The effectiveness of hand-held calculators for the remediation of basic multiplication facts (Doctoral dissertation, University of Maryland, 1978). Dissertation Abstracts International, 39, 5381A.
20. Fesharaki, M. (1979). A study of the effect of hand-calculators on achievement, estimation, and retention of seventh and eighth graders on decimals and percent. Dissertation Abstracts International, 39, 6004A. (University Microfilms No. 79-06885)
21. Fischman, M. L. (1976). New York City high school students' attitudes and concept learning in business arithmetic when using electronic calculators as contrasted with hand calculation. Dissertation Abstracts International, 37, 774A. (University Microfilms No. 76-19025)
22. Forsyth, R. A., and Ansley, T. N. (1982). The importance of computational skill for answering items in a mathematics problem-solving test: Implications for construct validity. Educational and Psychological Measurement, 42, 257-263.
23. Fugate, B. R. (1979). An assessment of attitudes, self-concept, and mathematical achievement resulting from the use of minicalculators (Doctoral dissertation, North Texas State University, 1978). Dissertation Abstracts International, 39, 6531A.
24. Gaslin, W. L. (1975). A comparison of achievement and attitudes of students using conventional or calculator-based algorithms for operations on positive rational numbers in ninth-grade general mathematics. Journal for Research in Mathematics Education, 6, 95-108.
25. Hawthorne, F. S., and Sullivan, J. J. (1975). Using hand-held calculators in sixth-grade mathematics lessons. New York State Mathematics Teachers Journal, 25, 29-31.
26. Herrington, J. P. (1980). Experimental analysis of cognitive style and information processing variables in mathematical problem-solving (Doctoral dissertation, Southern Illinois University at Edwardsville). Dissertation Abstracts International, 41, 973A.
27. Hopkins, B. L. (1978). The effect of a hand-held calculator curriculum in selected fundamentals of mathematics classes (Doctoral dissertation, University of Texas at Austin). Dissertation Abstracts International, 39, 2801A.

28. Hutton, L. A. (1977). The effects of the use of minicalculators on attitude and achievement in mathematics (Doctoral dissertation, Indiana University, 1976). Dissertation Abstracts International, 37, 4934A.
29. Hutton, L. A. (1978). An experimental study of the effects of mini-calculators on mathematics achievement and attitude in grades four, five, and six. Unpublished report. Indianapolis, IN: Indiana University-Purdue University.
30. Jamski, W. D. (1977). The effect of hand calculator use on the achievement of seventh-graders learning rational number-decimal-percent conversion algorithms (Doctoral dissertation, Indiana University, 1976). Dissertation Abstracts International, 37, 4934A.
31. Jones, E. W. (1976). The effect of the hand-held calculator on mathematics achievement, attitude and self-concept of sixth-grade students (Doctoral dissertation, Virginia Polytechnic Institute and State University). Dissertation Abstracts International, 37, 1387A.
32. Kasnic, M. J. (1978). The effect of using hand-held calculators on mathematical problem-solving ability among sixth grade students. Dissertation Abstracts International, 38, 5311A. (University Microfilms No. 78-01276)
33. Kenner, M. (1981). Teaching decimal math with calculators. (ERIC Document Reproduction Service No. ED 214 290)
34. Keough, J. J., and Burke, G. W. (1969). Utilizing an electronic calculator to facilitate instruction in mathematics in the 11th and 12th grades. Albany, NY: New York State Education Department, Division of Research. (ERIC Document Reproduction Service No. ED 037 345)
35. Kloupoushak, E. L. (1979). The use of hand-held calculators in grade 8 mathematics (Doctoral dissertation, University of Alberta, Canada, 1978). Dissertation Abstracts International, Volume X1979.
36. Kobrin, B. (1978). The hand-held calculator: Effects on intermediate grade mathematics achievement (Doctoral dissertation, Brigham Young University). Dissertation Abstracts International, 39, 3354A.
37. Kolpas, S. J. (1979). The use of electronic calculators as in-class instructional aids in a ninth-grade arithmetic program (Doctoral dissertation, University of Southern California). Dissertation Abstracts International, 39, 5293A.

38. Langbort, C. R. (1983). An investigation of the ability of fourth grade children to solve word problems using hand-held calculators. Dissertation Abstracts International, 43, 2914A. (University Microfilms No. 83-00563)
39. Laursen, K. W. (1978). Use of calculators in high school general mathematics: A study comparing achievement, attitude, and attendance of general mathematics students who used calculators with students who did not (Doctoral dissertation, Brigham Young University). Dissertation Abstracts International, 39, 733A.
40. Lawson, T. J. (1978). A study of the calculator's and altered calculator's effect upon student perception and utilization of an estimation algorithm (Doctoral dissertation, University of New York at Buffalo, 1977). Dissertation Abstracts International, 39, 647A.
41. Lewis, J., and Hoover, H. D. (1981). The effect on pupil performance of using hand-held calculators on standardized mathematics achievement tests. (ERIC Document Reproduction Service No. ED 204 152)
42. Lunder, D. A. (1978). The impact of the calculator on test anxiety for high school consumer mathematics students. Dissertation Abstracts International, 39, 6607A. (University Microfilms No. 79-10673)
43. Miles, M. M. (1980). A study of the interrelationship of the hand-held calculator, achievement in mathematical computation and problem-solving, and attitude toward mathematics of eighth grade students. Dissertation Abstracts International, 41, 974A. (University Microfilms No. 80-18831)
44. Miller, D. P. (1977). Effectiveness of using minicalculators as an instructional aid in developing the concept and skill of long division at the fifth grade level (Doctoral dissertation, Florida State University, 1976). Dissertation Abstracts International, 37, 6327A.
45. Moore, B. H. (1982). The effect of the hand-held calculator on attitude toward mathematics and mathematics achievement of third-grade learners (Doctoral dissertation, University of San Francisco). Dissertation Abstracts International, 43, 1457A.
46. Moser, J. M. (1979). The effect of calculator supplemented instruction upon the arithmetic achievement of second and third graders. Madison, WI: University of Wisconsin Research and Development Center for Individualized Schooling. (ERIC Document Reproduction Service No. ED 180 764)

47. Murphy, N. K. (1981). The effects of a calculator treatment on achievement and attitude toward problem solving in seventh grade mathematics. Dissertation Abstracts International, 42, 2008A. (University Microfilms No. 81-21439)
48. Nelson, D. W. (1976). Effects of using hand calculators on the attitudes and computational skills of children in grades four through seven (Doctoral dissertation, Arizona State University). Dissertation Abstracts International, 37, 3382A.
49. Noone, J. A. (1980). Effects of the use of hand-held calculators on mathematics achievement and attitude toward mathematics of seventh grade students (Doctoral dissertation, University of Virginia, 1979). Dissertation Abstracts International, 40, 3849A.
50. Papritan, J. C. (1979). The effects of pocket calculators versus paper and pencil procedures on algorithm achievement of vocational agriculture students (Doctoral dissertation, Ohio State University, 1978). Dissertation Abstracts International, 39, 4660A.
51. Pedersen, D. A. (1979). The effect of the calculator on the elementary mathematics student. Dissertation Abstracts International, 39, 4794A. (University Microfilms No. 79-02850)
52. Prigge, G., and Langemo, J. (1978). Effects of minicalculators on the pre- and co-requisite skills of intermediate school children. Grand Forks, ND: Department of Mathematics, University of North Dakota.
53. Quinn, D. R. (1976). The effect of the usage of a programmable calculator upon achievement and attitude of eighth and ninth grade algebra students. Dissertation Abstracts International, 36, 4234A. (University Microfilms No. 76-00874)
54. Schafer, P., Bell, M. S., and Crown, W. D. (1975). Calculators in some fifth-grade classrooms: A preliminary look. Elementary School Journal, 76, 27-31.
55. Schnur, J. O., and Lang, J. R. (1976). Just pushing buttons or learning?--A case for minicalculators. Arithmetic Teacher, 23, 559-562.
56. Schoen, H. L., Oehmke, T., and Tarr, J. (1980). The Iowa problem-solving project: Development and evaluation. (ERIC Document Reproduction Service No. ED 199 052)

57. Shively, J. C. (1981). An investigation of effects of the hand-held calculator on the mathematics achievement of students at the seventh grade level (Doctoral dissertation, University of Southern California, 1980). Dissertation Abstracts International, 41, 2921A.
58. Shult, D. L. (1980). The effect of the hand-held calculator on arithmetic problem-solving abilities of sixth grade students (Doctoral dissertation, University of Oregon, 1979). Dissertation Abstracts International, 40, 6179A.
59. Spencer, J. N. (1975). Using the hand-held calculator in intermediate grade arithmetic. Dissertation Abstracts International, 35, 7048A. (University Microfilms No. 75-10376)
60. Standifer, C. E., and Maples, E. G. (1981). Achievement and attitude of third-grade students using two types of calculators. School Science and Mathematics, 81, 17-24.
61. Standifer, C. E., and Maples, E. G. (1982). Mathematical achievement of third- and fourth-grade students in compensatory education when using two types of calculators. Psychological Reports, 51, 11-18.
62. Stewart, J. T., Jr. (1981). Using the hand-held calculator as a computing aid for instruction in word-problem solving with elementary grade students (Doctoral dissertation, University of Illinois, 1980). Dissertation Abstracts International, 41, 4634A.
63. Sutherlin, W. N. (1977). The pocket calculator: Its effect on the acquisition of decimal estimation skills at intermediate grade levels (Doctoral dissertation, University of Oregon, 1976). Dissertation Abstracts International, 37, 5663A.
64. Szetela, W. (1979). Hand-held calculators and the learning of trigonometric ratios. Journal for Research in Mathematics Education, 10, 111-118.
65. Szetela, W. (1980). Calculators and the teaching of ratios in Grade 7. Journal for Research in Mathematics Education, 11, 67-70.
66. Szetela, W. (1982). Story problem solving in elementary school mathematics: What differences do calculators make? Journal for Research in Mathematics Education, 13, 381-389.

67. Szetela, W., and Super, D. (1983). Calculators and strategies for problem solving in Grade 7: An implementation program and study (Report No. 83:3). Vancouver, British Columbia: Educational Research Institute of British Columbia, Vancouver. (ERIC Document Reproduction Service No. ED 228 036)
68. Toole, B. A. (1980). Evaluation of the effectiveness of calculator assisted curriculum and instruction in ninth grade general mathematics classes (Doctoral dissertation, University of California-Berkeley, 1979). Dissertation Abstracts International, 40, 3852A.
69. Townsend, G. C. (1979). The effect of programmable calculator use on probability estimation achievement and attitude toward estimation of students in second year algebra (Doctoral dissertation, Indiana University). Dissertation Abstracts International, 40, 1936A.
70. Vaughn, L. R. (1977). A problem of the effects of hand-held calculators and a specially designed curriculum on attitude toward mathematics, achievement in mathematics, and retention of mathematical skills (Doctoral dissertation, University of Houston, 1976). Dissertation Abstracts International, 37, 4938A.
71. Wajeeh, A. (1976). The effect of a program of meaningful and relevant mathematics on the achievement of the ninth grade general mathematics student. Dissertation Abstracts International, 37, 2801, 2802. (University Microfilms No. 76-26189)
72. Ward, D. E. (1979). The effect of the electronic calculator on problem-solving achievement and attitudes toward mathematics of general mathematics students (Doctoral dissertation, University of Southern California, 1978). Dissertation Abstracts International, 39, 4038A.
73. Wheatley, C. L. (1980). Calculator use and problem-solving performance. Journal for Research in Mathematics Education, 11, 323-334.
74. Wheatley, G. H., and Shumway, R. J. (1979). Impact of calculators in elementary school mathematics. Washington, D.C.: National Science Foundation, Grant No. NSF-SED-77-18077. (ERIC Document Reproduction Service No. ED 175 720)
75. Wheatley, G. H., and Wheatley, C. L. (1982). Calculator use and problem solving strategies of Grade Six pupils. Lafayette, IN: Purdue University. (ERIC Document Reproduction Service No. ED 291 250)

76. Whitaker, W. H. (1977). A study of change in achievement, interest, and attitudinal variates accompanying the use of electronic calculators in a first-grade mathematics curriculum (Doctoral dissertation, University of Southern California). Dissertation Abstracts International, 38, 97A.
77. Williams, D. E. (1979). The effect of the use of the mini-calculator and an associated curriculum supplement on computational skills and attitudes toward arithmetic of ninth grade non-college bound students (Doctoral dissertation, Temple University). Dissertation Abstracts International, 39, 6610A.
78. Wilson, A. W. (1978). The effects of the hand-held calculator upon achievement test scores of elementary school mathematics students. Dissertation Abstracts International, 39, 2116A. (University Microfilms No. 78-19154)
79. Zepp, R. A. (1976). Reasoning patterns and computation on proportions problems, and their interaction with the use of pocket calculators in ninth-grade and college (Doctoral dissertation, Ohio State University, 1975). Dissertation Abstracts International, 36, 5181A.

APPENDIX B

RESULTS OF DATA COLLECTION

This appendix records the data coded from the studies of Appendix A. Part 1 presents the data for the independent variables. Part 2 tabulates the effect magnitudes.

Part 1

The subsequent Table B-1 contains the bulk of information coded for the independent variables described within the "Method" of Chapter IV, "Representing Independent Variables." Insufficient data existed to merit tabulation for ethnicity, amount of drill and practice, time of calculator use per treatment session, teacher training, and teacher experience. For brevity, the remaining variables were assigned the reference designations listed in the legend of Table B-1. The data are segregated by study according to the identification (ID) numbers assigned in Appendix A. Symbols for the coded data are consistent with the descriptions in the "Method" of Chapter IV.

Evaluating research design quality required subjective judgments regarding eight design criteria. To yield consistency in ratings, each criterion was equated to an integer 1 through 6, under preestablished guidelines now described.

1. Problem definition. Clear hypotheses (3); no hypotheses but clear research questions (2); muddled (1).
2. Population description. Thorough (3); partial (2); little (1).

3. Sampling procedures. Fully random (6); existing population with randomized students, teachers, and classes (5); existing population with randomized students, otherwise not random (4); existing classes with randomized groups and teachers (3); existing classes with randomized groups, otherwise not random (2); existing classes (1).

4. Error control. Pretest-Posttest Control Group design (PPCG), with analysis of covariance (ANCOVA), control for pretest-treatment interaction, and low experimenter bias (6); PPCG with ANCOVA, with control for interaction, high bias (5); PPCG with ANCOVA, without control for interaction (4); PPCG without ANCOVA, with control for interaction, low bias (3); PPCG without ANCOVA, with control for interaction, high bias (2); PPCG without ANCOVA, without control for interaction (1). (Note: Designs in their diversities were judged for their equivalence to these guidelines.)

5. Instruments. Standardized (3); experimenter designed with reliability quoted (2); ad hoc (1).

6. Data analysis. Appropriate methods with full disclosure of results (3); appropriate methods with data missing (2); inappropriate methods, or much data missing (1).

7. Conclusions. Appropriate and tied to hypotheses (3); appropriate, but with digressions (2); inappropriate (1).

8. Report: Can a reader critically examine the evidence? Yes (3); more or less (2); no (1).

Data are now presented for the independent variables.

Table B-1
Data Tabulations for Independent Variables by Study Number

ID No.	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12 days	X13 wks	X14	X15
1	D	X	4	Mix		Averages	M	T	F	1/1	1980	7	4	1.8	
2	D	S	6-8	L			L	Sp	F	1/1	1976	20		2.3	
3	D	Xr	6	Mix	L-M	Decimals	L	T	F	1/1	1976	25	4	1.6	
4	D	S	7	Mix			L	T	F	1/2	1976	90		2.6	
5	D	S	12	L-A		Consumer Math	M	T	F	1/1	1977	95		2.2	
6	D	X	6	Mix	LM	Decimals	M	T	F	1/1	1976	20		1.6	
7	D	X	4	Mix		Whole Nos	L	Sp	F	1/1	1980	16		1.9	None
8	ERIC	S	5,6	Mix			L	T	F	1/2	1976	140		2.0	
9	J	S	9	L	M-UM		L	T	F		1972	35		2.3	
10	D	S	7,8	Mix	UL		M	T	F	1/1	1979	120		2.5	
11	D	S	11/12	H	M-UM	Trig	M	Sp	F	1/1	1981	90		2.5	
12	ERIC	S	K,2,4	Mix		Number Concepts	L	T	F	1/2	1980	100		1.7	
13	D	Xr	4	Mix		Mult Facts	L	T	F	1/2	1980	20	6	1.5	
14	ERIC	S	4-6	Mix			L	T	F	1/1	1980	130		2.1	
15	ERIC	S	8	Mix			M	T	F		1981	0		1.6	
16	D	S	4	L			L	T	F	1/2	1978	170		2.3	
17	D	S	1	Mix			L	T	F	1/1	1981	30		2.3	
18	D	S	6	Mix			L	T	F	1/1	1980	30		1.6	
19	D	X	7	L			L	T	F	1/1	1978	10		2.1	
20	D	Xr	7/8	Mix		Mult Facts	L	Sp	P-F1/1		1978	30		1.9	
21	D	S	10	Mix		Decimals	L	T	F		1976	160		2.2	
22	J	S	9,11	Mix		Bus Arith	L	T	F		1982	0		1.9	
23	D	S	4,5	Mix			L	T	F		1978	50		2.4	None
24	J	Xr	9	L		Fractions	M	T	F	1/1	1975	10	2	1.8	
25	J	S	6	Mix			L	T	F	1/1	1975	180		1.7	
26	D	X	7	Mix	L-UM		M	T	F	1/1	1980	3		1.7	None
27	D	S	9	L	L-M	Whole Nos	L	T	F	1/1	1978	20		2.7	
28	D	S	9	Mix		Alg I	L	T	F	1/1	1976	10		2.0	

Table B-1 (continued)

ID No.	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12 days	X13 wks	X14	X15
29	O	S	4	Mix	L-M		L	T	F	1/2	1978	180		2.3	
30	D	Xr	7	Mix	M		L	T	F		1976	15		1.7	
31	D	S	6	Mix	L-M		L	T	F	1/1	1976	45		2.4	FF
32	D	S	6	Mix	M		L	T	F	1/1	1977	9		1.7	
33	ERIC	Xr	7,8	Mix		Decimals	L	T	F		1981	10		1.7	
34	ERIC	S	11/12	A-H		Alg	L	T	F	1/2	1969	180		1.9	
35	D	Xr	8	Mix			M	T	F	1/1	1978	36		1.2	
36	D	S	4-6	Mix			L	T	F	1/1	1978	100		2.1	FM
37	D	X	9	L-A		Gen Math	M	T	F	1/1	1979	90		2.2	
38	D	Xr	4	Mix			M	T	F	1/1	1982	5		2.0	
39	D	Xr	10-12	Mix		Gen Math	L	T	F	1/1	1978	90		2.0	FM
40	D	X	7	Mix			M	T	F		1977	4		1.7	
41	ERIC	S	4,8	Mix			L	T	F		1981	0		1.9	
42	D	S	9-12	Mix		Cons Math	L	T	F	1/1	1978	50		2.5	None
43	D	S	8	Mix			M	T	F	1/2	1980	40		2.2	FF
44	D	Xr	5	Mix		Division	L	T	F	1/1	1976	12		2.1	
45	D	S	3	Mix	L	Mult	L	T	F	1/1	1982	15		1.9	
46	ERIC	Xr	2	Mix	M-U		L	T	F	1/1	1979	145		2.3	
47	D	Xr	7	Mix	M-U		L	T	F	1/2	1981	30	7	2.4	None
48	D	S	4-7				L	T	F	1/1	1976	20		2.2	
49	D	S	7	Mix			L	T	F		1979	40		2.3	
50	D	S	10	Mix		Ag Math	L	T	F	1/1	1978	5		2.2	
51	D	S	2,3,6	Mix			L	T	F	1/1	1978	160		1.9	
52	O	X	3	Mix			M	Sp	F	1/1	1978	20		1.6	
53	D	S	8,9	Mix		Alg I	L	T	F		1975	175		2.5	
54	J	S	5	H	M-U		L	T	F		1975	5		1.9	
55	J	S	3/4	L	L-M		L	T	F	1/2	1976	20		2.0	FF
56	ERIC	S	5,6,8	Mix			L	T	F	1/2	1980	20		2.5	
57	D	S	7	Mix	M-U		L	T	F	1/1	1980	110		2.4	
58	D	Xr	6	Mix	UM		M	Sp	F	1/1	1979	21		2.1	

Table B-1 (continued)

ID No.	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12 days	X13 wks	X14	X15
59	D	S	5	Mix	M		M	T	F	1/1	1974	32		2.2	None
60	J	S	3	A			L	T	F		1981	55		2.6	
61	J	S	3	L			L	T	F		1982	105	12	2.6	
62	D	X	5/6	Mix			L	T	F	1/1	1980	20		1.5	
63	D	X	5/6	Mix			L	T	F	1/1	1976	15	4	1.6	
64	J	Xr	9/10	L-A	M		M	Sp	P		1979	13		1.6	
65	J	Xr	7	Mix			M	T	F	1/2	1980	11		1.7	
66	J	Xr	3-8	Mix			L	T	F	1/2	1982	40		2.2	
67	ERIC	S	7	Mix			L	T	F	1/2	1983	195		1.7	
68	D	S	9	Mix			L	Sp	F		1979	25		2.3	
69	D	Xr	11	Mix			M	T	F		1979	20		1.9	
70	D	Xr	9	L			L	Sp	P	1/1	1976	40	2	1.8	
71	D	S	9				M	Sp	F	1/3	1976	75		1.9	
72	D	S	9	Mix			L	T	F		1978	75		2.5	None
73	J	X	6				L	T	F		1980	30		2.2	
74	ERIC	Xr	2-6	Mix			L	T	F	1/1	1979	67		2.5	
75	ERIC	S	6	Mix			L	Sp	F	1/1	1982	85		2.0	
76	D	S	1	Mix	M		L	T	F	1/2	1977	30		2.2	
77	D	S	9	Mix			L	T	F	1/1	1978	150		2.0	
78	D	S	3-6	Mix			L	T	F	1/2	1978	105		1.8	None
79	D	X	9	Mix			M	T	F	1/1	1975	10		2.0	

Coding Legend for Independent Variables:

X1--Publication form	X6--Subject matter	X11--Year of publication
X2--Instrument type	X7--Experimenter bias	X12--Length of treatment
X3--Grade level	X8--Curriculum	X13--Retention period
X4--Ability level	X9--Use of calculator	X14--Research design quality
X5--SES	X10--Calculator availability	X15--Sex-related differences

Part 2

This section records the effect magnitudes developed from the outcomes of studies used in the meta-analysis. These data are segregated by study according to the ID numbers assigned in Appendix A and are further subdivided as achievement and attitude dependent variables.

Tables B-2 through B-5 present effects for achievement dependent variables, where (E)xperimental and (C)ontrol groups had received equivalent instruction except for calculators. Table B-2 provides effects related to skills acquisition. Tables then appear (1 each) for skills retention, skills transfer, and estimation skills.

Tables B-6 through B-8 present effects for attitude dependent variables, where the E and C groups had received equivalent instruction except for calculators.

Tables B-9 and B-10 present effects for achievement and attitude dependent variables for special instruction with calculators versus traditional instruction without calculators.

Table B-2
Effect Sizes by Study Number for Achievement Dependent Variables Related to Skills Acquisition
(Equivalent Instruction Except for Calculators)

Study ID No.	Basic Operations						Problem Solving					
	Composite		Computation		Concepts		Composite		Productivity		Selectivity	
	Maintenance	Extension	Maint	Ext	Maint	Ext	Maintenance	Extension	Maint	Ext	Maint	Ext
1												
2	-.766	-.291										
	.203											
3	-.153											
4	.009		.212	1.411	-.056		.364					
	.088		.134	1.675	-.213		.098					
5												
								.022				
6	-.189											
7			.052				.121					
8	.198		-.080		.475		.523					
	.045		.090		0		.106					
9	-.092	1.294										
10	-.060	.061					.679					
	.120	.378					.948					
	.645	-.078										
11												
		.473										
12	.414	-.090	-.113		-.066		.074					
	.100		-.078		.278		.086					
13			-.322									
14	.153	.349	.113		.193		-1.028					
	-.182		.357		.341		-.230					
			-.506		.141		.411					
15				.640								
16	-.042		-.062		-.022		.135					
17	-.138		.085		-.361							
18							-.267	.263			.499	
							-.328	.168			.190	
							-.408	.317				

Table B-2 (continued)

Study ID No.	Basic Operations						Problem Solving					
	Composite			Computation			Concepts			Composite		
	Maintenance	Extension		Maint	Ext		Maint	Ext		Maintenance	Extension	
19												
21	-.018	.452		.055	.557 .485		-.090	.347				
22										.093	.156	
23	-.055	-.055										
24			-.124	-.603								
25	.494			.500			.488			-.040		
27		.399			.443 .709					.974	.758	
28	.624 .258 .509	-.067 -.107										
29	-.226	.270		-.479 .088						-.055 .337		
30		1.966 .393	.426									
31	.508			.604			.299					
32					.288 .312					.137	-.276	.142
33	-.106											
34	.342											
35	.509											
36	-.373 -.211 -.169	.302 -.085	.609									
37	.444			.091			.137			.173		
38				.010 -.001 .022	.498 .415 .581					.115 -.045 -.226	.185 -.023 .180	-.446 .183 -.221
										.906 .878 .622		.591 -.159 -.230
										1.135	-.134	-.080

Table B-2 (continued)

[illegible]

Effect Sizes by Study Number for Achievement Dependent Variables Related to Skills Transfer (Equivalent Instruction Except for Calculators)

[illegible]

Effect Sizes by Study Number for Achievement in Acquiring Estimation Skills (Equivalent Instruction Except for Calculators)

Study ID No.	Effect Size
35	.609
40	-.270
63	-.046
69	.031
	.562

Table B-6

[illegible]

Table B-7

Effect Sizes for Attitude Dependent Variables, Study Nos. 27-53
(Equivalent Instruction Except for Calculators)

[illegible]

Effect Sizes for Attitude Dependent Variables, Study Nos. 56-77 (Equivalent Instruction Except for Calculators)

[illegible]

Table B-10

[illegible]

APPENDIX C

MICROCOMPUTER PROGRAM FOR DERIVING STATISTICS RELATED
TO EFFECT SIZE GROUPS IN META-ANALYSIS

The process of meta-analysis is directed toward the estimation of means and confidence intervals for homogeneous groups of effect magnitudes. Hence, for analyzing a selected group of k effects, the following statistics must be calculated: H_T for testing homogeneity; a weighted effect size mean $\bar{g}^W$; and, confidence intervals about the population effect size at predetermined confidence levels. Within the present study, these statistics were derived by the Intertec Data Systems SuperBrain microcomputer, using a specially written program. This appendix describes that program and provides example calculations, pages 226-228.

The program was written for use with a printer. Input data for each effect are the effect size g_i , $i = 1, \dots, k$, and its associated experimental and control group sample sizes n_i^E and n_i^C . Following the input and printout of descriptive information (see example calculations), the computer asks for input data for the initial effect, e.g., $n_1^E = 10$, $n_1^C = 12$, and $g_1 = -.2$. A check for proper entry of data is then requested. After proper entries are assured, the computer calculates the variance of g_1 from the general equation,

$$\sigma_i^2(g_i) = \frac{n_i^E + n_i^C}{n_i^E n_i^C} + \frac{g_i^2}{2(n_i^E + n_i^C)} \quad (C-1)$$

Then the following calculations are performed:

$$P1 = \frac{g_i^2}{\sigma_i^2(g_i)} ; \quad Q1 = \frac{g_i}{\sigma_i^2(g_i)} ; \quad R1 = \frac{1}{\sigma_i^2(g_i)} . \quad (C-2)$$

These are printed out as SUB 1, SUB 2, and SUB 3, respectively, and calculations regarding the first effect are done.

The computer then asks for data regarding the second effect. Once proper data are entered, calculations through equation C-2 are repeated, and the previous SUBs 1, 2, and 3 are respectively added to the present P1, Q1, and R1 to yield a new SUB 1, SUB 2, and SUB 3 which are then printed out. This process is repeated for the third and subsequent effects until all k effects have been considered. For each effect, a printout is provided for the number k, the input data, g_i^2 , and the running subtotals,

$$\begin{aligned} (\text{SUB } 1)_i &= \sum_{i=1}^k \frac{g_i^2}{\sigma_i^2(g_i)} , \quad (\text{SUB } 2)_i = \sum_{i=1}^k \frac{g_i}{\sigma_i^2(g_i)} , \quad \text{and} \\ (\text{SUB } 3)_i &= \sum_{i=1}^k \frac{1}{\sigma_i^2(g_i)} . \end{aligned} \quad (C-3)$$

Basic statistics are then computed from the final subtotals as follows:

$$H_T = H = \text{SUB } 1 - \frac{(\text{SUB } 2)^2}{\text{SUB } 3} ; \quad \text{and,} \quad (C-4)$$

$$\bar{g}^w = M = \frac{\text{SUB } 2}{\text{SUB } 3} . \quad (C-5)$$

Where the mean effect size variance is computed from:

$$\sigma^2(\bar{g}^w) = \frac{1}{\text{SUB } 3} \quad , \quad (\text{C-6})$$

confidence intervals are derived from:

$$\bar{g}^w - z_{\alpha/2} \sigma(\bar{g}^w) \leq \delta \leq \bar{g}^w + z_{\alpha/2} \sigma(\bar{g}^w) \quad . \quad (\text{C-7})$$

Here, α is a desired significant level, and z is referred to the normal distribution. The computer automatically provides CI's at 95 and 90 percent confidence levels.

The computer program and example calculations are now presented, concluding the appendix and the dissertation.

```

10  ' PROGRAM NAME:  META1
15  '
20  REM ** INTRODUCTORY MATERIALS
25  '
30  ' PROGRAM AUTHOR:  RAY HEMBREE
40  ' COMPUTER IDENTIFICATION:  INTERTEC DATA SYSTEMS 'SUPERBRAIN'
41  ' WITH MICROSOFT EXTENDED BASIC
50  ' PROGRAM PURPOSES:  TO CALCULATE A STATISTIC FOR TESTING HOMOGENEITY,
52  ' THE WEIGHTED EFFECT SIZE MEAN, AND 95%/90% CONFIDENCE INTERVALS
54  ' FOR AN EFFECT SIZE GROUP IN META-ANALYSIS
60  ' PROGRAM PRODUCT:  PRINTOUT OF SAMPLE AND EFFECT SIZE INPUT DATA,
62  ' RUNNING SUBTOTALS OF TERMS WHICH LEAD TO THE HOMOGENEITY TEST
63  ' STATISTIC, THE WEIGHTED EFFECT SIZE MEAN, ITS 95 PERCENT
64  ' CONFIDENCE INTERVAL, AND ITS 90 PERCENT CONFIDENCE INTERVAL
70  '
80  ' PRIMARY VARIABLES:
85  ' AN$...USER RESPONSE VARIABLE (10 CHARACTERS MAX)
90  ' I$....IDENTIFYING INFORMATION (50 CHARACTERS MAX)
95  ' NE....EXPERIMENTAL GROUP SAMPLE SIZE
100 ' NC....CONTROL GROUP SAMPLE SIZE
105 ' G.....EFFECT SIZE
110 ' GG....EFFECT SIZE SQUARED
115 ' V.....EFFECT SIZE VARIANCE
120 ' P1....INDIVIDUAL TERM 1 OF TEST EQUATION
125 ' P2....SUBTOTAL TERM 1
130 ' Q1....INDIVIDUAL TERM 2 OF TEST EQUATION
135 ' Q2....SUBTOTAL TERM 2
140 ' Q3....Q2 SQUARED
145 ' R1....INDIVIDUAL TERM 3 OF TEST EQUATION
150 ' (RECIPROCAL OF VARIANCE V)
155 ' R2....SUBTOTAL TERM 3
158 ' K.....NUMBER OF EFFECT SIZES
160 ' H.....TEST STATISTIC FOR HOMOGENEITY
165 ' M.....WEIGHTED EFFECT SIZE MEAN
170 ' L1....LOWER 95% CONFIDENCE LIMIT
175 ' U1....UPPER 95% CONFIDENCE LIMIT
180 ' L2....LOWER 90% CONFIDENCE LIMIT
185 ' U2....UPPER 90% CONFIDENCE LIMIT
190 ' MV....VARIANCE OF EFFECT SIZE MEAN M
195 ' RT....SQUARE ROOT OF MV
200 '
210 ' CONSTANTS:
215 ' INITIAL VALUES K=P2=Q2=R2=0
220 ' Z1=1.960=FACTOR FOR 95% CONFIDENCE INTERVAL
225 ' Z2=1.645=FACTOR FOR 90% CONFIDENCE INTERVAL
1000 '
1010 REM ** PROGRAM UTILIZATION
1015 REM ** INITIALIZATION
1020 DIM I$(50)
1030 K=0:P2=0:Q2=0:R2=0
1035 Z1=1.96: Z2=1.645

```

```

1040 REM ** ENTRY OF IDENTIFYING INFORMATION
1050 LINE INPUT "PRESS 'RETURN' IF YOUR PRINTER IS READY"; AN$:PRINT
1060 LPRINT "STATISTICS FOR META-ANALYSIS"
1070 LINE INPUT "ENTER TODAY'S DATE: "; I$
1080 LPRINT "DATE: "; I$
1090 LINE INPUT "ENTER DEPENDENT VARIABLE (50 CHAR MAX): "; I$
1100 LPRINT: LPRINT "DEPENDENT VARIABLE: "; I$
1110 LINE INPUT "ENTER SUBCATEGORY: "; I$
1120 LPRINT"   SUBCATEGORY: "; I$: LPRINT: LPRINT
1130 LPRINT" K   NE   NC   G       G SQ           SUB 1           SUB 2           SUB 3"
1135 LPRINT
1137 PRINT CHR$(26)
1140 LINE INPUT "PRESS 'RETURN' TO INPUT DATA OR ENTER 'STOP': "; AN$
1150 IF AN$="STOP" THEN 2000
1170 '
1180 REM ** ENTRY OF DATA
1190 '
1200 INPUT "NE = "; NE
1203 INPUT "NC = "; NC
1206 INPUT "G = "; G
1210 PRINT "ARE "; NE; ", "; NC; ", "; G; "   WHAT YOU WANTED?"
1212 INPUT AN$: PRINT
1215 IF AN$="NO" THEN 1140
1220 GG=G*G:B=NE+NC:C=NE*NC:D=2*B:E=B/C:F=GG/D
1230 V=E+F:R1=1/V:P1=GG*R1
1260 Q1=G*R1:K=K+1:P2=P1+P2
1290 Q2=Q1+Q2:R2=R1+R2
1300 LPRINT K;TAB(5)NE;
1310 LPRINT TAB(10)NC;TAB(16)G;TAB(22)GG;TAB(34)P2;TAB(47)Q2;TAB(58)R2
1320 GOTO 1140
1325 '
2000 REM ** CALCULATION OF TEST STATISTIC
2010 '
2015 LPRINT "END OF DATA": LPRINT: LPRINT
2020 Q3=Q2*Q2:H=P2-(Q3/R2)
2040 LPRINT "NUMBER OF EFFECT SIZES: K = "; K
2050 LPRINT "TEST STATISTIC: H = "; H: LPRINT
2060 REM ** CALCULATION OF MEAN AND CONFIDENCE INTERVALS
2070 M=Q2/R2
2080 LPRINT "WEIGHTED EFFECT SIZE MEAN: M = "; M: LPRINT
2090 MV=1/R2
2100 RT=MV^.5
2110 L1=M-(Z1*RT): U1=M+(Z1*RT)
2120 LPRINT "95% CI: ("; L1; ", "; U1; ")": LPRINT
2130 L2=M-(Z2*RT): U2=M+(Z2*RT)
2140 LPRINT "90% CI: ("; L2; ", "; U2; ")
3000 END

```

STATISTICS FOR META-ANALYSIS

DATE: ANYTIME

DEPENDENT VARIABLE: EXAMPLE CALCULATIONS

SUBCATEGORY: NONE

K	NE	NC	G	G SQ	SUB 1	SUB 2	SUB 3
1	10	12	-.2	.04	.217105	-1.08553	5.42763
2	18	24	-.1	.01	.319837	-2.11284	15.7008
3	25	31	.1	.01	.458059	-.730619	29.523
4	16	36	.2	.04	.899256	1.47537	40.5529
5	40	40	.3	.09	2.67923	7.40862	60.3304
6	20	24	.4	.16	4.39074	11.6874	71.0273
7	29	27	.5	.25	7.78048	18.4669	84.5863
8	87	92	.6	.36	23.1852	44.1415	127.377
9	76	49	.7	.49	36.9779	63.8453	155.526
10	50	50	.8	.64	51.7928	82.3638	178.674

END OF DATA

NUMBER OF EFFECT SIZES: K = 10

TEST STATISTIC: H = 13.8252

WEIGHTED EFFECT SIZE MEAN: M = .460973

95% CI: (.314342 , .607604)

90% CI: (.337908 , .584039)

VITA

Ray Hembree has taken a circuitous pathway toward his work in education.

A native Tennessean, he attended public schools in Cumberland County, then spun nylon with DuPont in Chattanooga. Induction into service granted duty as a military policeman in Louisiana, North Carolina, and Germany. After separation, he enrolled at Tennessee Tech in Cookeville, suppressing a wish to teach mathematics in favor of the bachelor's degree in electrical engineering. His technical career then spanned some twenty years, first as systems engineer and engineering manager with the Martin-Marietta Corporation, Orlando, Florida. He returned to East Tennessee in 1976 as plant manager of Dalen Products, Inc., Knoxville.

There, his wish to teach mathematics was rekindled, and proximity to the University of Tennessee abetted the flame. He enrolled in its graduate school in 1981 and claimed a Master's degree in 1982. He then continued study for the Ph.D., believing that if he departed without that degree, return for its pursuit would be unlikely. Such is the nature of time and circuitous pathways.

He received the doctoral degree in education, August 1984. Upon graduation, he accepted appointment as assistant professor of mathematics at Adrian College, Michigan, his destination.