Quantitative literature review techniques: a replication and examination of their role in educational evaluation.

Maureen Dianne Bennett Gates

University of Windsor

Follow this and additional works at: https://scholar.uwindsor.ca/etd

Recommended Citation
https://scholar.uwindsor.ca/etd/3257

This online database contains the full-text of PhD dissertations and Masters' theses of University of Windsor students from 1954 forward. These documents are made available for personal study and research purposes only, in accordance with the Canadian Copyright Act and the Creative Commons license—CC BY-NC-ND (Attribution, Non-Commercial, No Derivative Works). Under this license, works must always be attributed to the copyright holder (original author), cannot be used for any commercial purposes, and may not be altered. Any other use would require the permission of the copyright holder. Students may inquire about withdrawing their dissertation and/or thesis from this database. For additional inquiries, please contact the repository administrator via email (scholarship@uwindsor.ca) or by telephone at 519-253-3000 ext. 3208.
NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

THIS DISSERTATION HAS BEEN MICROFILMED EXACTLY AS RECEIVED

AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

LA THÈSE A ÉTÉ MICROFILMÉE TELLE QUE NOUS L'AVONS RÉCEVE
QUANTITATIVE LITERATURE REVIEW TECHNIQUES: A REPLICATION
AND EXAMINATION OF THEIR ROLE IN EDUCATIONAL EVALUATION

by

Maureen Dianne Bennett Gates

A Thesis
submitted to the Faculty of Graduate Studies
through the Faculty of Education in Partial Fulfillment
of the requirements for the Degree of Master of Education at
The University of Windsor

Windsor, Ontario, Canada

1984
DEDICATION

To my friend and husband Peter.
ACKNOWLEDGEMENTS

I would like to acknowledge my committee members, Dr. Ian Crawford, Dr. Wilf Innerd, and Dr. Sheila Cameron for their input and suggestions. I am also very grateful to Kent Campbell for his assistance in locating articles and for permitting his own reference texts to be borrowed for extended periods of time and to John Bennett for proof reading the final draft. I would also like to thank Dr. Donald Laing for his many helpful comments. My deepest thanks go to my husband, Peter, for his continual support, patience, and belief in my potential.
# TABLE OF CONTENTS

ABSTRACT ......................................................... vii

DÉDICAT...ON ..................................................... i

ACKNOWLEDGEMENTS ....................................... ii

LIST OF TABLES ............................................... v

LIST OF FIGURES ............................................... vi

Chapter One  INTRODUCTION 1

  Background to the Problem ................................. 1
  Statement of the Problem ................................ 3
  Relevance to Education ................................. 6
  Purpose of the Study ................................ 9
  Scope and Delimitations .............................. 10
  Summary ................................................. 12

Chapter Two  REVIEW OF RELATED LITERATURE 14

  Methods of Literature Review .......................... 14
  Narrative Method ........................................ 15
  Voting Method ........................................... 17
  Chi-Square Method ..................................... 19
  Probability Methods .................................. 21
  Fail-safe N ............................................... 26
  Effect Size Method ................................. 27
  Combining Quantitative Techniques .................. 29
  Section Summary ....................................... 32

  Applications of Quantitative Review Techniques 32
  Historical Overview .................................. 32
  General Applications ................................ 35
  Educational Applications ......................... 36
  Section Summary ...................................... 43
Chapter Three  METHOD

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Defining the Sample</td>
<td>45</td>
</tr>
<tr>
<td>Sample Size</td>
<td>45</td>
</tr>
<tr>
<td>Method of Coding Individual Studies</td>
<td>47</td>
</tr>
<tr>
<td>Characteristic of Studies</td>
<td>48</td>
</tr>
<tr>
<td>Quantifying Findings from Studies</td>
<td>48</td>
</tr>
<tr>
<td>Effect Sizes</td>
<td>53</td>
</tr>
<tr>
<td>Derived Zs</td>
<td>59</td>
</tr>
<tr>
<td>Specific Coding Concerns</td>
<td>63</td>
</tr>
<tr>
<td>Quantifying Non-significant Results</td>
<td>64</td>
</tr>
<tr>
<td>Negative Outcome Measures</td>
<td>64</td>
</tr>
<tr>
<td>Design of Analysis</td>
<td>65</td>
</tr>
<tr>
<td>Computer Programme for the Analysis</td>
<td>66</td>
</tr>
<tr>
<td>Summary</td>
<td>67</td>
</tr>
</tbody>
</table>

Chapter Four  RESULTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preliminary Analysis</td>
<td>69</td>
</tr>
<tr>
<td>Data from All Studies</td>
<td>71</td>
</tr>
<tr>
<td>Data by Outcome Measures</td>
<td>78</td>
</tr>
<tr>
<td>Data by Category of Exceptionality</td>
<td>80</td>
</tr>
<tr>
<td>Data by Type of Classroom Setting</td>
<td>83</td>
</tr>
<tr>
<td>Correlations of Study Features with Effect Size</td>
<td>83</td>
</tr>
<tr>
<td>Summary</td>
<td>85</td>
</tr>
</tbody>
</table>

Chapter Five  DISCUSSION

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interpretation of Results</td>
<td>86</td>
</tr>
<tr>
<td>Efficacy of Special Versus Regular Class</td>
<td>86</td>
</tr>
<tr>
<td>Placement</td>
<td>86</td>
</tr>
<tr>
<td>Replicability of Quantitative Reviews</td>
<td>90</td>
</tr>
<tr>
<td>Role of Quantitative Reviews in Evaluation</td>
<td>99</td>
</tr>
<tr>
<td>Conclusions</td>
<td>102</td>
</tr>
<tr>
<td>Implications for Future Research</td>
<td>104</td>
</tr>
<tr>
<td>Summary</td>
<td>107</td>
</tr>
</tbody>
</table>

APPENDIX

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appendix A  Codebook</td>
<td>108</td>
</tr>
<tr>
<td>Appendix B  Programme 1</td>
<td>112</td>
</tr>
<tr>
<td>Programme 2</td>
<td>113</td>
</tr>
</tbody>
</table>

BIBLIOGRAPHY

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>General References</td>
<td>116</td>
</tr>
<tr>
<td>Studies in Quantitative Review Sample</td>
<td>122</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Probit Transformations Used in Estimating an Effect Size</td>
<td>60</td>
</tr>
<tr>
<td>2. Description of Studies in the Analysis</td>
<td>73</td>
</tr>
<tr>
<td>3. Description of the Overall Results in the Replication and Carlberg and Kavale's Study</td>
<td>77</td>
</tr>
<tr>
<td>4. Average Effect Size by Domain of Outcome Measure</td>
<td>79</td>
</tr>
<tr>
<td>5. Average Effect Size by Category of Exceptionality</td>
<td>81</td>
</tr>
<tr>
<td>6. Comparison Between replication and Carlberg and Kavale's Data in Terms of Category of Exceptionality</td>
<td>82</td>
</tr>
<tr>
<td>7. Correlations of Study Features with Effect Size</td>
<td>84</td>
</tr>
<tr>
<td>Obtained by Carlberg and Kavale and in the Replication</td>
<td></td>
</tr>
<tr>
<td>Figure</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td>1. Illustration of the Findings of a Meta-analytic Review on the Effects of Psychotherapy Reported by Smith Glass, and Miller, 1980</td>
<td>30</td>
</tr>
<tr>
<td>2. Model for the Estimation of E.S. from Percentage Data</td>
<td>58</td>
</tr>
</tbody>
</table>
ABSTRACT

QUANTITATIVE LITERATURE REVIEW TECHNIQUES: A REPLICATION AND EXAMINATION OF THEIR ROLE IN EDUCATIONAL EVALUATION

by

Maureen Dianne Bennett Gates

A package of quantitative literature review techniques, consisting of Glass's effect size, Stouffer's Z, and a fail-safe N, was used to replicate Carlberg and Kavale's meta-analysis of the efficacy of special versus regular class placement of exceptional children. The literature reviewed in the replication included twenty-six of the studies in Carlberg and Kavale's sample and four from an updated literature search. Whereas Carlberg and Kavale found special classes to be inferior to regular class placement, this review obtained an overall effect size of 0.105 which indicated that students in special education were functioning at a level approximately 3.5% above their counterparts who had experienced a regular classroom placement. The one-tailed probability of such an outcome occurring by chance was less than .014 and at least 127 studies with disconfirming results would be needed to raise the probability to more than .05. The discrepancies between the findings of the replication and the original study...
were examined through direct comparisons of pairs of effect sizes. Several coding practices in the quantitative approach were identified as being ambiguous, and consequently, subject to a variety of interpretations. The differences between interpretations demonstrated the influence of a reviewer's judgments and documented the extent of subsequent biasing in the outcomes. The implications of reviewer biasing were discussed in terms of the utility of quantitative reviews in the planning of educational policies and programmes.
INTRODUCTION

Background to the Problem

Literature reviews are a constant companion to the research process, either as a mandatory component of new proposals or as an independent scholarly work. They fulfil a variety of purposes, including the establishment of facts, the identification of inconsistencies among studies, gleaning support for some theories while denouncing others, and to provide the justification for new endeavours (Cook and Leviton, 1980; Jackson, 1980; Borg and Gall, 1983). As it is not uncommon to find that several hundred studies have been conducted on any given topic, the traditional approach of summarizing the salient features of each individual study is no longer viable.

Most authors attempt to rectify the situation by narratively reviewing just a portion of the available literature. The intuitive impressions of related studies, that comprise such a review have been extensively criticised. On various occasions the method has been described as being highly subjective and frequently misrepresentative (Glass, 1977; Rosenthal, 1978; Cooper, 1979; Cook and Leviton, 1980; Kavale and Glass, 1981). Blatant biases are also repeatedly associated with qualitative reviews. Glass (1976) observed that:
A common method for integrating several studies with inconsistent findings is to carp on the design or analysis deficiencies of all but a few studies—those remaining frequently being one's own work or that of one's students or friends—and then advance the one or two "acceptable" studies as the truth of the matter. (p. 4)

Accordingly, such flawed reviews fail to meet the needs of the research community. Rather than assembling the information at hand, they add to the already confusing puzzle. The important issues remain unresolved, theoretical progress is impeded, and new research proceeds in a haphazard manner. For every narrative review that is published in support of one position, another can be found with opposing results, and yet a third will state that the evidence is inconclusive. As the number of primary studies continues to increase, the deficiencies of the narrative method will be exacerbated.

Numerous methods have been devised over the years to permit a quantitative, in contrast to a qualitative, review of literature. Some of the processes, which are described in detail in the next chapter, are simple, requiring nothing more difficult than keeping a tally of the number of studies falling into a specified category. Other methods, involving complicated calculations, are best undertaken with computer assistance. The basic premise for a quantitative review is that separate reviewers "using the same synthesizing procedure should arrive at the same statistical output" (Pillemer and Light, 1980, p. 177). Since the results are derived from systematic methods, the quantitative approach is considered to
be less subjective. These techniques also allow the amalgamation of the results from an unlimited number of studies sharing a common hypothesis, thereby surpassing the narrative review in terms of both its feasibility and objectivity.

The utility of the quantitative approach has been widely demonstrated. Green and Hall (1984) have recently speculated that quantitative reviews "probably number in the hundreds" (p. 39). In addition to augmenting traditional narrative reviews, they have been used as the sole means to organize and understand vast bodies of literature. It has even been suggested that quantitative reviews may become a compulsory feature in dissertations, grant proposals, research reports, and the sole form of review in professional journals (Cook and Leviton, 1980). If this is to be the case, then those involved in any aspect of the research process, will require a thorough understanding of the various quantitative methods.

Statement of the Problem

The quantitative review format has the appearance of being such an attractive alternative to the narrative method, that it is being employed in hundreds of studies without being fully understood. The assumed superiority of the quantitative procedures has yet to be completely substantiated. Are these techniques the panacea they appear to be? The mathematical
derivations of the formulae have been established, but all the postulated benefits have not.

One key problem lies in the objectivity of the quantitative procedures. Although various authors have lauded the quantitative techniques as being more scientific and providing more rigorous information (Cook and Leviton, 1980; Pillemer and Light, 1980; and Glass, 1977), there have been very few attempts to either demonstrate the replicability of the procedures or verify the basic assumption that they are indeed objective. Rosenthal (1983) has noted that contrary to its allusion of objectivity:

...the meta-analytic enterprise is neither mechanical nor perfectly reliable. The process is made up of a series of judgments just as is the case in any other data analysis. For that reason meta-analyses are just as much in need of replication as are individual studies that make up the meta-analysis. (p.298)

Since most reviewers exercise their own judgment in the definition of the topic under consideration, the location and inclusion of studies, and finally in the choice of the appropriate measurement tool, the quantitative techniques are open to subjective biases similar to those identified in the narrative method (Cook and Leviton, 1980; Jackson, 1980). As long as reviewers continue to apply the quantitative techniques to as many new topics as possible, without replicating previous works, the effects of these biases will not be fully understood.

The other main problem is the inconsistent use of the nomenclature associated with the quantitative techniques.
Glass (1977) suggested the term 'meta-analysis' as a means of distinguishing the "statistical analysis of the findings of many individual analyses" (p. 352) from the traditional narrative review. As such, meta-analysis is not a specific technique. Instead, it "is a perspective that uses many techniques of measurement and statistical analysis" (Glass, McGaw, and Smith, 1981). Despite the fact that all quantitative procedures are subsumed under this definition, most authors (for example Kavale, 1982; Kavale and Carlsberg, 1980; Luiten, Ames, and Ackerson, 1980; Hearold, 1979) conducting a meta-analysis, focus on a determination of the magnitude of the effects utilizing a particular method developed by Glass (1977) exclusively. As a consequence, meta-analysis is becoming synonymous with this one technique. The implications of this phenomenon are extensive.

By considering the individual techniques to be mutually exclusive, that is, only using one technique with a given set of studies, reviewers have inadvertently restricted the scope of their analyses. Conclusions drawn from only aggregate probability values are limited to confirming the existence of a relationship and cannot provide any indication of its size. In opposition to this, reviewers utilizing just an effect size calculation document the magnitude of the relationship without indicating its significance. Reviews based on a single quantitative technique, while purporting to be comprehensive, are actually incomplete and through their omissions, sometimes misleading. The maximum potential of the
quantitative (or meta-analytic) approach can only be realized from judicious combinations of several techniques into one review.

Relevance to Education

In recent years, there has been an increasing demand for accountability in education, particularly, through programme evaluations. Educators need to know the extent to which a programme has achieved its goals and also document any unanticipated outcomes before they can decide if the programme should be continued, modified, expanded, curtailed, or completely abandoned. Theoretically, an evaluation is a systematic investigation of the functioning, quality and effects of a programme, that provides information to assist educators in making such decisions. The results of an evaluation should lead to improvements in curriculum planning that would, in turn, provide a relevant education, in better and more efficient ways.

This, however, is not always the case (Pillemer and Light 1980). In fact there are very few instances in which the results of a study are implemented. Resistance to change and the politics of the situation are only partially to blame, as are inadequate evaluations that cannot supply the information necessary for effective programme innovation. The main impediment is actually the inherent problem that a single
evaluation study does not provide an irrefutable statement of a programme's operation. Instead, the findings are frequently inconclusive or ambiguous, and highly dependent on the specific conditions in effect at the same time as the programme (Weiss, 1972). Often additional studies evaluating similar programmes furnish conflicting results that only serve to heighten the problem.

A thorough quantitative review of the existing evaluative studies on any given type of programme would clarify the situation for educational administrators. Information gathered from several evaluations would reveal insights into the programme functioning that would otherwise be undetected and identify issues warranting supplementary research. Debates about the relative merits of programmes, such as infant stimulation, computerized instruction, and special education could be resolved. In addition to meeting the specific needs of curriculum planners, quantitative reviews would facilitate theoretical progress in other areas of education. Such reviews could make significant contributions to understanding various psychological issues that have a bearing on child development, learning, and the sociology of education.

Mainstreaming, which is the practice of placing exceptional children in regular classes, exemplifies the type of issue that can be investigated using quantitative methods. At one point in time, it was a widely held belief that exceptional children, such as those classified as mentally
retarded, learning disabled, behaviourally disordered, and emotionally disturbed, should be educated apart from 'normal' children (Hallahan and Kauffman, 1970). In a segregated classroom, the exceptional children could receive more individual attention from teachers who were highly trained in special education. Conversely, the normal children would not be subjected to the disruptive presence of exceptional children in their classroom. The efficacy of the special classroom came under scrutiny in the 1960's. It was suggested that the segregated special classrooms only served to accentuate the differences between the exceptional and normal children. This, in turn, aggravated the academic and social problems the exceptional children experienced. (Christopolus and Renz, 1969; Dunn, 1968; Johnson, 1962).

The manifest evidence required to establish which environment best meets the needs of the exceptional child has not yet been provided. Evaluative studies of individual programmes and traditional literature reviews "have routinely studied numerous hypotheses, techniques, diagnoses, outcomes, and mediating variables which have resulted in ambiguous results" (Carlberg and Kavale, 1980, p. 296). Consequently, administrators have had to rely on either the economics of the situation, biased information, or pure intuition in choosing whether to integrate or segregate exceptional children. A complete quantitative review of the pertinent literature would assist educators in ascertaining which type of teaching arrangement allows the social and academic potential of
exceptional children to be maximized.

Purpose of the Study

The purpose of this study is to provide a much needed replication and expansion of an existing, single technique, meta-analysis. A select combination of quantitative review techniques, similar to the 'package' suggested by Cooper (1979) and the analysis conducted by Johnson et al. (1981) will be used to re-examine the studies in Carlberg and Kavale's (1980) review of the literature on mainstreaming. The techniques utilized in this study will include a Stouffer Z score, a fail-safe N, and Glass's measure of effect size. The results of the replication will be compared with the original review to ascertain the influence of reviewer judgments. Any discrepancies between the conclusions drawn from the initial and replicated reviews will be apparent. The updated information on the efficacy of special as opposed to regular class placement of exceptional students will be of benefit to curriculum planners currently involved in the 'mainstreaming' debate. Findings based on this extensive, multi-faceted review, will supply additional information to assist in the decision process. Furthermore, the BASIC computer programmes written for the calculations that are involved in this study will provide a convenient means of either replicating other existing reviews or approaching new
topics for a quantitative analysis.

Scope and Delimitations

Since this study is a practical demonstration of quantitative review techniques, it will not delve into the mathematical derivation of the various procedures. Several articles and books (Hunter, Schmidt, and Jackson, 1982; Rosenthal, 1979; Glass, 1977; Jackson, 1980; Pillemer and Light, 1980; Glass, McGaw, and Smith, 1981; Cooper, 1984; Rosenthal, 1984) are available on the underlying theories and should be consulted by those seeking a better knowledge of the procedures. Suffice to say, that at this point in time, there is no argument about the mathematical development of the procedures. All controversy centres on the way in which the methods are implemented. Therefore, only issues relevant to the practical application of quantitative reviews will be discussed in this study.

The question of whether or not special students should be placed in regular classrooms was chosen for replication because the initial meta-analysis by Carlberg and Kavale (1980) employed only the effect size method which does not provide sufficient information to meet the needs of educational planners. The history and debate surrounding the issue of mainstreaming are not of concern in this study and will only be discussed if it necessary to clarify a point in the replication. The fact that mainstreaming is still
controversial indicates that the usual programme evaluations have not been able to satisfactorily resolve the issue and that a quantitative review of the literature is in order. Such terms as integrative review and quantitative review, since they are not entirely interchangeable, must be clearly defined prior to this study.

Integrative Review

Integrative reviews are aimed at "inferring generalizations about substantive issues from a set of studies directly bearing on those issues" (Jackson, 1980, p.348). This type of review combines the findings from individual studies into a comprehensive statement that identifies important, unresolved, issues and indicates the direction of future research. If the integrative review is thorough, it should have synthesized knowledge from all pertinent sources and be a description of the 'state of the art'.

Quantitative Review

This term will be used to encompass all methods of reviewing literature that are based on statistical theory. The quantitative review must "describe numerically the characteristics of a body of evidence" (Cooper, 1979, p. 131). Quantitative reviews may be aimed at recording methodological developments in a given field, verifying theories, combining
knowledge from different disciplines, or integrating information as described in the previous section.

Summary

The traditional narrative review no longer meets the needs of educational researchers and planners. Contradictory conclusions and biased information arising from such reviews hinder both the development of new projects and the modification of existing programmes. Light and Smith (1971) noted that "when the purpose of the review is to develop specific policies...such inconsistencies can paralyze attempts at public action" (p. 430). While quantitative reviews appear to be the most plausible alternative, they are not without faults.

Two features of the quantitative approach have contributed to its usefulness in extracting information from a large body of data. Reviewers have a highly diversified repertoire of quantitative techniques that can be applied to almost any research question. Secondly, the results of these procedures claim to provide objective, rather than subjective, information. These features, unfortunately, also form the basis for two main problems associated with the quantitative methods. There is almost a 'fad' in the application of the effect size method to new topics. As a result, the possible contributions of the other qualitative techniques are ignored and investigations of the approach's objectivity are
neglected. Reviewers tend to "throw everything [and anything] into the 'knowledge pot', stir a bit, and hope that an answer will bubble to the top" (Corday and Bootzin, 1983, p. 287).

This study addresses both problems. It incorporates several quantitative techniques into one review 'package' and in so doing, replicates an existing review. In addition to providing valuable information on the issue of mainstreaming, the techniques and computer programmes presented in this study will be of benefit to those either planning or reading a quantitative review.
REVIEW OF THE LITERATURE

The chapter on relevant literature is included in a thesis to provide an understanding of the previous research pertaining to a given problem (Borg and Gall, 1983). In order to facilitate an understanding of quantitative review procedures, this chapter will be separated into two sections. The first will address the methodological techniques involved in literature reviews and will include a discussion of the benefits and limitations associated with each of the techniques. The applications of the quantitative techniques will be presented in the second section. Consideration will be given to general as well as educational applications.

Methods of Literature Review

Over the years several procedures have been developed to determine the overall significance of the research findings of related studies. Even though the various procedures differ in their required information, complexity of calculation, and applicability, they have one aspect in common. The outcome of the review is only as good as its sample of studies. Unless the review is based on either an exhaustive survey of the literature or a selection of representative studies related to a given hypothesis, the combined results will be biased. This
would defeat the purpose of an integration, which is to provide an as accurate as possible representation of the topic under consideration.

Narrative Method

In the traditional narrative review, the results of each separate study are individually interpreted, then summarized into an overall conclusion about the topic under consideration. When there are only one or two dozen studies investigating a specific hypothesis, a complete narrative review is possible, and if it follows certain guidelines it can be quite adequate. However, with the vast majority of research questions, primary studies are too numerous to allow a complete narrative review. An incomplete, or partial sampling of the available studies has rarely furnished an accurate picture of the issue under investigation.

Unfortunately, the scope of the reviews are extremely diverse, and depending on the comprehensiveness and accuracy, an author can surmise almost any conclusion. Some of the problematic features of this method of synthesizing research into one body of information have been outlined by Jackson (1978). He noted that:

1. Reviewers frequently cited previous literature surveys on the same topic without critically examining the methods, evidence, and conclusions that were presented. Of the 27 studies, from Jackson's sample of 36 that used existing reviews, only 6% examined them critically.
2. Reviewers often ignore the full range of available studies, and instead, focus their attention on an unspecified subset. Such a sample can easily be misrepresentative of the topic as a whole. Only 3% of Jackson's sample analyzed the full set of available studies. A fair sampling of the literature was selected in 22% of the reviews, while a mere 3% acknowledged the use of a published index—e.g. Psychological Abstracts and ERIC—in locating articles.

3. The fact that random sampling error can cause variable findings among studies was frequently ignored by reviewers.

4. Most reviewers used a priori judgments to preclude some works from their surveys without discussing the influences of the perceived design flaws.

5. Reviewers fail to provide a sufficient description of their methods of reviewing to allow an assessment of the validity of the conclusion they have drawn.

Without the inclusion of the details of the review procedure, replication is impossible. Reviewers hold such widely differing opinions, that there is not even agreement about the interpretation of findings from specific studies. It was observed by Miller (1977) that among three narrative reviews, based on virtually the same literature, there was not even agreement as whether or not a particular study demonstrated positive, negative, or an inconclusive results.

Clearly, "the process is dangerously vulnerable to the injection of prejudice and bias" (Glass, Smith, and Miller, 1980, p.8). However, this does not imply that the narrative review should be entirely discredited as a research tool. The method offers a "certain contextual richness" (Pillemer and Light, 1980, p. 178) that is missing from the numerical methods. For example, a narrative review of the historical antecedents to the issue under investigation can be incorporated into any quantitative analysis.
Voting Method

The first attempt to provide a more technical means for reviewing research is the "voting method". This method simply involves classifying all studies into three or four categories based on whether or not a statistically significant difference was reported. According to Light and Smith (1971) the results of each study has one, out of three, possible outcomes. The relationship between independent and dependent variable is statistically significant in either a positive or negative direction, or else demonstrates no significance. Other scholars choose to divide the third category into non-significant positive and non-significant negative trends (Jackson, 1980; Borg and Gall, 1983). Regardless of which framework is used, the number of studies within each category is then counted. Whichever category contains a plurality of studies is considered to be the winner. It is assumed that the modal category gives the best estimate of the true relationship in question.

Light and Smith (1971) also described the main difficulty with this procedure. It neglects the effect of the sample size. More statistically significant findings are produced in large samples than in small ones. Given that nine small-sample studies demonstrate a trend in the expected direction without reaching significance, and one large sample study reported a significant finding. The score would be one 'for' and nine 'against', which is contrary to what would be
expected. The obvious solution is to weight the results of a study according to the sample size. The question then arises, as to how the results should be weighted.

Another deficiency of the voting method is that it determines the existence of a relationship, but not its strength or importance. To use the example of one 'for', nine 'against', the winner is quite apparent, but there is not an accurate indication of the closeness of the contest. Moreover, Hedges and Olkin (1980), noted that the probability of concluding that there is not a positive effect using the vote counting strategy, increases as the number of studies in the integration increases. While the voting method, also referred to as a 'box-score' analysis (Kavale and Glass, 1981), is easy to use and allows the inclusion of virtually every study, it can produce very, misleading results.

A modified version of this method, called the 'sign approach' (Walberg and Haertel, 1980) focuses only on the direction of the results. Any statistical significances of the effects are ignored. The probability of obtaining the observed proportion of studies is calculated, using a two-tailed distribution, and compared to the conventional .05 level of significance. Although simple to use, even this form of the vote counting method can produce erroneous results, in that it fails to detect an effect despite the presence of a large size effect.
Chi-Square Method

Prior to describing this means of consolidating research, discrepancies pertaining to its classification and appellation must be clarified. Some authors treat this method as being distinct from other procedures utilizing one-tailed probabilities based on z or t statistics (Borg and Gall, 1983; Glass, 1977). In opposition to this, the chi-square method is viewed as being the "best known" (Walberg and Haertel, 1980, p. 8) and the "most discussed of all the methods of combining independent probabilities" (Rosenthal, 1978, p. 188). To further complicate the matter, various terms, including Fisher method, method of adding logs, and Pearson's chi-square, have been used to label this format. Since it is such a prominent method of literature analysis, it shall be consistently referred to as the 'chi-square' method for the purpose of this paper and described separately from other probability theories.

The chi-square (\(X^2\)) procedure ameliorates some of the problems encountered with the voting method and permits the integration of research findings into one body of evidence to test the null hypothesis in question. Rather than rely solely on the direction and statistical significance of the individual study findings, the chi-square method includes both the size of the sample and the magnitude of the reported relationships. It is derived from the fact that the natural logarithm of a given probability (\(\log_{e}p\)), when multiplied by \(-2\), follows a chi-square distribution with two degrees of
freedom (df). When combined with the additive properties of independent chi-squares, the following equation to determine the overall level of significance of the N studies being reviewed is possible:

\[ \chi^2(\text{df} - 2N) = \sum_{i=1}^{N} (-2 \log p_i) \]

According to Gage (1978), the inferential statistics reported in each study are converted into exact probability values through the use of standard tables or complicated calculations. Any probabilities derived from results that are contrary to the expected direction are subtracted from 1.00 so that they are reversed. Then the chi-square values and the degrees of freedom are summed across all studies in the integration. In order to ascertain the overall significance of the combined studies, the probability associated with the summed chi-square value is compared to a specified level of significance.

Despite its widespread use, this method has several notable disadvantages. Primarily, it is only applicable when the studies in the review are independent of one another. In some instances, a study will conduct several significance tests on the one set of subjects. Since such tests are not independent, only one probability value can be used from each study. One approach to this problem (Gage, 1978), is that the least significant result from each study should be used to
obtain a conservative estimate of the issue. Alternatively, the one result from each study that would be most relevant to the problem under investigation could be used in the computation of the aggregate chi-square. Another problem with this method is that it can yield results that are inconsistent with other simple tests (Rosenthal, 1978) and in fact demonstrate significance even when all the original studies failed to do so (Walberg and Haertel, 1980). However, it can be argued that the purpose of a literature integration is to reveal obscured information. Comparisons between just two studies pose a particular problem for the chi-square method. If the two studies obtained strongly significant, but opposing results, this procedure would support the significance of either outcome. The combination of a $p$ of .001 for $A>B$ and a $p$ of .001 for $B>A$ is less than .01 for either $A>B$ or $B>A$ (Rosenthal, 1978).

Probability Methods

Several means of combining individual study results are actually subsumed under this title. The chi-square method, mentioned in the previous section, is merely one of the forms of pooling exact one-tailed probabilities. Adding ts, Adding Zs, Adding Weighted Zs, Testing the Mean $p$, and Testing the Mean $Z$ are all options available within the probability method.

Adding probabilities is one of the more recently
published methods based on this theory. Edgington (1972) noted that the chi-square method, since it was based on the addition of logs, determined the significance of the product of the probabilities, and not the significance of their sum. He proposed that the null hypothesis could be tested on the basis of a summation of the observed $p$ values, that is, then raised to an exponent equal to the number of studies in the literature survey ($n$), and subsequently divided by $n!$. This formula is restricted to surveys wherein the total of the summed probabilities does not exceed 1. The correction factor described by Edgington (1972), for use when the overall probability value is greater than unity is:

$$\frac{S^n}{n!} - \binom{n}{1} \frac{(S-1)^n}{n!} + \binom{n}{2} \frac{(S-2)^n}{n!} - \binom{n}{3} \frac{(S-3)^n}{n!} + \ldots$$

where $n$ is the number of $p$s being added and $S$ is the sum of these $p$s.

Additional terms are employed until the value that is raised to the exponent $n$, is less than 1. This modification gives the exact summed probability, assuming that the individual probabilities are from a continuous distribution. When used as a framework for a literature analysis, the probabilities will have been derived from the statistics reported in each of the sample studies. As such, the individual probabilities are from a discrete distribution, and
the resulting summed value is greater than the exact probability. Theoretically, this method supplies a valid measure of the overall probability of the issue being investigated that yields results comparable to the chi-square method. However, the practicality of the measure for large scale reviews is questionable. Edgington states that "Ordinarily one would not combine ps from as many as seven experiments" (1972, p. 353), and Rosenthal, (1978) also acknowledges that this method has limited applicability and precommends that it is used when the number of sample studies is small.

Another method that is limited by too few, rather than too many studies in the review, is adding ts. First described by Winer (1971), it is based on the knowledge that the variance of the t distribution for any given degree of freedom is $df/(df-2)$. The square root of this value is then used as the denominator, with the sum of all obtained ts acting as the numerator to form the following equation:

$$z = \frac{\sum t_i}{\sqrt{(\sum t_i^2)/(df-2)}}$$

The calculated Z score is associated with a specific probability level that supplies a measure of the survey's results. The restriction associated with this method is not a real hinderance in most situations warranting a literature integration. The procedure of adding ts is viable when each
component of the denominator contains a positive value. As such, the degrees of freedom for each sample in the survey must be equal to, or greater than three. The probability values obtained with this method are typically larger than the results from either of the two preceding formulae.

One simple version of the probability method that has no restrictions on its usage is Stouffer's method of adding Zs. This procedure is similar to the previously described probability methods, in that it assumes the data has: 1) a linear relationship; 2) a normal distribution; 3) homogeneity of variance; 4) independence among events; and 5) requires the pooling of one-tailed probabilities. The last requirement, necessitates a decision about the direction of the hypothesis being tested, prior to implementing the procedure. Hence, any two-tailed p reported in the individual studies must be altered to comply with this criterion. Two-tailed p values in support of the stated hypothesis must be halved before the change into Z scores can be made. A p opposed to the hypothesis, is also halved, but its associated Z score is given a negative value. In the original method described by Stouffer (1949), each p value is transformed into its corresponding Z score, as described above, and then used in the following equation to calculate the standard normal deviate (Zma) for the combined research:

\[
Z_{ma} = \frac{Z_{a1} + Z_{a2} + \ldots + Z_{an}}{(n_a)^{1/2}}
\]
The total number of studies in the sample is represented by \( n \) and \( Z_{an} \) is the term for each study's standard normal deviate. Statistical tables are used at the beginning of the procedure to obtain the \( Z \) scores for each individual study, and again when the computation is complete, to determine the probability that the results of the reviewed studies were generated by chance.

Critics of this analysis note that no consideration is given to the size of sample in each of the reviewed studies. Such complaints are easily rectified. Any size effects can be incorporated into the formula by the inclusion of weighting factors. Adjustments to the formula can also be instituted to allow for the quality of the studies included in the integration. Studies that meet specified methodological requirements can be weighted more heavily than those studies that are demonstrated to be unsound. Weighting the Stouffer method, when sample size is being considered, is accomplished with the equation being modified to:

\[
Z_{an} = \frac{N_{a1}Z_{a1} + N_{a2}Z_{a2} + \ldots + N_{an}Z_{an}}{(N_{a1}^2 + N_{a2}^2 + \ldots + N_{an}^2)^{1/2}}
\]

The quantities of \( N_{a1}, \ldots, N_{an} \), as the number of subjects in each reviewed study, are the only alterations from the original formula. Weighting the procedure in this manner, leads to a lower probability level being associated with larger studies having large \( Z \) scores, and concomitantly, a
higher probability when smaller studies produce larger Z scores. Cooper (1979) notes that this refinement, developed by Mosteller and Bush in 1954, increases the precision of the method and therefore provides a more objective answer to the question "What is the probability that a set of studies exhibiting these results could have been generated if no actual relation existed?" (Cooper, 1979, p. 134).

Fail-safe \( N \)

The Stouffer method can also be manipulated into a format that yields a 'fail-safe' number (Cooper, 1979). The fail-safe number, \( (N_{Fs}) \), indicates how many studies with a demonstrated total Z score of zero are required to raise the summed probability to a given significance level, usually of .05. This, in turn, is a measure of the 'strength' of the review's conclusions, in terms of the perceived completeness of the sample of studies drawn from the population. From the original procedure, the number of studies necessary to raise the integrated probability level to more than .05, can be shown as:

\[
N_{Fs,.05} = \frac{(Z_1 + Z_2 + \ldots Z_n)^2}{1.645} - (N_{S})
\]

The figure of 1.645 is the Z score of a two-tailed .05 probability. The resulting \( N_{Fs,.05} \) is the actual number of additional studies, with a summed null relationship, that
would be have to be included in the review to increase the overall probability to a level of significance above .05. The .05 value is arbitrary, and any desired significance level can be substituted in its place.

Effect Size Method

Although Glass's (1977) effect size method is a recent approach to combining research outcomes, its use has become widespread in all areas of social inquiry. The popularity is a consequence of its applicability to almost all primary studies (only those failing to report any form of statistical measures are excluded), and comparative ease of calculation and interpretation. The method involves the conversion of the reported statistics of each study, into one common scale, called an 'effect size'. By such a transformation, comparisons among studies utilizing different measures and research methodologies are possible. The basic form of calculating the effect size (E.S.) is:

\[
E.S. = \frac{\bar{X}_{exp} - \bar{X}_{con}}{SD_{con}}
\]

in which \(X_{exp}\) is the mean from the treatment/experimental group and both \(X_{con}\) and \(SD_{con}\) are, respectively, the mean and standard deviation from the control/comparison group.

More complicated versions of this formula have been
developed for use in obtaining an E.S. from various types of inferential statistics that are frequently reported in the literature. Regardless of which formula is used, an E.S. is calculated for each of the studies in the literature review, the mean of which provides an estimate of typical relationship under investigation. Moreover, for each study, additional characteristics can be related to the E.S. using regression techniques or the base metric could be changed to a product-moment correlation. The effect size method is therefore one of the more versatile combinatory techniques currently available.

It also one of the more controversial (Cooper, 1981; Glass, McGaw, and Smith, 1981; Cahen, 1980; Cordray and Bootzin, 1983). Exception has been taken to the position that well designed and poorly designed studies yield similar findings and that a meaningful effect size can be obtained from such studies (Mansfield and Busse, 1977). Critics also note that the effect size method tends to oversimplify the results of the sample studies (Hedges, 1983). It has also been argued that combining material from widely differing studies is a 'most dubious' practice (Maher, 1983). Cook and Leviton (1980) attribute the controversy to the less conservative nature of the effect size method. The results of an E.S. analysis frequently demonstrate effects that were not apparent from a traditional narrative review. This is exemplified by Plaum et al. (1980), who reported a significant difference in literacy development methods from a
quantitative analysis that had not been demonstrated in previous reviews utilizing a narrative format. Narrative reviewers had also concluded that psychotherapy was ineffective (Eysenck, 1966), yet a study by Smith, Glass, and Miller (1980) revealed a significant improvement in outcomes by the treatment groups. In their meta-analysis of more than 500 independent studies, an overall effect size of 0.85 standard deviations (s) was obtained. Such an effect size indicates that the performance of a subject who received treatment, as measured on the aggregated outcome variables, was at 0.85s above the mean of the untreated controls. This in turn, may be translated as an average improvement of more than 29% after intervention (see Figure 1).

Combining Quantitative Techniques

Quantitative reviewers have a wide range of techniques to choose from when developing an analytical ‘package’. Some of the techniques can be readily dismissed because of intrinsic flaws. The vote counting technique, for example, suffers from many of the same problems as found in a narrative review and the method of adding probabilities is restricted to very small sample sizes. Other techniques can be excluded from a review if they provide redundant information. As such, there is no need to have more than one means to test the significance of the probability in any given review. In order to provide a complete picture of the
Figure 1. Illustration of the Findings of a Meta-analytic Review on the Effects of Psychotherapy Reported by Smith, Glass, and Miller, 1980
issue under investigation, all the studies that incorporate several quantitative techniques into one review package must include an estimation of the size and significance of any obtained relationships. To meet this last criterion, most reviewers employ the methods developed by Glass (1977) and Stouffer (1949).

Compared to the single technique reviews, there is a dearth of combinatorial studies. Works by Rosenthal and Rubin (1978); Cooper (1979); Arkin, Cooper, and Kolditz (1980); Johnson et al. (1981) represent the majority of efforts in this field. With the exception of the study by Johnson et al., all of these reviews include calculations of Stouffer's Z score, Glass's effect size, and the fail-safe number of studies with 'no effect' conclusions that would be needed to invalidate the obtained probability level. Johnson and his colleagues also use the voting method in their meta-analysis of the effects of goal structure on achievement. No justification for this choice is provided in their review, and the limited, if any benefit from its inclusion is questionable in view of the additional work it entails. Cooper extends the review package to also include a weighted version of Stouffer's technique. Ordinarily, the influence of sample size can be determined through an analysis of study characteristics in conjunction with Glass's effect size procedures and the inclusion of a weighted Stouffer's is unnecessary.
Section Summary

From this section it is quite evident that there are many, highly diversified means of conducting a quantitative review. The number of original studies, time constraints, inherent disadvantages, and the benefits associated with each method are all factors that must be considered in the planning stage. Of these techniques the Stouffer Z score, fail-safe N, and Glass's effect size method provide the most information without being redundant and they permit the inclusion of almost every study that reports statistical findings.

Applications of Quantitative Review Techniques

Quantitative techniques, as discussed in the previous section, were not developed in isolation from applied research. They are, for the most part, responses to specific problems encountered by various researchers, and not just theoretical exercises. As a result, the quantitative approach to literature review has been used in many highly diversified areas.

Historical Overview

The concept of quantitative literature analysis is not new. Since the method of adding Z scores was published by Stouffer in 1949, there has been interest in the idea and
controversy surrounding its application. However, until the numbers of studies to be reviewed became too great, there was no pressure to turn away from the traditional qualitative method. By the 1970's, reviews covering more than 1,600 hundred individual works on a given topic were not uncommon (Glass, 1977), and the need for a more viable means of keeping abreast of the literature in all the social sciences became evident. The interest subsequently generated in quantitative review procedures has led to the development of a wide assortment of analytic techniques that are applicable to a variety of situations.

Stouffer’s method itself, was the result of an overabundance of information, and the need to sift out relevant items. The Second World War provided copious quantities of statistics on the American soldiers and how they perceived military service. The number of returned servicemen, and the need to understand what they had been through, prompted research that Stouffer (1949), compiled into a definitive volume on military life.

An investigation of the relationship between the number of lists serving as interference and memory recall (Underwood, 1957) was another of the early examples of combining research findings in something other than a qualitative manner. Without employing any complicated statistical techniques, Underwood was able to aggregate the findings from 16 studies into one conclusion. Underwood noted
a similarity of research designs among the studies when the variations in the type of material that constituted a list were disregarded, and decided to simply graph the percent recall that was recorded versus the number of previous lists presented by each experimenter. The graph revealed that as the amount of interference increased, the ability to recall the original information decreased. This was an important discovery in the field of human memory and clearly demonstrated the benefits in using a quantitative review procedure.

The next major uses of quantitative techniques were reviews in the mid-1970's. Sudman and Bradburn (1974) and Rosenthal (1976) considered aspects of the research process in their reviews. Both included several hundred studies in their works, respectively on: response effects in survey research; and experimenter expectancy effect in behavioural research.

The following year, meta-analysis provided the format for an evaluation of the outcomes of psychotherapy (Smith and Glass, 1977). This was an attempt to resolve the debate on the effectiveness of psychotherapy in an objective and unbiased fashion. An effect size was calculated for each dependent variable included in the 520 surveyed studies. A mean effect size of .85 was obtained, which according to Glass and Smith indicated that the outcome of psychotherapy was superior to a non-treatment control group. The analysis did not provide a definitive statement on the effectiveness of the therapeutic process. Instead, it initiated theoretical arguments about
the procedure, as well as serving as a catalyst for further meta-analytic applications.

General Applications

Quantitative reviewing has become a widespread practice in most areas of sociology and psychology. Some topics that have been investigated include: the neuropsychological assessment of children (Davidson, 1978); sex differences in decoding verbal cues (Hall, 1978); sex differences in conformity research (Cooper, 1979); test validity in personnel selection (Pearlman, 1979); effects of television on social behaviour (Hearold, 1979); validity of employment tests with respect to racial differences (Hunter, Schmidt, and Hunter, 1979); the efficacy of stuttering treatment (Andrews, Guitar, and Howie, 1980); an investigation of Fiedler's contingency model of leadership effectiveness (Strube and Garcia, 1981); sex differences in influenceability (Eagly and Carli, 1981); motivational biases in the attribution of responsibility for an accident (Burger, 1981); effects of competitive, individualistic goal structures on achievement (Johnson, Maruyama, Johnson, Nelson, and Skon, 1981); cognitive gender differences (Rosenthal and Rubin, 1982b); and the effectiveness of attention and rejection as coping styles (Mullen and Suls, 1982). Some additional aspects of psychotherapy that have also been quantitatively reviewed are: drug therapy and psychological disorders
(Miller, 1977); psychotherapy and medical utilization
(Schlesinger, Mumford, and Glass, 1979) which included the
relationship between psychotherapy and asthma, and also with
alcoholism; the treatment of stuttering (Andrews, 1979); and
sex bias in counseling and psychotherapy (Smith, 1980).
Further studies of psychotherapy outcomes (Landman and Dawes,
1982) and psychotherapy versus placebo (Piroleau, Murdock, and
Brody, 1983), are actually replications of Glass and Smith's
study.

Educational Applications

Quantitative procedures have also been applied to
several issues in education. Hartley's review in 1977, one of
the first applications of meta-analysis in educational
research, compared four methods of individualized instruction
in mathematics. She examined peer and cross age tutoring,
computer assisted instruction, individualized learning
packets, and programmed instructional material from a sample
of 153 studies. Based on Glass's average effect size,
tutoring proved to be the most effective method, in that
approximately 72 percent of the tutored group demonstrated
greater gains than the average student in the control groups.

Since Hartley's review was published, meta-analysis
has been extensively employed in the synthesis of educational
research. Over 45 papers on systematic integration in
education were presented in 1980 alone (Walberg and Haertel,
Some of the areas that have been examined are: the classroom environment; methods of instruction; teacher attitudes; student characteristics; and programme evaluations. One such review investigated the relationships among the socio-psychological environment in the classroom and the cognitive, affective, and behavioural learning outcomes reported in a sample of 12 studies (Haertel, Walberg, and Haeretel, 1979). All of the studies measured the socio-psychological environment by having the students complete a questionnaire on how the classroom was perceived. Typical areas of concern were: the students' enjoyment of their work in class; and were the goals of the class clear to the students. The results from the individual studies were reviewed with meta-analysis, including simple tabulations, correlation-weighted one-way analyses of variance, and multiple regression techniques. The analysis concluded that there was a positive relationship among gains in student achievement, performance, and self-concept and the perceived class cohesiveness, satisfaction, task-difficulty, formality goal direction, democracy and material environment. Whereas a negative relationship was found with cliqueness, apathy, disorganization, and classroom discord. The magnitude of the relationships was influenced by the country in which the sample study was conducted, but was not dependent on the subject taught, the type of learning domain tested, or by statistical adjustments, based on student ability or pre-test score.
The area of instructional methodology has received a great deal of attention from integrative reviewers. Peterson (1980) addressed the issue of whether or not the direct traditional approach to teaching is more effective than the open method. In the studies exploring the influence of classroom style on students' cognitive skills, a computation of effect sizes revealed that students had higher achievement in mathematics and reading when taught with a more traditional approach. Students receiving open teaching demonstrated slightly higher level of creativity than did their peers in a traditional classroom. The affective domain was also examined through the results of effect size calculations. The open classroom fostered more positive levels of self-esteem and attitude towards school than did instruction in a direct approach. Peterson noted that the sample of studies reporting measures of anxiety, curiosity, independence, and attitude toward teacher was too small to provide a stable measure for the mean effect size.

The teaching methods associated specifically with science have also been subjected to a meta-analytic review. Again using average effect sizes, El-Nemr (1979) integrated 59 studies on the outcome of 'inquiry' teaching in biology. When compared to traditional methods, the inquiry approach demonstrated a superiority in all aspects of achievement, including science process abilities, scientific attitudes, critical thinking, and laboratory skills. No relationships associated with the descriptive variables, such as class size,
age of students, and grade, were apparent in this review.

Plaum et al. (1980), used a calculation of effect sizes in combination with a tabulation of a sign test, to investigate the relative merits of various methods of teaching reading. Rather than attempting to use the entire body of 5241 published studies as the sample, a random selection of 97 studies, involving 341 experimental/control group comparisons, that met specific criteria formed the basis of the review. Studies with fewer than five instructional sessions, no control group, or providing no inferential statistics were excluded from the sample. In addition, each study in the sample was weighted by the reciprocal of the number of comparisons it contained. Thus, with only one comparison a study was weighted as one, instances of two comparisons were weighted by one-half. The sign test showed no significant results regardless of the variables being considered. The authors interpreted this as an indication that "the treatment affected participants positively no matter what treatment" (p. 122). Only the use of a sound/symbol blending approach to beginning reading demonstrated a positive gain from the effect size calculation. However, because of the large number of variables, the authors concluded that the apparent benefit of a sound/blending technique, was likely the result of chance.

Other aspects of the instructional process that have been quantitatively analyzed include: modern versus traditional math instruction (Bredderman, 1980), the effects
of teachers using higher cognitive level questions in the classroom (Rousseau and Redfield, 1980), and the relationship between instructional time and student performance (Frederick and Walberg, 1980). Of which, the first two studies used effect size computations as the principle means of review. The latter one described the results of studies using correlation and regression techniques.

The characteristics of teachers and students that are germane to the educational process have also been examined using quantitative review techniques. Smith (1980) considered the research on the effect of teacher expectation on students' abilities, attitudes, and achievements in her review whereas White (1982) specifically investigated the role of students' socio-economic status on their academic performance.

In Smith's review, the results of 47 studies on teacher expectations were analysed in an attempt to resolve the arguments surrounding the "Pygmalion" effect. The review included studies in which students were designated as having an arbitrary intellectual rank, as well as those in which the teachers were given actual information on the students' intellectual abilities. The calculation of effect sizes confirmed Rosenthal and Jacobson's original finding that judgments of students' performance and abilities reflect the labelling information the teachers received (E.S. = .69 averaged over 35 effects). Moreover, it was noted that teachers provided more 'learning opportunities' to students considered to be functioning at a higher intellectual level.
(E.S. = 1.00 averaged over 8 effects). Teacher expectations
had a more moderate effect on student achievement. The
average effect size of .38 obtained from 44 comparisons,
revealed on closer investigation that reading ability,
language arts, and social studies were influenced to a greater
extent than mathematics. Students' intellectual abilities
were minimally influenced by teacher expectations. Two areas
not subject to any prejudice from intellectual labelling were
the teacher's willingness to instruct the students and the
teacher's perceptions of student interest and motivation.

Correlation coefficients from 100 independent studies
were used in White's examination of the relationship between
academic achievement and socio-economic background. In this
review the existence of the relationship was not in doubt.
White instead sought to explain the variability of the
correlation's magnitude and provide a reasonable estimate of
'true or expected' relationship between socio-economic status
and academic performance. White's results ($r = .22$),
corroborated the existence of the relationship, but at a lower
level than most previous research which had reported
correlation coefficients as high as .80. By using individuals
as the unit of analysis, White was also able to discern 'age'
and 'subject' trends in the relationship. The older the
student, the smaller the correlation ($r$ decreasing from .25 in
the early grades to .15 at the end of high school). Verbal
ability correlated higher ($r = .49$) with socio-economic status
than did math achievement measures ($r = .24$). According to
White the variation among reported correlations was attributable to differences in the unit of analysis, type of achievement measure, socio-economic indicator (such as parental education, occupation, income, or school resources, and home atmosphere), age of students, and year of study. Almost 30 per cent of the variance among studies resulted from the manner in which the students' socio-economic level was measured.

Two quantitative reviews, conducted by Kavale (1982, 1981), have also utilized correlational information to study reading achievement. In one review (1981), the findings from 106 studies were used to investigate the relationship between reading ability and auditory perceptual skills. The other review (1982) examined the influence of visual perceptual ability on reading proficiency through an integration of the results from 161 studies. The results of these two reviews showed that both auditory and visual skills are factors predicting reading ability.

Programme evaluation is another area of education in which researchers have employed quantitative review techniques. One example is the review of nutrition-education programmes by Levy, Iverson, and Walberg (1980). Their preliminary quantitative review, which was limited to a sample of 6 studies, demonstrated some degree of 'knowledge gain' in terms of nutrition awareness. The effect sizes ranged from a low of .48 to a high of .82, indicating that the magnitude of the knowledge gains varied across different classroom
situations. As well as attesting to the efficacy of nutrition-education programmes, these results, in conjunction with the qualitative information from an additional 12 studies, identified areas of weakness in the design of the evaluations. Variables associated with the programmes, the quantity of instruction, and the teaching styles were the main factors under investigation. Evaluators tended to ignore the influence of the students' home and classroom environments on the acquisition of nutrition knowledge. Levy et al. observed that the results of a 'comprehensive comparison', such as those obtained in quantitative reviews, "suggest modifications which can strengthen future research designs" (p. 46).

Quantitative review procedures have also been used to examine evaluation techniques. One such study investigated whether or not student ratings of instruction are a valid means of evaluating courses. Cohen (1981) synthesized the findings of 41 independent studies, which included data on 68 separate multi-sectioned courses, into one body of evidence that provided strong support for the use of student ratings of teaching effectiveness.

Section Summary

Primarily, the sample of studies presented in this section demonstrates both the range of topics and the depth of information that can be obtained from a quantitative review.
On closer inspection, however, this section has revealed that there is a precedent for using quantitative techniques in the fields of education and evaluation and that the vast majority of reviewers directed their attention to new areas of investigation. In fact there were only two replications in the sample. It would therefore appear that there is a need to channel review efforts into the re-examination of existing studies and the verification of the procedures.
METHOD

Defining the Sample

Since the original quantitative analysis by Carlberg and Kavale was published in 1980, it was not surprising to find that all of the documents in their sample were written prior to 1977. Rather than exactly duplicating Carlberg and Kavale's work, this replication extended the sample to include more recent findings. Standard literature-search procedures, assisted by computer searches of Psychological Abstracts and ERIC, were used to identify studies pertaining to the topic of mainstreaming exceptional children. Once located, each study had to meet all of Carlberg and Kavale's criteria before it could be added to the sample. The four criteria employed by Carlberg and Kavale are as follows:

1. The study had to investigate educational placement for an identifiable category of exceptionality.

2. The study had to examine special class placement.

3. The study had to include a comparison group (e.g., regular class) even if the comparison group was the same as the special class group (as in a correlated group pretest-posttest design).

4. The study had to report results in a fashion that could be translated into a form appropriate for meta-analysis. (p. 298)
The fourth criterion does not restrict the sample to only those studies that reported means and standard deviations. Almost any type of statistical information that is reported in a study can be converted into a form suitable for an effect size analysis. Techniques are available to glean the necessary information from studies that supply either $t$ or $F$ statistics and the sample sizes, or the levels of significance (alpha) and the sample sizes. Only the studies that simply state that one group's performance exceeded that of another, or provide only the alpha levels must be excluded.

An additional constraint on the selection of studies for the sample was the 'locateability' of any given document. Papers, theses, and dissertations that were not available through either the university or public library systems were excluded.

Some authors, notably Eysenck (1978) and Mansfield and Busse (1977), have suggested that the studies should also be screened on the basis of research quality prior to their inclusion in the review sample. The rationale for this position is that effect sizes derived from a combination of rigorous and nonrigorous research threatens the validity of the quantitative review procedure (Stock, Okun, Haring, Miller, Kinney and Ceuvorst, 1982). It can be argued, however, that the deletion of nonrigorous studies from a review on the basis of arbitrary a priori decisions poses more of a threat to the validity than their inclusion. Such judgments introduce
unwarranted subjectivity into the reviewer's sample selection procedure. Reviewers usually neglect to mention the means used to determine which studies are methodologically inadequate nor do they provide any evidence that the deficiencies had biased the results (Glass, McGaw and Smith, 1981). Furthermore, "since most studies have some weaknesses, these reviewers often end up reporting inferences based on only a few studies" (Hunter, Schmidt, and Jackson, 1982, p. 151). The optimal review procedure includes all available studies in the sample, and if there is some reason to suspect that the rigor of research design has prejudiced the findings, the characteristics associated with the methodological inadequacies can become variables that are examined in the analysis. Therefore, as in the original review, the rigor of research design in an individual study was not a criterion for inclusion in the replication sample.

Sample Size

Although Carlberg and Kavale reported using 50 studies in their sample, only 48 were identifiable in the references which accompanied the article. The studies used in the analysis were "represented in the reference list with their ESs following the citation" (p. 305), and as such were easy to identify. From these 48 studies 26 were located for the replication's sample. Another 4 were added to the sample as a result of the updated literature search. The studies included
the review sample are listed in a separate section of the
bibliography.

Method of Coding Individual Studies

A statistical integration of a body of literature is
based on a transformation of both the characteristics and
the findings from each individual study into a format suitable
for the analysis. Due to the importance of this initial step,
the method of coding of the two types of requisite information
from each study will be discussed separately.

Characteristics of Studies

Carlberg and Kavale chose to examine four groups of
study characteristics in their review: Subject, Treatment,
Design, and Outcome. These four groups included both
substantive and methodological characteristics (Glass, 1981).
The categories of 'Subject' and 'Treatment' can be classified
as substantive features in that they "are those
characteristics of studies that are specific to the problem
studied" (Glass, McGaw, and Smith, 1981, p. 78), whilst the
'Design' and 'Outcome' categories consider the more general
methodological attributes of the studies. The examination of
these two main types of characteristic provides information on
the overall body of literature as well as revealing the
individual variability among studies.
A brief outline of the conventions used for coding studies in the replication are presented in the following paragraphs. Features that were not mentioned, or else did not occur in sufficient number, in the original analysis but have been added to the replication are identified with the symbol (*). The complete codebook is included in Appendix A.

Substantive

Major category of exceptionality. The educational/diagnostic label used to identify the student's exceptionality and/or determine classroom placement: 1) Educable Mentally Retarded (EMR), 2) Slow-learning (SL), which encompasses poor readers, slow learners, and the perceptually handicapped, 3) Behaviourally Disordered (BD), 4) Emotionally Disturbed (ED).

Categorical combination if more than one category. Students who could not be diagnosed/classified on the basis of a single exceptionality were denoted by a combination of exceptionalities, such as BD/ED.

Mean IQ, special and regular classes. The average intelligence scores of the students in the regular and special classes in conjunction with the source of the IQ scores were noted i.e. whether the score was assessed by the experimenter, inferred from the students' placement in an exceptional class, or reported from another source.

Mean age, special and regular classes. A record was made of the average age, in months, of the students in each class.

Percent male in total sample. For each study the percentage
of the total sample who were male students was noted. In this way, the outcome data could be examined for any confounding due to differential learning between the sexes.

Socioeconomic status. Based on information provided in the studies, each class was identified as having either a high, middle, or low socioeconomic status (SES).

Type of treatment (*). The type of treatment was classified as follows: 1) physical/perceptual practice, 2) special reading instruction, 3) social/psycho-drama, 4) resource room, 5) preschool, 6) academic remediation, 7) behavioural modification.

Duration of treatment. The duration of special class placement was recorded in terms of number of weeks.

Training level of instructors. The teachers' educational background was coded: 1) classroom aide, 2) undergraduate, 3) graduate, 4) regular classroom teacher, 5) special education training/experience.

Experience of instructors (*). The number of years of actual teaching experience was recorded.

Class sizes. The number of students in each special and regular class was noted.

Student/instructor ratio (*). The student/teacher ratio was computed for each special and regular class.

Methodological

Publication form (*). Each study was classified on the basis of the form in which it appeared: 1) journal article, 2) book,
or 3) thesis.

Publication date (*). The date associated with the publication of each study was recorded.

Solicitation of subjects. The students in the special classes were classified according to whether: 1) selected by experimenter, 2) referred/recommended to a special class, 3) already placed in a special class.

Solicitation of instructors (*). For each study the manner of obtaining the instructors participation was noted.

Group assignment of students. Random assignment of students to the classes would be the ideal means of ascertaining the effects of special education. It is not always, however, a feasible or ethical means of determining a child's educational experience for the duration of a study, and consequently it is rarely employed. Designs that use matched-pairs of subjects and 'convenient samples' provide viable alternatives. Every study was coded according to the type of group assignment that was used.

Assignment of instructors. The coding of this study characteristic followed the same outline as Group assignment of students described previously.

Percent attrition, special and regular classes. The percent of attrition was recorded for each class. In situations where these figures were not reported, they were determined from a comparison of the degrees of freedom in the posttest analysis with the original group size.

Type of comparison group. The comparison group was coded as
either 1) control or 2) a second treatment.

**Number of threats to the internal validity** The internal validity of a study is the extent to which the extraneous variables (as described by Borg and Gall, 1983) were controlled by the researcher. The number of extraneous variables that could have confounded the results was listed for each study.

**Type of outcome measure.** The type of outcome measure used to ascertain changes in student performance were coded as: 1) standardized achievement test, 2) *ad hoc* instructor/experimenter designed achievement test, or 3) observed student behaviour.

**Domain of outcome measure.** The area of student performance, such as 1) cognitive, 2) achievement, 3) aptitude, 4) social/personality skills, and 5) behaviour, that a measure investigated was also noted.

**Posttreatment delay of measurement.** The number of weeks from the conclusion of instruction to the measurement of outcomes was recorded for each class.

**Reactivity of outcome measure.** The obtrusiveness of the measure was noted in terms of whether it had a low, medium, or high influence on the subjects' behaviour.

**Blindedness of outcome measure.** The rater of the outcome measures used in a study was coded as being: 1) blind to the subject's status within the experiment, or 2) non-blind, that is to say, knowing whether a subject was in the comparison or treatment group.
Quantifying Findings from Studies

This section describes the methods used to obtain the information required from each study to complete the calculations of Glass's effect sizes, Stouffer's Z, and a fail-safe N. Although there are three quantitative techniques within this study, only two main types of information are needed. The effect size calculations utilize some form of the mean and standard deviation, while the Stouffer's Z and the fail-safe N are based on Z scores derived from the probabilities that are reported in the sample studies. The methods of quantifying the data will, therefore, be discussed separately in this section.

Effect Size Calculations

The most informative and straightforward measure of effect size is the mean difference between the experimental (X_exp) and control (X_con) groups divided by the within group standard deviation (SD_w), assuming that there is homogeneity of variance between groups. This can be expressed as:

\[
E.S. = \frac{\bar{X}_{exp} - \bar{X}_{con}}{SD_w}
\]

Rather than use this precise version of computing E.S., Carlberg and Kavale substituted the control group's standard deviation (SD_con) in place of SD_w. This modification, which was first suggested by Glass (1977) for use in situations where the within group variance is
heterogeneous, provides an E.S. that is applicable to many situations and readily understood. In this case, as well as Carlberg and Kavale's, the special education classes are considered to be the treatment groups, with the regular classes assuming the role of the control group. When the control group's variability is used in the E.S. calculation (E.S.C), the resulting value is the amount of improvement that must be gained by the comparison group to reach parity with the treatment group. If a given study investigates two (or perhaps more) special classes with a regular class and the variances for the three groups differ, the use of E.S.C provides for a direct comparison between the three teaching methods. For example:

<table>
<thead>
<tr>
<th>Method A</th>
<th>Method B</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Means</td>
<td>50</td>
<td>40</td>
</tr>
<tr>
<td>Standard Deviations</td>
<td>6</td>
<td>3</td>
</tr>
</tbody>
</table>

\[
E.S.\text{Treatment A} = \frac{50 - 45}{5} = \frac{5}{5} = 1
\]

\[
E.S.\text{Treatment B} = \frac{40 - 45}{5} = \frac{-5}{5} = -1
\]

Subjects receiving treatment A performed at a level one standard deviation above the average control group subject, who likewise, functioned one standard deviation above the
average subject receiving treatment B.

The preceding example also illustrates that there may be more than one E.S. associated with a specific study investigating several treatment groups. Multiple effect sizes are also present in studies that have the combination of only one treatment group with a variety of outcome measures. In this replication, the sample of 30 studies yielded 198 effect sizes.

Not all of the effect sizes could be calculated with the simple formula that has been discussed. The vast majority of studies in the sample neglected to report group means and standard deviation. Consequently, most E.S.s were obtained through the following complex manipulations of the available data.

Given: \(t\) statistic

sample size

When the \(t\) statistic and sample size is known for the treatment and control groups, E.S. can be derived in the following manner:

\[
t = \frac{\bar{x}_{\text{exp}} - \bar{x}_{\text{con}}}{\sqrt{\frac{(n_{\text{exp}}-1)S_{\text{E}}^2 + (n_{\text{con}}-1)S_{\text{C}}^2}{(n_{\text{exp}}-1) + (n_{\text{con}}-1)} \left[ \frac{1}{n_{\text{exp}}} + \frac{1}{n_{\text{con}}} \right]}}
\]

assuming that

\[
t = \frac{\bar{x}_{\text{exp}} - \bar{x}_{\text{con}}}{\sqrt{\frac{S^2(n_{\text{exp}} + n_{\text{con}} - 2)}{(n_{\text{exp}} + n_{\text{con}} - 2)} \left[ \frac{1}{n_{\text{exp}}} + \frac{1}{n_{\text{con}}} \right]}}
\]
\[ t = \frac{\bar{x}_{\text{exp}} - \bar{x}_{\text{con}}}{\sqrt{\frac{1}{n_{\text{exp}}} + \frac{1}{n_{\text{con}}}}} = \frac{\bar{x}_{\text{exp}} - \bar{x}_{\text{con}}}{S} = E.S. \]

It should be noted that this is an approximation of E.S. since the denominator is a pooled estimate of the control group's standard deviation.

Given: sample sizes
level of significance for \( t \) tests

A table of critical \( t \) values is consulted to find the actual value of \( t \) that corresponds to the stated level of significance. One-tailed tests are assumed unless otherwise stated. The resulting figure is then used in the following equation to estimate E.S.:

\[ E.S. = t \sqrt{\frac{1}{n_{\text{exp}}} + \frac{1}{n_{\text{con}}}} \]

This is a conservative estimate of the actual E.S.

This is demonstrated as follows:

If there are 10 subjects in the treatment group and 20 in the control, the values of \( t \) are:

\[ t_{01} = 1.701 \quad \text{and} \quad t_{02} = 2.048 \]

It can be assumed that stated by the author that \( p < .01 \), \( t \) does not exceed the value of 2.048. Therefore, the value of E.S. is within the range defined by:

\[ 1.701 \sqrt{\frac{1}{10} + \frac{1}{20}} \leq E.S. \leq 2.048 \sqrt{\frac{1}{10} + \frac{1}{20}} \]

\[ .66 \leq E.S. \leq .79 \]
The smaller limit of E.S. is used for any comparisons in order that any errors will be on the conservative side.

If the group differences are non-significant, yet it is observed that one mean is higher than the other, a standard is used in the estimation of E.S.

If group differences are non-significant, and there is no mention of any differences between the means, E.S. is assigned a value of zero.

Given: \( F \) in a two group comparison or
\( F \) in a three (or more) group comparison group means

In view of the relationship between \( F \) and \( t \),
\[ t^2 = F \]

Techniques similar to those described for \( t \) can be used.

Given: sample sizes significance levels for non-parametric statistics

The degrees of freedom associated with the statistic are noted. A \( t \) table is then entered at the appropriate level of significance for the non-parametric statistic and the specified degrees of freedom. The resulting estimate of E.S. is extremely conservative, in that, the critical \( t \) is an underestimate of E.S. that has been underestimated from the non-parametric findings. This is may be attributed to the fact that non-parametric tests are not as powerful as the parametric tests.

Given: percentage statement of results

If the author reports the outcomes in percentage terms, such as 50% of the treatment group subjects showed an improvement, while 20% of the control group demonstrated similar gains, a probit analysis is used to estimate an E.S. This type of analysis assumes that the results of both groups are normally distributed in the manner illustrated in Figure 2.
Figure 2. Model for the Estimation of E.S. from Percentage Data.

$C_x =$ point defining improvement
Using this assumption the following derivation provides an estimate of \( E.S. \):

\[
Z_{\text{exp}} = \frac{C - \bar{x}_{\text{exp}}}{S_{\text{exp}}}
\]

and

\[
Z_{\text{con}} = \frac{C - \bar{x}_{\text{con}}}{S_{\text{con}}}
\]

and if \( S_{\text{exp}} = S_{\text{con}} \)

then

\[
Z_{\text{con}} - Z_{\text{exp}} = \frac{\bar{x}_{\text{exp}} - \bar{x}_{\text{con}}}{S} = E.S.
\]

The values for \( Z_{\text{con}} \) and \( Z_{\text{exp}} \) are standard normal deviates that can be located from normal curve tables. Table 1 provides the probit transformations for the most common percentages reported in studies.

Derived Z Scores

As its name implies, the Stouffer Z method is based on \( Z \) scores, otherwise known as standard normal deviates. These scores were obtained from the sample studies and used to calculate an overall Stouffer's \( Z \) in the following five steps:

1) The direction for testing the hypothesis should be chosen.

2) Either, record the exact probability (\( p \)) level or compute the exact \( p \) value of the test statistics, for every outcome measure in each study in the review. Any reported \( ps \) associated with a two-tailed test must be halved before proceeding to the next step, and if the \( p \) value is a disconfirmation of the hypothesis it is subtracted from the sum of the
<table>
<thead>
<tr>
<th>Percent Control</th>
<th>Percent Experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.36</td>
</tr>
<tr>
<td>0.10</td>
<td>0.44</td>
</tr>
<tr>
<td>0.15</td>
<td>0.52</td>
</tr>
<tr>
<td>0.20</td>
<td>0.65</td>
</tr>
<tr>
<td>0.25</td>
<td>0.79</td>
</tr>
<tr>
<td>0.30</td>
<td>0.91</td>
</tr>
<tr>
<td>0.35</td>
<td>1.04</td>
</tr>
<tr>
<td>0.40</td>
<td>1.17</td>
</tr>
<tr>
<td>0.45</td>
<td>1.30</td>
</tr>
<tr>
<td>0.50</td>
<td>1.43</td>
</tr>
<tr>
<td>0.55</td>
<td>1.56</td>
</tr>
<tr>
<td>0.60</td>
<td>1.69</td>
</tr>
<tr>
<td>0.65</td>
<td>1.82</td>
</tr>
<tr>
<td>0.70</td>
<td>2.09</td>
</tr>
<tr>
<td>0.75</td>
<td>2.33</td>
</tr>
<tr>
<td>0.80</td>
<td>2.59</td>
</tr>
<tr>
<td>0.85</td>
<td>2.83</td>
</tr>
<tr>
<td>0.90</td>
<td>3.07</td>
</tr>
<tr>
<td>0.95</td>
<td>3.34</td>
</tr>
</tbody>
</table>

Table 1. Probit Transformations Used in Estimating an Effect Size
3) A Z score is computed for each p value, or in some simple cases the p may be located in a table of standard normal deviates (Z tables) and the associated Z score can be read directly from the table.

4) Sum the Z scores across all studies in the sample. The resulting figure is then divided by the square root of the number of findings involved to provide the standard normal deviate \( Z_{mn} \). The following equation illustrates the calculation:

\[
Z_{mn} = \frac{Z_{m1} + Z_{m2} + \ldots + Z_{mn}}{(N_m)^{1/2}}
\]

5) The overall Z score, \( Z_{mn} \), is then referred back to the Z tables and the appropriate probability level is noted.

The final probability value provided by this procedure indicates the likelihood that the results from all the reviewed studies were nothing more than a chance occurrence. Since some studies fail to mention specific t, F, or chi-squared scores, their Z scores are derived from nominal instead of exact p values. Consequently, the overall results of the analysis are underestimated.

The basic method of computing Stouffer's Z can be modified to compensate for the quality of research in the review, as well as the size of the sample used in each individual study. Studies that are deemed to be more rigorous can be weighted more heavily than those that are felt to be unsound, while results reported from studies with relatively small samples can be weighted in a way that minimizes their effect in the overall calculation.

Neither of these two modifications were incorporated
into this analysis. The disadvantages associated with the reviewer determining the influence of research quality on a study's outcome for the purpose of choosing the sample (as discussed earlier in this chapter) are also valid in this situation. Although the second modification has a surface appearance of a 'scientific' foundation, it is actually a re-statement of the first suggested alteration. If all matters of research quality are held constant, it can be mathematically demonstrated, that large sample sizes reduce the probability of forming erroneous conclusions. When comparisons are performed among 'real-life' studies, however, this observation is not always true. It is quite possible for a well designed, small study to provide more rigorous information than an extremely large, but flawed, study. The decision to give more weight to findings based on large samples, is in fact, a judgment of research quality. As such it is a procedure that is best omitted from a quantitative review. Furthermore, any effects of sample size and research quality are determined through the analysis of the E.S. data and are considered in the interpretation of Stouffer's Z.

The third step of this study is a calculation of a fail-safe $N$. It will only be included in the analysis if the overall probability level obtained from the Stoufer's $Z$ is significant.

The equation used for the calculation of the fail-safe $N$ in this analysis differed slightly from the version described by Cooper (1979). In his paper the following
equation is given:

\[ N_{a_{.05}} = \frac{Z_{a1} + Z_{a2} + \ldots + Z_{an}}{1.645} - N_a \]

where \( N_{a_{.05}} \) represents the number of additional studies, with summed \( Z \) scores totaling zero, that are required to increase the overall probability level above .05, the value 1.645 is defined as "the \( Z \) score associated with \( p < .05, \) two-tailed" (p. 135), and the remaining terms are defined as in the Stouffer \( Z \) method. The figure of 1.645 is not the value of \( Z \) at the .05 level for a two-tailed test. Consulting tables of standard normal deviates revealed that the \( Z \) score at the .05 level of a one-tailed test is 1.645. The corresponding two-tailed value is 1.96. Since the fail-safe \( N \) should be calculated with \( Z \) scores from a two-tailed test, the value of 1.96 was substituted into the equation for this study. Cooper (1979) does note that other significance levels can, at the reviewer's discretion, replace the value specified in the formula.

Specific Coding Concerns

The coding conventions employed by Carlberg and Kavale failed to mention how non-significant results and outcomes of negative attributes were quantified in their review. Since the reviewer can interpret these two types of results in a variety of ways, the specific conventions used in this study
are described in the following paragraphs.

Quantifying Non-significant Results

In situations where the author only reported that an outcome had a non-significant result, an E.S. of zero was assumed and used in the analysis. For the calculation of the Stouffer Z score, the probability of such an outcome was deemed to be .50, which in turn, implied an associated Z score equal to zero. If an author stated that the results were non-significant, yet also specified exact figures, such as the means and standard deviations, or the value of t then an exact calculation of the E.S. and probability were included in the review. This use of exact results leads to a more precise evaluation of the effects of special as opposed to regular education.

Negative Outcome Measures

The calculation of the E.S.s is defined by Carlberg and Kavale in terms whereby "a positive E.S. favors the special class; a negative E.S. favors the regular class" p. 299). The majority of outcome measures used to assess children's performance are designed so that a higher score equates to a higher level of functioning. Thus, a high score by the special group, and a lower score by the regular group, would provide a positive E.S. which favours special education. On a scale that measures a negative attribute, typically
rejection or anti-social behaviour, a higher score corresponds to a lower level of functioning. In such a situation, a high score by the special group, and a lower score by the regular group, while providing a positive E.S., actually means that the special group were more rejected or more anti-social than the regular class; hardly evidence in favour of special education.

In order to conform to Carlberg and Kavale's position, the 'sign' of any E.S. calculated for a negative attribute measure was reversed in this study. Consequently, an E.S. arising from a high score by the special class and a lower score by the regular class, was assigned a negative, rather than a positive, value. By following this convention, all positive E.S.s favoured the special class and all negative E.S.s favoured the regular class.

Design of Analysis

The preliminary analysis consisted of an examination of the distribution of various study characteristics. At this point, variables which were not recorded in sufficient instances among the sample, or for which there appeared to be only one category of response were eliminated from the study.

The main analysis, since one purpose of this study was to investigate the replicability of the quantitative review process, was patterned after the procedures used by Carlberg and Kavale. The effect sizes obtained from each study within
the sample were the basis of a calculation of an overall mean, range, and standard deviation that could be directly compared with Carlberg and Kavale's findings. Variables that survived the preliminary analysis were either subjected to an analysis of variance (ANOVA), which was followed in some instances by a Scheffe test of comparisons between means, and/or correlated with the effect sizes. These results were then used in additional comparisons with the original quantitative review.

The results of the Stouffer Z and the fail-safe N calculation were considered in the second stage of the analysis. This stage consisted of a comparison of the types of information obtained from each method. Thereby, demonstrating a 'package' of techniques for use in a quantitative review that would provide the most beneficial information for educational researchers and administrators, meeting another goal of this study.

Computer Programme for the Analysis

Two BASIC programmes, listed in Appendix B, were written to meet the specific needs of the analysis. They were designed to calculate the individual effect sizes from the outcome data presented in the sample studies as well as performing the statistical analysis. All the equations used in the statistical analysis were taken from Linton and Gallo (1975). The procedures for deriving the E.S.s followed the methodology developed by Glass, McGaw, and Smith (1981).
If the calculation of E.S.s was selected, the programme requested the user to identify what type of data would be used, for example means and standard deviations, ts, or Fs. On the basis of the response, the programme would request that the appropriate information be entered, perform the calculation, and show the E.S. on the screen. No allowances for data storage were incorporated into this programme. Information was copied from the screen onto coding forms, and subsequently transferred into the data storage file.

The programme for the statistical analysis was divided into four sections. The user could select 1) a computation of means, standard deviations, t tests, 2) a printout of the information obtained in 1), 3) Pearson product moment correlation, 4) ANOVA. For each selection, the user identified which variables were of interest, and how the data was to be grouped. The data was obtained from a separate file, in which the information was stored in a card-like format.

Summary

The analysis employed in this study is designed to facilitate comparisons with Carlberg and Kavale's original quantitative review of mainstreaming. The results of such comparisons provide an indication of the replicability and
subjectivity of the quantitative approach to reviewing.

Additional information, that can be of assistance to programme administrators, is furnished in the obtained Stouffer's Z and fail-safe N.
RESULTS

Preliminary Analysis

The results of the initial frequency distributions indicated that several variables were not suited for further analysis. Since only one study in the total sample specifically defined its subjects as having more than one educational handicap, there was no need to have a distinct 'combined exceptionality' variable. The four levels of the 'combined exceptionalities' variable were incorporated, as one group, into the 'category of exceptionality' variable for the main analysis.

The percentage of boys in each study's sample was originally recorded to assist in the determination of any sex-biasing in the outcomes. Of the 30 studies in the sample, 18 failed to report the number of boys and girls involved, 8 used equal proportions, and 4 studies had sample groups in which there were almost twice as many boys as girls. In the last group of studies, where biasing would be anticipated, there were no significant interactions between the educational treatment and sex. The one study which obtained significant interactions between treatment and sex, had an equally proportioned sample of 'normal' children who rated the social acceptability of their integrated and segregated EMR peers. The percentage of boys in each study's sample was, therefore, eliminated from the main analysis due the combination of
insufficient information and no apparent biasing of the results.

The variables that examined students' treatment status, instructors' training level, instructors' teaching experience, student/instructor ratio, solicitation of both subjects and instructors, and assignment of students and instructors were also excluded from the main analysis. These variables could potentially influence the classes' performance and the associated outcome E.S.s, and as such warranted inclusion in this review. Unfortunately, most researchers failed to provide the requisite information to code these variables (for example, only three of the sample studies described either the instructors' teaching experience or training level).

Although every study in the sample detailed the domain of student performance being investigated and the types of measures being used, none of the authors clearly stated whether the person/people conducting the outcome measurements were aware of the groups' experimental status (control or experimental) at the actual time of testing. For the vast majority of studies, no reference was made to the people involved in the measurement procedure. Without this knowledge, even the assumption that any teachers or researchers who were responsible for the outcome measurement would have been 'non-blind', cannot provide any information. In a few instances where school psychologists, as opposed to the experimenters and teachers, were given the task of assessing the students' performance, no additional information
describing their 'blindness' was included. Consequently, the variable examining the blindedness of the outcome measure was also omitted from the main analysis.

One variable that became meaningless in this review was 'publication form'. Although studies from books, dissertations, and theses, were eligible for inclusion in this review's sample, none were present. The entire sample consisted of research that had been published in journals. Analysis of a variable that has only one category of response cannot provide additional insight and its continued examination would have become a futile exercise.

The last variable excluded from the analysis at the preliminary stage was 'reactivity of outcome measure'. Descriptions of both the time and setting of the outcome assessment were either absent or else sufficiently vague that the information couldn't be accurately coded.

Data from All Studies

The 30 studies in this review documented the performance of over 2,500 students from special and regular classrooms. These students had an average I.Q. of 78.3, and although the mean age was 11.35 years, the sample actually ranged from pre-schoolers through to 21 year-old adults. The average duration of special educational intervention was 51 weeks.
The aggregated results across the 30 studies in the sample yielded 198 effect sizes that had a mean of 0.105. A mean effect size (M.E.S.) of this magnitude indicates that on the average, students placed in some type of special educational setting were functioning at a level approximately 3.5% above their counterparts being maintained in a regular classroom. Table 2 presents the average E.S. calculated for each study in the review, and where possible, also provides the E.S.s reported by Carlberg and Kavale (1980). A calculation of inter-rater reliability based on a Pearson product moment correlation demonstrated that a significant positive relationship (r = .338, df = 27, p = .01) exists between the 29 average E.S.s common to this study and Carlberg and Kavale's. When the level of inter-rater agreement between the two samples was examined, with the criterion for matching being set at the second decimal place, slightly more than 41% of the E.S. pairs were the same. A comparison between the two studies using all the E.S.s noted in Table 2 failed to disclose any significant differences (t = 1.0186, df = 94).

The individual outcome E.S.s ranged from -1.571 to 1.71 with a standard deviation of .612. While a majority of 63% of the E.S.s were above zero, only 37% (n = 74) of the E.S.s actually evidenced the superiority of special education over regular classroom placement. The 95% confidence interval surrounding the mean extended from approximately .020 to .190. Since the limits of the confidence interval exclude zero, it can be argued that the mean should be considered significantly
TABLE 2. Description of Studies in the Analysis

<table>
<thead>
<tr>
<th>Author</th>
<th>Exceptionality</th>
<th>Placement</th>
<th>Calculated E.S. (a)</th>
<th>Carlberg &amp; Kavale E.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bersoff, et al.</td>
<td>SL/BD</td>
<td>Regular class</td>
<td>-0.691(2)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td></td>
<td>with tutoring</td>
<td>1.145(1)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Special class</td>
<td>-1.072(2)</td>
<td>-0.55(2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.342(1)</td>
<td></td>
</tr>
<tr>
<td>Blatt</td>
<td>EMR</td>
<td>Special class</td>
<td>0.763(3)</td>
<td>0.76(3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-0.091(2)</td>
<td>-0.09(2)</td>
</tr>
<tr>
<td>Budoff &amp; Gotlieb</td>
<td>EMR</td>
<td>Regular class</td>
<td>0.396(3)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td></td>
<td>with tutoring</td>
<td>0.526(2)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.676(1)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Special class</td>
<td>0.142(3)</td>
<td>0.04(3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.261(2)</td>
<td>0.29(2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.085(1)</td>
<td>--</td>
</tr>
<tr>
<td>Carroll</td>
<td>EMR</td>
<td>Special class</td>
<td>-0.710(3)</td>
<td>-0.71(3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-0.684(2)</td>
<td>-0.35(2)</td>
</tr>
<tr>
<td>Cawley &amp; Goodman</td>
<td>EMR</td>
<td>Special class</td>
<td>-0.860(1)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Special class</td>
<td>0.255(1)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td></td>
<td>with tutoring</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gampel, et al.</td>
<td>EMR</td>
<td>Regular class</td>
<td>-0.270(3)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td></td>
<td>with tutoring</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Special class</td>
<td>0.015(3)</td>
<td>-0.55(3)</td>
</tr>
<tr>
<td>Goodman, et al.</td>
<td>EMR</td>
<td>Special class</td>
<td>0.304(3)</td>
<td>0.20(3)</td>
</tr>
</tbody>
</table>
## Table 2. (cont.)

<table>
<thead>
<tr>
<th>Author</th>
<th>Exceptionality</th>
<th>Placement</th>
<th>Calculated E.S.(a)</th>
<th>Carlberg &amp; Kavale E.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gottlieb &amp; EMR</td>
<td>Special class</td>
<td>0.131(3)</td>
<td>0.00(3)</td>
<td></td>
</tr>
<tr>
<td>Budoff</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gottlieb &amp; EMR</td>
<td>Special class</td>
<td>0.00(3)</td>
<td>-0.27(3)</td>
<td></td>
</tr>
<tr>
<td>Davis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gottlieb, et al.</td>
<td>EMR</td>
<td>Special class</td>
<td>0.958(3)</td>
<td>0.95(3)</td>
</tr>
<tr>
<td>Guerin &amp; Szatlocky</td>
<td>EMR</td>
<td>Special class</td>
<td>-0.506(4)</td>
<td>--</td>
</tr>
<tr>
<td>Haring &amp; Krug</td>
<td>Special Class</td>
<td>-0.405(3)</td>
<td>-0.571(2)</td>
<td></td>
</tr>
<tr>
<td>Jordan</td>
<td>EMR</td>
<td>Special class</td>
<td>-0.124(2)</td>
<td>0.21(2)</td>
</tr>
<tr>
<td>Jordan &amp; DeCharms</td>
<td>EMR</td>
<td>Special class</td>
<td>0.480(3)</td>
<td>0.11(3)</td>
</tr>
<tr>
<td>Kern &amp; Pfaeffle</td>
<td>EMR</td>
<td>Special class</td>
<td>0.510(3)</td>
<td>0.45(3)</td>
</tr>
<tr>
<td>Lapp</td>
<td>EMR</td>
<td>Special class</td>
<td>-0.166(3)</td>
<td>-0.16(3)</td>
</tr>
<tr>
<td>Mayer</td>
<td>EMR</td>
<td>Special class</td>
<td>-0.242(3)</td>
<td>-0.25(3)</td>
</tr>
<tr>
<td>Myers</td>
<td>EMR</td>
<td>Special class</td>
<td>0.455(3)</td>
<td>0.45(3)</td>
</tr>
</tbody>
</table>
Table 2. (cont.)

<table>
<thead>
<tr>
<th>Author</th>
<th>Exceptionality</th>
<th>Placement</th>
<th>Calculated E.S. (a)</th>
<th>Carlberg &amp; Kavale E.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Porter &amp; Milazzo</td>
<td>EMR</td>
<td>Special class</td>
<td>0.253(3)</td>
<td>--</td>
</tr>
<tr>
<td>Rubin, Krus &amp; Balow</td>
<td>EMR</td>
<td>Special class</td>
<td>-0.426(3)</td>
<td>-1.31(2)</td>
</tr>
<tr>
<td>Sabatino</td>
<td>SL</td>
<td>Regular class</td>
<td>0.265(2)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td></td>
<td>with tutoring</td>
<td>0.479(1)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Special class</td>
<td>0.420(2)</td>
<td>0.86(2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.815(1)</td>
<td>0.28(5)</td>
</tr>
<tr>
<td>Schurr, et al.</td>
<td>EMR</td>
<td>Special class</td>
<td>0.630(3)</td>
<td>0.60(3)</td>
</tr>
<tr>
<td>Sheare</td>
<td>EMR</td>
<td>Special class</td>
<td>-1.059(3)</td>
<td>-1.06(3)</td>
</tr>
<tr>
<td>Spollen &amp; Ballif</td>
<td>EMR</td>
<td>Special class</td>
<td>0.079(1)</td>
<td>--</td>
</tr>
<tr>
<td>Trimble</td>
<td>SL</td>
<td>Special class</td>
<td>-0.045(2)</td>
<td>-0.05(2)</td>
</tr>
<tr>
<td>Trippi</td>
<td>EMR</td>
<td>Special class</td>
<td>0.527(3)</td>
<td>0.53(2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.53(5)</td>
</tr>
<tr>
<td>Vacc</td>
<td>ED</td>
<td>Special class</td>
<td>-0.095(4)</td>
<td>-0.04(2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.189(2)</td>
<td>0.58(3)</td>
</tr>
<tr>
<td>Vacc</td>
<td>ED</td>
<td>Special class</td>
<td>-0.491(4)</td>
<td>-0.09(3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.033(3)</td>
<td>-0.70(2)</td>
</tr>
</tbody>
</table>
Table 2. (cont.)

<table>
<thead>
<tr>
<th>Author</th>
<th>Exceptionality</th>
<th>Placement</th>
<th>Calculated E.S.(a)</th>
<th>Carlberg &amp; Kavale E.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walker</td>
<td>'EMR</td>
<td>Special class</td>
<td>-0.121(3)</td>
<td>0.01(3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-0.404(2)</td>
<td>-0.27(2)</td>
</tr>
<tr>
<td>Ito</td>
<td>SL</td>
<td>Special class</td>
<td>0.224(2)</td>
<td>--</td>
</tr>
</tbody>
</table>

(a) The number in parentheses refers to the domain of outcome measure as follows: 1) cognitive; 2) academic/achievement; 3) social/personality; 4) behaviour; 5) other.

Different from zero and in support of special education being a better alternative to regular classroom placement for exceptional children. It should be noted, however, that the confidence intervals are an approximation due to the fact that not all E.S.s were statistically independent. These results are presented in Table 3 in conjunction with Carlberg and Kavale's overall findings.

An initial comparison of the data presented in Table 3 disclosed findings that appear to be contradictory to the results reported by Carlberg and Kavale. Although the M.E.S.s calculated from each sample are of similar magnitude, they provide evidence for opposing positions. While Carlberg and Kavale noted that their results "demonstrated that special
class placement is an inferior alternative to regular class placement" (p. 304), the replication findings refute such a statement. The results also seem to be inconsistent with the non-significant differences obtained between the two samples' outcomes.

Table 3. Description of the Overall Results in the Replication and Carlberg and Kavale's Study.

<table>
<thead>
<tr>
<th>Result</th>
<th>Replication</th>
<th>Carlberg &amp; Kavale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td>Mean Age</td>
<td>11.35</td>
<td>11.0</td>
</tr>
<tr>
<td>Mean I.Q.</td>
<td>78.3</td>
<td>74.0</td>
</tr>
<tr>
<td>Mean Length of Intervention in weeks</td>
<td>51</td>
<td>69</td>
</tr>
<tr>
<td>Total Number of E.S.s</td>
<td>198</td>
<td>322</td>
</tr>
<tr>
<td>M.E.S.</td>
<td>.105</td>
<td>-.12</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>.612</td>
<td>.65</td>
</tr>
<tr>
<td>Range</td>
<td>-1.571</td>
<td>-1.31</td>
</tr>
<tr>
<td></td>
<td>1.71</td>
<td>1.98</td>
</tr>
<tr>
<td>95% Confidence Interval</td>
<td>.020</td>
<td>-.18</td>
</tr>
<tr>
<td></td>
<td>.190</td>
<td>-.06</td>
</tr>
</tbody>
</table>
An unweighted Stouffer's $Z_{max}$ of 2.219 was obtained from the integration of the probabilities associated with the outcomes presented in each study. A score of 2.219 equates to a one-tailed probability of less than .0148, which implies that the likelihood of the sampled studies' results being generated by chance is less than 1.5 in a hundred. The calculation of the fail-safe $N$ determined that at least 127 studies that either confirm the null hypothesis or report evidence of regular classrooms being superior to special education would be needed to raise the observed probability score to more than 5 chances in a 100.

Data by Outcome Measures

Analyzing the E.S.s on the basis of the domain of the outcome measure provided the results shown in Table 4. Special class placement was inferior to regular class placement in the areas of academic achievement and behaviour. Students who were educated in the special classes were considered to be slightly more negative in their behaviour (2.8%), and functioning at an academic level 2.5% below the students in regular classes. In opposition to this, the children in special classes demonstrated gains on the cognitive (14%) and social/personality (4%) measures. The differences among the four domains of outcome measure were found to be significant on the basis of an ANOVA ($F[3,194] = 5.978, p < .01$). Allowing for the fact that an ANOVA is not as
Table 4. Average Effect Size by Domain of Outcome Measure

<table>
<thead>
<tr>
<th>Domain of outcome measure</th>
<th>Number of effect sizes</th>
<th>Mean effect size</th>
<th>Standard deviation</th>
<th>Percentile status of median special class student compared to regular class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive</td>
<td>35</td>
<td>.418</td>
<td>.623</td>
<td>64</td>
</tr>
<tr>
<td>Academic achievement</td>
<td>66</td>
<td>-.074</td>
<td>.563</td>
<td>47</td>
</tr>
<tr>
<td>Social/personality</td>
<td>83</td>
<td>.128</td>
<td>.606</td>
<td>54</td>
</tr>
<tr>
<td>Behaviour</td>
<td>14</td>
<td>-.084</td>
<td>.251</td>
<td>47</td>
</tr>
</tbody>
</table>

rigorous when there are heterogeneous variances and unequal numbers in each sample group, the differences would be still significant at the .05 probability level.

Multiple comparisons of the M.E.S.s using individual t tests significantly differentiated the outcomes of the domain groups. The M.E.S. of the cognitive outcome category was significantly different to academic achievement ($t = -3.993$, $p < .001$), social/personality ($t = -2.333$, $p < .05$), and behavioural categories ($t = -2.862$, $p < .05$). Even though the median special class student's performance ranged from the 47th to the 54th percentile among the latter three groups, the
differences were not significant at the .05 probability level.

The data, therefore, suggests that special class placement enhanced the average child's cognitive functioning by 14%. In addition, the special class placement did not substantially alter the student's performance on the academic achievement, social/personality, and behavioural scales. The students would have obtained approximately the same results had they been in the regular class. This last statement is in agreement with the remarks in Carlberg and Kavale's article.

The outcome results were also analyzed in terms of measure type. The measures which were categorized as being either standardized, ad hoc (experimenter designed), or observational had M.E.S.s of .143, −1.28, and .96 respectively. An ANOVA on these groups did not indicate any significant differences ($F(2,195) = 1.276$).

Data by Category of Exceptionality

An overwhelming majority (80%) of the studies in the sample focused on the effects of special education in EMR students. Of the remaining studies, three considered SL, two had ED subjects, and only one diagnosed its sample as having a combined SL/BD exceptionality. Table 5 provides a summary of the results for students in special versus regular classes. The calculated $F$ of $4.685[3,194]$ was statistically significant at the .01 probability level, which indicated that the M.E.S.s were not equal. In fact the
Table 5. Average Effect Size by Category of Exceptionality

<table>
<thead>
<tr>
<th>Category of exceptionality</th>
<th>Number of effect sizes</th>
<th>Mean effect size</th>
<th>Standard deviation</th>
<th>Percentile status of median special class student compared to regular class</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMR</td>
<td>145</td>
<td>.054</td>
<td>.599</td>
<td>52</td>
</tr>
<tr>
<td>SL</td>
<td>37</td>
<td>.414</td>
<td>.497</td>
<td>64</td>
</tr>
<tr>
<td>ED</td>
<td>10</td>
<td>-.146</td>
<td>.384</td>
<td>45</td>
</tr>
<tr>
<td>SL/BD</td>
<td>6</td>
<td>-.173</td>
<td>1.041</td>
<td>44</td>
</tr>
</tbody>
</table>

differences obtained from a series of individual t tests formed a chain-like pattern. The EMR group differed significantly from the SL group (t = 3.53, p < .01), which in turn differed from the both the ED group (t = -3.238, p < .01) and the combined SL/BD group (t = -2.161, p < .05). The ED and SL/BD groups failed to demonstrate any statistical differences between each other (t = -.069). Children diagnosed as being slow learners, which included poor readers and the perceptually handicapped, responded so well to special class placement that their performance across all outcome measures averaged 14% better than their peers in the regular classes. Whereas a marginal, 1.8% improvement was shown by the EMR students in special classes, the ED and SL/BD students were worse off in the special classes. The performance these students corresponded, approximately, to the 45 percentile of
their counterparts in the regular classrooms.

The results of Carlberg and Kavale's analysis of the category of exceptionality are listed in Table 6. Unfortunately, Carlberg and Kavale presented insufficient information to allow a precise examination of the discrepancies between the two studies shown in Table 6. An approximate calculation of a Pearson product moment correlation indicated that there was a relationship between the findings in the two studies. Despite the size of the relationship, which was $r = -0.889$, it was not significant at the .05 level of probability. Carlberg and Kavale further analyzed the data for interactions between the category of exceptionality and domain of outcome measure. This type of

<table>
<thead>
<tr>
<th>Category of exceptionality</th>
<th>Percentile status of median special class student compared to regular class: Replication</th>
<th>Percentile status of median special class student compared to regular class: Carlberg &amp; Kavale</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMR</td>
<td>52</td>
<td>44</td>
</tr>
<tr>
<td>SL</td>
<td>64</td>
<td>37</td>
</tr>
<tr>
<td>ED</td>
<td>45</td>
<td>--</td>
</tr>
<tr>
<td>Combined</td>
<td>44</td>
<td>61</td>
</tr>
</tbody>
</table>
analysis was more than was required in the replication and was not pursued.

Data by Type of Classroom Setting

Another analysis was performed on the data using a framework of educational setting. The E.S.s were placed into groups according to the type of instruction the students received: 1) regular with tutoring; 2) special; and 3) special with tutoring. The students receiving one-on-one tutoring, regardless of their main classroom experience, showed an overall improvement of more than 7.5%. The M.E.S.s for the regular with tutoring and special with tutoring groups, which were .224 and .226 respectively, did not differ significantly from each other on the basis of a t test. Special education alone seemed to be of least benefit to the students since the M.E.S. was only .069. Despite the conspicuous disparity among the groups' M.E.S.s an ANOVA performed across all the groups found no significant differences (F[3,194] = .8018).

Correlations of Study Features with Effect Size

In an attempt to discern which study features were related to the obtained E.S.s, a series of Pearson product moment correlations were performed on the data. The r values obtained from the correlations all tended to be low, and in
only one relationship could reliably be considered as nonzero.

Age of the subjects was correlated significantly with the
E.S.s at the 5% probability level ($r = .190$). Table 7 lists

<table>
<thead>
<tr>
<th>Study feature</th>
<th>Correlation obtained in replication</th>
<th>Correlation obtained by Carlberg &amp; Kavale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean I.Q.</td>
<td>.13</td>
<td>-.07</td>
</tr>
<tr>
<td>Mean age</td>
<td>.19*</td>
<td>.01</td>
</tr>
<tr>
<td>Socio-economic status</td>
<td>-.06</td>
<td>--</td>
</tr>
<tr>
<td>Length of time between treatment and measurement</td>
<td>-.04</td>
<td>.02</td>
</tr>
<tr>
<td>Attrition from special class</td>
<td>-.01</td>
<td>.04</td>
</tr>
<tr>
<td>Attrition from regular class</td>
<td>.00</td>
<td>-.01</td>
</tr>
<tr>
<td>Size of sample in special class</td>
<td>.04</td>
<td>.06</td>
</tr>
<tr>
<td>Size of sample in regular class</td>
<td>.00</td>
<td>-.01</td>
</tr>
<tr>
<td>Date of publication</td>
<td>-.11</td>
<td>.02</td>
</tr>
<tr>
<td>Number of threats to internal validity</td>
<td>.12</td>
<td>.14</td>
</tr>
</tbody>
</table>

* indicates significance at the $p < .05$ level
the correlations that were tested with their accompanying $r$
values obtained in this study as well as Carlberg and
Kavale’s.

Summary

The results of the quantitative review indicated that
the children in a special classroom were functioning, on the
average, one-tenth of a standard deviation above similar
children in a regular setting. The greatest improvement was in
the children’s cognitive ability, with the median special
group children functioning at a level that would place them in
the 64th percentile of the regular group. An examination of
the data by category of exceptionality revealed that children
diagnosed as SL made the most improvement in a special
classroom, 14%, while children labeled as either ED, or having
combined exceptionalities, demonstrated a decrease in their
functioning (5%). Individual tutoring in addition to the
regular and special classroom experience did not significantly
alter the children’s overall performance.

The overall results of the replication were a complete
reversal of Carlberg and Kavale’s findings. The original
review obtained a M.E.S. that indicated approximately a one-
tenth of a standard-deviation inferiority in performance of
children in a special education setting. A comparison of the
M.E.S.s calculated for the individual sample studies also
revealed many differences that lead to an inter-rater
agreement of only 41%. 
DISCUSSION

Interpretation of Results

The results from the study provide insight into two issues in education. One interpretation focuses on whether or not mainstreaming benefits exceptional children, whilst the other demonstrates the replicability of a quantitative review. When the results are considered from both points of view, the viability of quantitative reviews as a means of evaluating educational programmes can be determined.

Efficacy of Special Versus Regular Class Placement

Considering the amount of controversy that has surrounded the issue of mainstreaming, the review results were quite surprising. There was no evidence of the deleterious consequences nor the copious benefits that had been predicted by those arguing about the relative merits of special education. The integration of the data from the sample of 30 studies revealed a 3.5% improvement in the functioning of exceptional children who had been placed in special classroom when compared to similar children who had remained in regular classes. Since the limits of the 95% confidence interval associated with the outcome did not span zero, the improvement, although small, could be considered significant. The calculation of a confidence interval, (which was used by
Carlberg and Kavale in their analysis) however, assumes that the data being used is independent; that is to say, the occurrence of outcome E.S.s were in no way related to each other. In the review this assumption was violated. Consequently, the limits of the confidence interval are approximate and may actually span zero. More accurate information on the significance of the overall findings was provided by the Stouffer's Z that had been incorporated into the review. The obtained Z of 2.219 indicates that the overall results could be expected to occur less than 1.5 times in a hundred on the basis of chance alone.

Of the main categories of outcome measures considered in this review, only cognitive functioning appeared to change significantly. Academic achievement, social/personality, and behaviour varied slightly, but not to the point where they could be considered as either a statistical improvement or worsening of student performance. Therefore, following the same reasoning as employed by Carlberg and Kavale, it can be concluded that an exceptional child's self-concept, social interactions, academic abilities, and classroom behaviour are unaffected by the type of educational experience. The M.E.S. of the cognitive measures established that the special class students were performing at a level 14% higher than the students in the regular classes.

Students who had been diagnosed or classified as SL also demonstrated a 14% improvement in their functioning after
experiencing special class placement. There were no
differential placement effects discernable among the EMR, ED,
and combined SL/BD categories of exceptionality. In these
three groups the status of the median special class child
compared, respectively, to the 52nd, 45th, and 44th percentile
in the regular class. Such results suggest that only the
educational difficulties experienced by SL children can be
ameliorated, to some extent, by a special class placement.
Consequently, the overall functioning of the majority of
exceptional students is neither helped nor hindered by a
special class placement.

The minimal effects of special education were evident
in all three categories of intervention. Children in either
regular or special classes who received additional tutoring
showed gains of approximately 7.5% in their overall
functioning. Education in a special classroom alone resulted
in only a 2% improvement.

For the most part study characteristics were not
related to the outcome E.S.s. The students' mean I.Q., socio-
economic status, delay in performing outcome assessment,
attrition, individual study sample sizes, date of publication,
and threats to internal validity, failed to show any effects
on the treatment outcomes. Mean age was the only significant
relationship documented in this study. The obtained $r$ of .19
indicates that the older the child, the greater the
improvement in their general functioning.

The calculation of the fail-safe $N$ indicated that at
least 127 'disconfirming' studies would have had to have been incorporated into the review in order to change the findings. Although this figure seems quite large, it is not sufficient to conclude that the results of the review are resistant to "unretrieved null results" (Cooper, 1984, p. 94). Rosenthal (1978) proposed that the $N_{\text{min}}$ must be equal to, or exceed, 5 times the number of studies in the sample plus 10, before the review results can be considered resistant to the effects of unpublished works or studies not included in the sample. Accordingly, an $N_{\text{min}}$ of at least 160 would have been required to maintain the review's resistance to unretrieved null results. Cooper (1984) speculates that Rosenthal's rule isn't steadfast. Instead, he suggests that the resistance hinges on the comprehensiveness of the search strategy used to locate studies for the review in conjunction with the size of the $N_{\text{min}}$. Since the replication's sample was limited to available literature, and as such only included 30 studies compared to the 860 documents originally identified by Carlberg and Kavale, it cannot be considered complete, nor, therefore, resistant.

The review was also limited in terms of information. Unfortunately, the studies included in the review provided incomplete information on research design and procedures, which precluded a detailed investigation of special education. The results that have been described thus far, could have been influenced, for example, by the subjects' previous special education experience, the training of the instructors in
charge of the special classrooms, student/teacher ratio in the classrooms, and the means of obtaining teacher participation in the research studies. Without the necessary information, the relationships among these variables and the outcome E.S.s cannot be determined.

The findings from this quantitative review of the literature on the efficacy of special education suggests that the present policy of returning exceptional children to a regular classroom may not be in the children's best interests. The detrimental effects envisaged by the opponents of special education were not observed in this review. Instead, special education proved to be beneficial to the overall functioning of the students. The improvements shown after special intervention, although small, are present and many administrators may be 'throwing the baby out with the bathwater' in their haste to reform the educational system. Echoing Carlberg and Kavale the "present policies require re-examination" (p. 305) before major decisions are made.

Replicability of Quantitative Reviews

Since the overall findings of this study are approximately equal, but in the opposite direction of Carlberg and Kavale's study, the replicability of the quantitative approach becomes questionable and must be investigated. In order to accomplish this, the exact nature of the differences between the two reviews must first be
established.

When the 26 studies in common to both samples are compared statistically, there are no significant differences between the two groups ($t = 1.0186, df = 84$). The differences only become apparent after a comparison of the 29 pairs of corresponding outcome measures. Such a comparison reveals that a mere 41% of the 29 E.S. pairs are the same value when matched to the second decimal place. Most of the mismatches are major discrepancies and they do not follow any discernible pattern. If, for example, all the E.S.s derived from a given type of outcome data had failed to match or else differed by a constant amount, then the differences could be attributed to a specific methodological flaw in calculating E.S.s. (In one instance the magnitude of the E.S.s were identical, but one was positive, while the other was negative.) The observed differences could be a consequence of several reviewer judgments, and perhaps reviewer error, that can occur during the three main stages of a quantitative review; the stages being 1) inclusion of studies in the sample, 2) calculation of E.S.s from the reported outcome data, 3) interpretation of the results.

A combination of evidence suggests that the differences between the two groups were not due to the study selection processes. The non-significance found between the E.S.s in the two groups implies that the results are representative of one population, not two, as would be inferred from significant results. Additionally, both reviews
operated under the same criteria in the selection of studies for inclusion in their samples. An examination of the reliability of the quantitative review techniques (Stock et al., 1982) found that given clear standards on which to base decisions, coders can achieve an 88% agreement on the inclusion of studies in a sample.

Stock et al.'s study also documented inter-rater agreement that averaged more than 90% in the calculation of data points to be used in a quantitative analysis. The coders in Stock et al.'s study followed an 80 page codebook that 'was written to facilitate the extraction and calculation of data points' (p. 11). Although an 80 page codebook required a lot of work to prepare, the authors felt that the complexity of the analysis demanded an extensive and comprehensive set of coding guidelines. The study by Carlberg and Kavale outlined the review procedures, including the coding, in less than three pages. The brevity of their coding instructions in combination with the diversity of the statistics reported in the sample studies provided ample opportunity for different E.S.s to have been calculated in this replication.

Two primary differences in the coding procedures appear to have been the placement of students into categories of exceptionality and the classification of the outcome measures. In the replication, the category of exceptionality was determined on the basis of the diagnosis or classification used in the sample studies. From the information provided in the original review, it is not possible to be certain if a
similar method was used or if the studies were coded by Carlberg and Kavale on the basis of the subjects' I.Q. Carlberg and Kavale used broader categories for classifying the outcome measures than the replication. In the replication cognitive skills, academic achievement, personality/social attributes, and observed behaviour were separated into distinct categories. Carlberg and Kavale only reported outcome measures as being either social/personality, achievement, and other. Without the specifications of exactly what was subsumed under each title, several disparate coding decisions were made in the replication. For example, a study that examined the differential effects of special class placement on the suggestibility of EMR students (Trippi, 1973) was considered social/personality in the replication, yet was given a dual classification of 1) achievement and 2) other by Carlberg and Kavale in their review. Such differences precluded, in many instances, direct comparisons of the E.S.s. between the studies that were in both the original and replication samples.

Another area in which different coding conventions appeared to have been employed was in the estimation of E.S.s from non-significant results. The study by Gotlieb and Davis (1973) noted that the subjects "did not differ significantly in the number of times they chose an integrated (8) or a segregated (6) EMR child" (p. 142). From this statement alone, Carlberg and Kavale estimated the E.S. of the study to be -.27, whereas it was assigned a value of zero in the
replication. There was, however, no consistency in Carlberg and Kavale's method of arriving at E.S.s. In another study (Gottlieb and Budoff, 1973) where actual means and standard deviations were reported in conjunction with the statement "no significant differences were found between partially integrated and segregated EMR children in the percentage of times they were selected as friends" (p. 17), Carlberg and Kavale obtained an E.S. of zero. The replication used the exact figures whenever they were provided to calculate the E.S.s, consequently Gottlieb and Budoff's study had an overall E.S. of .131.

Some differences between the two sample groups could also have resulted from the manner in which negative outcome scales were coded. The replication followed the convention of reversing the sign associated with any negative outcome measure that has been described previously in the Method chapter (p. 63). Since this problem was not discussed by Carlberg and Kavale there could be no assumptions made as to how such outcomes were coded in the original review.

Yet another area that was subject to coding differences was the actual conversion of the raw data presented in the sample studies into E.S.s. The data usually conformed to one of three possible formats. Either it was 1) a direct comparison between special and regular groups after a specified period of intervention, 2) a repeated measures design performed at pre, post, and sometimes follow-up on the special group alone, or 3) a combined effort that
used repeated measures with both regular and special classes. The first two types of data presented no real problems. Effect sizes were obtained using the basic formula of:

$$E.S. = \frac{\bar{x}_{exp} - \bar{x}_{con}}{SD_{con}}$$

If the data was from a repeated design study, the pretest outcome scores served as the control group in the calculation of the E.S.s. Data based on the combined design, however, could have been used in two, radically different, methods of determining the E.S.s.

Adherence to the above formula would yield an outcome E.S. as illustrated in the following example.

<table>
<thead>
<tr>
<th></th>
<th>Special class</th>
<th>Regular class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretest mean scores</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Pretest standard deviation</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Posttest mean scores</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>Posttest standard deviation</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

$$E.S. = \frac{9 - 10}{3} = \frac{-1}{3} = -0.333$$

The negative E.S. would then be interpreted as evidence supporting the regular class placement of exceptional children. Although such a calculation appears to be a legitimate application of the basic formula, it is in practice
quite insensitive to the pretest group differences and, thus it paints an inaccurate picture of the study's results. When the pretest differences are considered, it is clear that the children in the special class achieved overall gain of 6, while the control group only increases their score by 5.

Glass, McGaw and Smith (1981) proposed an alternative method of determining more representative E.S.s in such situations. Using the suggested modifications, the above formula becomes:

\[
E.S. = \frac{(\bar{X}_{\text{exp}} - \bar{X}_{\text{con}}) - (\bar{Y}_{\text{exp}} - \bar{Y}_{\text{con}})}{SD_g}
\]

The terms \(X_{\text{exp}}\) and \(X_{\text{con}}\) still refer to the mean posttest outcome scores, \(Y_{\text{exp}}\) and \(Y_{\text{con}}\) are introduced as the mean pretest outcome scores, and \(SD_g\) is the variance of the gain scores. An estimation of the variance of the gain scores (\(SD_g\)), using the known or 'guessed' correlation between the control group's pretest-posttest scores \(r_{xy}\), is obtained from:

\[
SD_g = SD_x[2(1 - r_{xy})]^{1/2}
\]

Substituting the data from the preceding example into the alternative formula yields an E.S. as follows:
For the purpose of the example the value .95 was given as the correlation between pretest and posttest scores. It was noted by Glass, McGaw, and Smith that in situation where the correlation is not specified, nor can it be calculated, a "reasonable guess can probably be made if something is known about the tests involved" (p. 118).

The E.S.s obtained from these two formula are radically different. Whereas the initial E.S. seemingly proved the regular group's superiority, the modified E.S. revealed that the special group's performance was actually superior in terms of overall benefit. The modified version was utilized with the appropriate studies throughout the replication. It is impossible to determine from the information supplied by Carlberg and Kavale whether or not the modified formula was incorporated into their review. If the simpler, posttest only, form was used, it would account for a good proportion of the deviations in reported E.S.s.

Lastly, some of the differences could have been nothing more than oversights in the interpretation of the sample studies' results or omissions in the proofing and printing of Carlberg and Kavale's manuscript. The study by Trippi (1973) serves as one example. Trippi's study investigated suggestibility, or more specifically, how EMR children's judgments of visual stimuli were influenced by
knowledge of the judgments made by others. His conclusion that "EMR children who have been placed in special classes were less suggestible than retarded children who remained in the regular grades" (p. 221), was interpreted in Carlberg and Kavale's review as being a positive E.S. on two outcome scales; achievement and other. Carlberg and Kavale's results are inconsistent with the fact that only one outcome measure was reported in Trippi's paper. It could be indicative of a genuine misinterpretation of the information presented by Trippi or just a mere typographical error in the printing of the manuscript.

Carlberg and Kavale's interpretation of the overall results may have also been inaccurate. The conclusion that "special class placement is an inferior alternative to regular class placement" (p. 304), was based on the limits of the 95% confidence interval bounding the overall M.E.S. Since the interval extended from -.10 to -.06, it was considered to be significantly different from zero in the negative region, and thus favoured regular class placement. There would be no argument about their conclusion if all 322 E.S.s had been independent. As most studies reported multiple outcome E.S.s, the statistically assumption of independence was violated, and at best the range of the confidence interval was only an approximation. The conclusions drawn from such an approximation are only tentative. Carlberg and Kavale noted this problem, yet still used the results of the 95% confidence interval as the sole platform for their conclusions. Without
an other, more appropriate analysis, it is impossible to tell whether or not Carlberg and Kavale's overall findings actually differ from the results obtained in the replication.

The 41% inter-rater agreement among the average study E.S.s obtained in the original review and the replication is not surprising in view of the numerous sources of potential disagreement. Most of the discrepancies reported between Carlberg and Kavale's review and this replication appear to be the result of insufficient information about the coding process. The calculation of E.S.s from the data presented in the sample studies is one stage of the coding process that affords the most opportunities for reviewer disagreement. The E.S.s obtained in any given review can vary extensively depending on which method was chosen to extract the relevant information from the sample studies. The problem is exacerbated by the fact that most reviewers fail to discuss the procedures and assumptions they have used in their analyses. The widespread practice of completely describing all coding conventions and specifying the exact methods used to calculate the E.S.s would greatly increase the replicability of the quantitative review process.

Role of Quantitative Reviews in Evaluation

The replication has demonstrated that two quantitative reviews, ostensibly based on the same literature but differing in the finer aspects of the analysis, can lead
to diametrically opposed conclusions. Concerns about possible inconsistencies amongst coders and reviewers, expressed by Glass, McGaw & Smith (1981) and more recently, Green & Hall (1984), seem to be well founded. The complexity and diversity of the procedures available in a quantitative approach provide reviewers with many opportunities to make methodological mistakes and erroneous interpretations of the available data. Each calculation of an E.S. is an individual exercise, and in some circumstances, only obtained from an extensive reanalysis of the information provided in a sample study. In such situations "errors, or at least differences in judgment, are inevitable" (Green & Hall, 1984, p. 49). Until reviewers adopt a set of standard procedures, programme administrators must be made aware of the subjectivity that is possible in a quantitative review and the subsequent need to examine the objectivity and validity of all reviews prior to incorporating their findings into planning policies.

Despite this encumberance, quantitative reviews still have several advantages over the traditional narrative method. Effects that are typically hidden in individual studies with small sample sizes or nonsignificant results are more readily detected in a quantitative review of a large body of literature. This is of particular importance to educational evaluators. New curricula and programmes, which are usually evaluated on the basis of a test-run using a relatively small number of children, tend to yield significant results only if
there are large differences between the 'control' and 'treatment' groups. Modest gains that are obscured by nonsignificant findings can be identified from the aggregation of information in a quantitative review.

The results of a quantitative review can also present a fairly comprehensive picture of a given issue. In fact, variables that were not directly investigated in any of the individual studies may be incorporated into a quantitative review. This point can be illustrated by the mainstreaming studies used in the replication. The relationship between class size and student performance was never addressed in any of the sample studies; yet through the quantitative analysis it was possible to determine that the correlation coefficient of these two variables was .04 and .00 for special and regular classes respectively.

A close examination of a quantitative review's procedures, specifically the variables mentioned in the results, can also assist educational administrators in identifying issues that deserve further investigation. Both the original study by Carlberg and Kavale and the replication have demonstrated the quantitative review's ability to identify the deficiencies that exist within a body of research. For example, the literature reviewed on mainstreaming typically ignored the effects of teacher training, students' socioeconomic status, and teacher/student ratio. Knowledge of such omissions can be used by administrators in developing additional research projects.
There is definitely a place for quantitative reviews in educational evaluations and research. The approach, and its affiliated techniques, should not be dismissed because of potential reviewer subjectivity. Instead, the method should be used in any appropriate situation, but used with caution.

Conclusions

Although most literature reviews still follow the intuitive, subjective, narrative style that has traditionally been used, the quantitative approach is becoming so popular that it has been called "a revolution in the making" (Rosenthal, 1984, p. 16). The systematic techniques involved in the quantitative approach allow for a more explicit and exhaustive review of a body of literature than has previously been possible. The variety of techniques that are available within the quantitative framework can be used individually to provide limited information on a given topic, or combined in numerous ways to furnish a more thorough understanding of the issue at hand. Such reviews can serve many purposes. They can summarize past research and draw overall conclusions about a specific hypothesis, test which theory is most consistent with the known information, and examine the research methods that have been employed in any given area.

The use of a quantitative approach, however, does not preclude erroneous conclusions or guarantee replicability. Reviewers can use the techniques inconsistently and
inappropriately, either of which will provide misleading results. Educators must be alerted to the hazards of unconditionally accepting the findings from a quantitative review. The approach is not infallible. A flawed quantitative review can, because of its assumed objectivity and superiority, be more injurious to the research endeavour than an overtly biased narrative review. A comprehensive, methodologically accurate quantitative review can be of benefit in virtually any aspect of education, and especially in the planning and evaluating process. Consequently, researchers and administrators must evaluate the conclusions drawn from each quantitative review in the same way as they would examine the results derived from individual investigations. Furthermore, if the reviews are to be of practical benefit, the exact procedures used to aggregate the literature will have to be described in sufficient detail to allow the validity of the results to be determined.

The package of techniques used in the replication documented the type of information that can be obtained as well as the variability that can occur at each stage in quantitative reviews. Due to the fact that it was a replication, the statistical analysis followed the pattern outlined in Carlberg and Kayale's original study. A more extensive analysis, although possible, was not required for comparative purposes and as such was not included in the replication.
Implications for Future Research.

The results of the replication, specifically the 41% inter-rater agreement with Carlberg and Kavale's study, provided some insight into how the integration of a body of literature can be inadvertently determined by the reviewer. Both reviews employed Glass' methodological framework, yet a direct comparison between the E.S.s calculated for individual studies revealed considerable differences in the actual practical applications of the technique. Given that such differences can arise from Glass' method, additional research is needed to document the variability within the other effect size procedures.

Rosenthal (1984) describes three categories of techniques that can all be used to ascertain the E.S.s from sample studies in a quantitative review. Glass' method is only one of the three techniques that are based on the standardized differences between means. The other two categories encompass 7 distinct methods of deriving E.S.s from product moment correlations (r) and functions of r as well as the differences between proportions. These additional methods require different types of statistical information, thereby allowing E.S.s to be retrieved from many more studies. The combined use of these methods could increase the scope of reviews and assist researchers and administrators in gaining a more complete understanding of a given issue, but only if the researchers and administrators are aware of any differential
results or disadvantages associated with each technique. Each procedure, therefore, deserves further investigation.

Another area that should be considered in the future is the independence of the E.S.s. Due to the presence of multiple outcome measures, the 30 studies in the replication yielded 198 E.S.s, while the original review reported 322 E.S.s from a sample of 50 studies. Carlberg and Kavale (and likewise the replication) analyzed the review sample as though each E.S.s was a separate entity that had no bearing on any of the others and subsequently noted that this may have influenced statistical calculations that assumed independence. The results of this type of analysis should be compared to a weighted analysis, one in which all outcome E.S.s reported in a study are combined into a single measure. Such measures could then serve as the units in an analysis without violating the independence assumptions.

Even though the debate about special versus regular class placement for exceptional children was a side concern in this study, it is one area of education that definitely requires further research. The quantitative analysis of the sample studies identified several issues that researchers have generally ignored, yet need to be examined if the efficacy of special education is to be more fully understood. Some issues that warrant additional research are 1) the impact of the special education and regular teachers' training level and classroom experience on the exceptional students' abilities,
2) how the size of the student/teacher ratio effects the exceptional students’ performance in segregated and integrated settings, 3) and the influence of parental motivation and support in the children’s education, which was one area that was never mentioned in the sample studies.

Most of the studies included in the replication considered the short-term consequences of educational placement, that is to say, changes in student abilities over the span of an academic year. Apart from a couple of studies that provided results from a subsequent assessment in the succeeding year, there were no efforts directed towards determining the long-term effects of special education. Ethical and practical constraints may have prevented such research in the past, (as noted earlier this study did not delve into special education), but that does not prevent it from being an issue that must be included in deciding which form of education will maximize an exceptional child’s potential, and as such, worthy of further investigation.

Additional information on mainstreaming could be obtained from a more detailed quantitative review. Rather than merely extending the analysis of the data presented in the replication, any future endeavour should re-examine the entire body of literature. The package of techniques should be expanded to include all appropriate forms of calculating E.S.s, which would then permit a the inclusion of a greater number of studies in the sample, so that a more complete picture of the issue would be possible. The scope of the
statistical analysis could also be increased, through techniques such as an ANCOVA, to provide more information.

Summary

The package of review techniques demonstrated the type of information that can be obtained from a literature review. The results of reviewed studies indicated that exceptional children benefit from special education intervention over a regular class placement. The comparison between Carlberg and Kavale's study and the replication documented the variability in results that can occur when reviewers have different interpretations of the procedures. The information provided from a quantitative review can assist educators in evaluating programmes that have been the topic of many research efforts, if, and only if, the specific merits and deleterious consequence of the techniques are known. The quantitative approach to literature review is still evolving and in view of its potential benefits and increasing popularity, it should be the topic of considerable research in the future.
APPENDIX A

CODEBOOK

Note: the data was entered in 'card' format, with one card being completed for each outcome effect size, and at least one card being completed for each group.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Possible Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Study identification number, ranged from 1 - 30</td>
</tr>
<tr>
<td>Group</td>
<td>Group type</td>
</tr>
<tr>
<td></td>
<td>1 regular classroom</td>
</tr>
<tr>
<td></td>
<td>2 regular classroom with additional tutoring</td>
</tr>
<tr>
<td></td>
<td>3 special classroom</td>
</tr>
<tr>
<td></td>
<td>4 special classroom with additional tutoring</td>
</tr>
<tr>
<td>Maj.</td>
<td>Major category of exceptionality</td>
</tr>
<tr>
<td></td>
<td>1 EMR</td>
</tr>
<tr>
<td></td>
<td>2 SL, includes poor readers, percept. handicapped</td>
</tr>
<tr>
<td></td>
<td>3 BD</td>
</tr>
<tr>
<td></td>
<td>4 ED</td>
</tr>
<tr>
<td></td>
<td>9 NK (not known, not specified)</td>
</tr>
<tr>
<td>Com.</td>
<td>Categorical combination</td>
</tr>
<tr>
<td></td>
<td>1 EMR/BD</td>
</tr>
<tr>
<td></td>
<td>2 EMR/ED</td>
</tr>
<tr>
<td></td>
<td>3 BD/ED</td>
</tr>
<tr>
<td></td>
<td>4 SL/BD</td>
</tr>
<tr>
<td></td>
<td>5 SL/ED</td>
</tr>
<tr>
<td></td>
<td>9 NK</td>
</tr>
<tr>
<td>IQ</td>
<td>Mean IQ for group i.e. 100.4, 999 = NK</td>
</tr>
<tr>
<td>Age</td>
<td>Mean age for group i.e. 72.4 mos, 999 = NK</td>
</tr>
<tr>
<td>Male</td>
<td>Percent male in total sample i.e. 45.3%, 999 = NK</td>
</tr>
<tr>
<td>Soc</td>
<td>Socio-economic status group</td>
</tr>
<tr>
<td></td>
<td>1 low</td>
</tr>
<tr>
<td></td>
<td>2 middle</td>
</tr>
<tr>
<td></td>
<td>3 high</td>
</tr>
<tr>
<td></td>
<td>9 NK</td>
</tr>
</tbody>
</table>
Trt  Type of treatment
1 physical/perceptual practice
2 special reading instruction
3 social/psychodrama
4 resource room
5 preschool
6 academic remediation
7 behavioural modification
9 NK, or not applicable

Dur  Duration of treatment, # of weeks in special class
999 = NK

Ins  Instructor's training level
1 classroom aide
2 undergraduate
3 graduate
4 regular classroom teacher
5 special ed training/experience
9 NK

InsE Instructor's experience, # years teaching experience
999 = NK

Num  Number of students in group i.e. 35, 09
999 = NK

SIR  Student/Instruction ratio i.e. 25 would indicate 25
     students to 1 instructor.
     999 = NK

Pub  Publication form
1 journal article
2 book
3 thesis

Date Publication date i.e 34, would indicate 1934,
     and 84 would correspond to 1984

SolS Solicitation of subjects
1 selected by experimenter
2 referred/recommended to a special class
3 already placed in a special class
9 NK

SolI Solicitation of Instructors
1 selected by experimenter
2 volunteered
3 assigned by school administrator
9 NK
AssS  Assignment of students
1 random
2 matched pairs
3 convenient sample
9 NK

AssI  Assignment of instructors
1 random
2 matched pairs
3 convenient sample
9 NK

Att  Percent attrition in group i.e. 12.4%  
999 = NK

Comp  Type of comparison group
1 control
2 second treatment

Valid  Number of threats to the internal validity
From the following list, total the number of 
extraneous variables that are applicable:
0 none present
1 history
2 maturation
3 testing
4 statistical regression
5 instrumentation
6 differential selection
7 experimental mortality
8 selection-maturation interaction
9 John Henry effect

Dom  Domain of outcome score
1 cognitive
2 academic achievement
3 social/personality
4 behavioural
5 other
9 NK

Out  Type of outcome measure
1 standardized
2 ad hoc, experimenter/instructor designed
3 observational

Time  Time of outcome measurement
1 pretest
2 posttest
3 follow up

Dely  Delay in weeks from the conclusion of instruction to 
measurement, 999 = NK
<table>
<thead>
<tr>
<th>Reac</th>
<th>Obrusiveness of outcome measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>low</td>
</tr>
<tr>
<td>2</td>
<td>medium</td>
</tr>
<tr>
<td>3</td>
<td>high</td>
</tr>
<tr>
<td>9</td>
<td>NK</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Prob</th>
<th>Derived Z score, 999 = NK</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>E.S.</th>
<th>Effect size of outcome, 999 = NK or not applicable in the case of the regular classroom group</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Blnd</th>
<th>Blindedness of outcome measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>blind to the subject's status within the experiment</td>
</tr>
<tr>
<td>2</td>
<td>non-blind, had knowledge of subjects group assignment</td>
</tr>
<tr>
<td>9</td>
<td>NK</td>
</tr>
</tbody>
</table>
APPENDIX B

Programme 1

10 CLS: PRINT "Choose the required option for calculation of E.S.*
20 PRINT: PRINT 1) Mean of Experimental and Control Groups with SD.*
30 PRINT 2) t-statistic and sample sizes.
40 PRINT 3) F-statistic and sample sizes.
50 PRINT: PRINT "Type of test"
55 A$=INKEY$: IF A$="" GOTO 55
60 CLS: I=VAL(A$): ON I GOTO 100, 150, 200
70 END
100 INPUT "Mean of experimental group " ; XEXP
110 INPUT "Mean of control group " ; XCON
120 INPUT "Standard Deviation of Control" ; SCON
130 ES=(XEXP-XCON)/SCON: GOTO 500
150 INPUT "t-value " ; TCON
160 INPUT "n of control group " ; NCON
170 INPUT "n of experimental group " ; NEXP
180 ES=TCON*SQR(I/NCON+I/NEXP): GOTO 500
200 INPUT "F-value " ; F
210 INPUT "n of control group " ; NCON
220 INPUT "n of experimental group " ; NEXP
230 TCON=SQR(F)*ES=TCON*SQR(I/NCON+I/NEXP): GOTO 500
500 PRINT: PRINT "Effect size is " ; ES
510 A$=INKEY$: IF A$="" GOTO 510
520 GOTO 10
Programme 2

10 Z$=CHR$(13)+CHR$(10):ZZ$=CHR$(160):NF=25
20 DIM A(NF),E1(5,NF),E2(5,NF),N(5,NF),V5(NF),V1(NF),V2(NF)
30 DIM SD(5,NF),SG(5,NF),FLAG(5,NF),FLAGA(NF)
40 CLS:INPUT"Data file required = "CF$
50 GOSUB 1140:DEF FNA(X)=INT(X+.5)
60 CLS:PRINT"1) Calculate Means,t,f"Z$"2) Print"
70 PRINT "3) Pearson correlation"Z$"4) 1-way Anova"Z$"5) End"Z$
80 A$=INKEY$:IF A$=""GOTO 80
90 A=VAL(A$):ON A GOTO 100,290,670,900,2000:GOTO 100
100 CLS:INPUT "independent variable";MV
110 INPUT "of sample groups";SI:CLS
120 FOR I=1 TO SI:PRINT "Sample ";I:PRINT:INPUT "Range Low,High ";RL(I),RH(I)
130 INPUT "Description ";NP$;I:PRINT:PRINT:NEXT:GOSUB 1170:FLAG=0:SNBR=0
140 FOR I=1 TO SI:FOR J=1 TO NV:EI(I,J)=0;E2(I,J)=0;NI(I,J)=0:NEXT J,I
150 GOSUB 1180:IF A>999 THEN 60
160 IF A(I)=SNBR THEN 190
170 LOCATE 1,7:PRINT A(I)
180 FOR L=1 TO NV:FLAGA(L)=0:FOR M=1 TO 4:FLAG(M,L)=0:NEXT M,L
190 IF I=1 TO SI:FOR J=1 TO NV:IF V1(J)=0 OR A(J)>999 THEN 280
200 IF A(NV)<RL(I) OR A(MV)>RH(I) OR J=MV OR A(MV)>999 THEN 280
210 ON V2(J)+1 GOTO 260,220,250
220 IF FLAGA(J)=1 THEN 280
240 FLAGA(J)=1 GOTO 280
250 IF FLAGA(J)=1 THEN 280
270 FLAGA(J)=1 GOTO 280
280 NEXT J,I
290 CLS:PRINT "Print Function"
300 P1$="###

310 P2$="S$ N Sum x Sum x^2 Sum x/N Sum x^2/N SD 0"^2"
320 P3$="---

330 P4$="Test 1-2 1-3 2-3 1-4 2-4 3-4"
340 P5$="b & 

350 P6$="---

360 A=14:IF S1<2 THEN A=35:IF S1>3 THEN 380
370 P4$=LEFT$(P4$);A:PS$=LEFT$(PS$);A:PS$=LEFT$(PS$);A:PS$=LEFT$(PS$);A:
380 FOR I=1 TO SI:LEFT$"Sample ";I="NP$;I:NEXT:PRINT
390 FOR I=1 TO NV:IF I=MV THEN NEXT
400 IF V1(I)=0 THEN 460
410 LPRINT Z$"Variable ";I="V$;I:LPRINT Z$"2"Z$P3$ 
420 FOR J=1 TO SI:N=N(J,I);SD(J,I)=SDR(E2(J,I)-V1(J,I))/N-(E1(J,I)/N)^2
430 ON N+1 GOTO 460,450
440 SD(J,I)=SD(J,I)_2xN(N-1):IF SD(J,I)=0 THEN SD(J,I)=1
450 LPRINT USING P1$;J,N,E1(J,I),E2(J,I),E1(J,I)_N,E2(J,I)_N,SD(J,I),SG(J,I)
460 NEXT:IF V2(I)=1 THEN 660
470 LPRINT Z$"P$"P$"PS$;L=0:FOR J=2 TO SI:FOR K=1 TO J-1:IF K=J THEN 660
480 N1=N(J,I);N2=N(K,I);L=L+1
490 IF N1+N2=2=0 OR N1=0 OR N2=0 THEN T(L)=0:F(L)=0:GOTO 660
500 TI=E1(J,I)/NI-E1(K,I)/N2
510 T=TE2(J,I)_E1(J,I)_2/N1+2E2(K,I)_E1(K,I)_2/N2/(N1+N2-2)*(N1+N2)/(N1*N2)
520 IF T=0 THEN T(L)=0:GOTO 540
530 T(L)=T1/SBR(T)
540 G1=E1(J,1)+E1(K,1):G2=N(J,1)+N(K,1):G3=E1^2/G2
550 G4=E2(J,1)+E2(K,1):G5=G4-G3
560 Gb=E1(J,1)*2/N(J,1)+E1(K,1)*2/N(K,1)
570 MSA=66-63:G7=5-MSA
580 IF G2=2=0 OR G7=0 THEN F(L)=0:GOTO 600
590 MSE=67/(G2-2):F(L)=MSA/MSE
600 NEXT K,;ON SI GOTO 660,610,620,640
610 LPRINT USING P5$;"t",T(1),T(L),T(2),T(3)
620 LPRINT USING P5$;"f",F(1),F(2),F(3),F(4)
630 LPRINT USING P5$;"r",T(1),T(2),T(3),T(4),T(5),T(6)
640 LPRINT USING P5$;"F",F(1),F(2),F(3),F(4),F(5),F(6)
650 LPRINT:NEXT K1:LPRINT CHR$(12):GOTO 60
660 P0="r
670 P1="r
680 P2="r
690 P3="r
700 P4="r
710 CLS:INPUT "Variables for Correlation (2 values)";A1,A2:PRINT
720 INPUT "Do you wish to have a conditional variable (Y/N)?";A$ 
730 IF A$="n" OR A$="N" THEN RL=-1000:RH=1000:CV=0:GOTO 750
1030 \( G4 = E2(I1, I) + E2(I2, I) + E2(I3, I) + E2(I4, I) \): \( G5 = G4 - G3 \)

1040 \( G6 = E1(I3, I) \cdot \frac{2}{N(I3, I)} + E1(I4, I) \cdot \frac{2}{N(I4, I)} \): \( A = (I1 \times 1000 + I2 \times 100 + I3 \times 10 + I4) \)

1050 \( G6 = G6 + E1(I2, I) \cdot \frac{2}{N(I2, I)} + E1(I1, I) \cdot \frac{2}{N(I1, I)} \)

1060 \( G7 = G6 - G3; G8 = G5 - G7 \)

1070 \( G9 = SGN(I1) + SGN(I2) + SGN(I3) + SGN(I4); MSA = G7 / (G9 - 1) \)

1080 IF \( G2 - G9 = 0 \) THEN MSE = 0; \( F = 0 \); GOTO 1110

1090 IF \( G2 = 0 \) OR \( G8 = 0 \) THEN MSE = 0; \( F = 0 \); GOTO 1110

1100 MSE = \( G8 / (G2 - G9) \); \( F = MSA / MSE \)

1110 LPRINT USING C3; I, A, G7, G5, G8, G9, G2 - G9, MSA, MSE, F

1120 NEXT I, I4, I3, I2, I1

1130 NEXT I: LPRINT CHR$(12); GOTO 60

1140 OPEN "I", #1, CF$+, "VAR"; NV = 1

1150 INPUT #1, V$(NV), V1(NV), V2(NV); IF V$(NV) <> "END" THEN NV = NV + 1; GOTO 1150

1160 NV = NV - 1; CLOSE; RETURN

1170 OPEN "I", #1, CF$+", DTA"; RETURN

1180 INPUT #1, A; IF A > 998 THEN CLOSE; RETURN

1190 A(1) = A; FOR II = 2 TO NV; INPUT #1, A(II); NEXT II

1200 RETURN

2000 END
BIBLIOGRAPHY

General references


National Science Foundation for Grant #DIS 76-20398. Washington, DC: Social Research Group, George Washington University.


Studies in the quantitative review sample


Blatt, B. (1958). The physical, personality, and academic status of children who are attending special classes as compared with children who are mentally retarded attending regular classes. American Journal of Mental Deficiency, 62, 810-818.


VITA AUCTORIS

Since my parents are teachers, my twofold interest in education can probably be attributed to their influence. Due to their guidance, I recognized the importance of my own education and pursued a diverse course of studies when I entered the University of Toronto as an undergraduate. As a result I completed both a Bachelor of Science and a Bachelor of Arts as well as a Bachelor of Education. I then followed my interest in the evaluation of educational programmes by working with Dr. Richard Volpe at the Institute of Child Study and subsequently at the Sacred Heart Child and Family Centre. After three years of practical research experience I decided to continue my academic studies with a Master of Education from the University of Windsor. I have now been accepted into the Ph.D. programme at the Ontario Institute for Studies in Education and will be registering in January 1985.