A Meta Analysis of Comparative Research on the Effects of Desegregation on Academic Achievement

Ronald A. Krol

Western Michigan University

Follow this and additional works at: http://scholarworks.wmich.edu/dissertations

Part of the Bilingual, Multilingual, and Multicultural Education Commons

Recommended Citation
Dissertations. 2750.
http://scholarworks.wmich.edu/dissertations/2750

This Dissertation-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Dissertations by an authorized administrator of ScholarWorks at WMU. For more information, please contact maira.bundza@wmich.edu.
A META ANALYSIS OF
COMPARATIVE RESEARCH ON THE EFFECTS
OF DESEGREGATION ON
ACADEMIC ACHIEVEMENT

by
Ronald A. Krol

A Dissertation
Submitted to the
Faculty of The Graduate College
in partial fulfillment
of the
Degree of Doctor of Education

Western Michigan University
Kalamazoo, Michigan
December 1978
ACKNOWLEDGEMENTS

In writing this dissertation, many people have provided me with the moral support, guidance, and expertise that is necessary to complete a document such as this. I can not mention them all, but I would like to take this opportunity to thank those people who were most instrumental in providing me with the encouragement I needed.

First of all, I would like to thank my wife, Linda, who provided me with the emotional support and love that I needed during my years of graduate training and especially during the time this dissertation was being written. I truly could not have done it without her.

To my committee, I would like to express my genuine thanks for the patience, encouragement, and constructive criticism that was provided to me during the writing of this document. Without their technical expertise this dissertation would have been less than it is. Specifically, I would like to thank my chairman, Dr. Daniel Stufflebeam, for his guidance and advice in completing this dissertation. I would also like to thank him for financial support and the opportunity to learn which was provided to me at the Evaluation Center. I would like to thank Dr. James Sanders for the technical insight and constructive criticisms he provided. His viewpoints certainly made this a better
document. Finally, I would like to thank Dr. Subhash Sonnad who has been a constant source of encouragement throughout the completion of this dissertation.

In closing, I would like to thank Ms. Debbie Gareau and the interlibrary loan department at Western Michigan University; without their help I could not have accumulated the necessary data to complete this dissertation.

Ronald A. Krol
INFORMATION TO USERS

This was produced from a copy of a document sent to us for microfilming. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help you understand markings or notations which may appear on this reproduction.

1. The sign or “target” for pages apparently lacking from the document photographed is “Missing Page(s)”. If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure you of complete continuity.

2. When an image on the film is obliterated with a round black mark it is an indication that the film inspector noticed either blurred copy because of movement during exposure, or duplicate copy. Unless we meant to delete copyrighted materials that should not have been filmed, you will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., is part of the material being photographed the photographer has followed a definite method in “sectioning” the material. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.

4. For any illustrations that cannot be reproduced satisfactorily by xerography, photographic prints can be purchased at additional cost and tipped into your xerographic copy. Requests can be made to our Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases we have filmed the best available copy.

University Microfilms International
300 N. ZEEB ROAD, ANN ARBOR, MI 48106
18 BEDFORD ROW, LONDON WC1R 4EJ, ENGLAND

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
KROL, RONALD ALLAN
A META ANALYSIS OF COMPARATIVE RESEARCH ON
THE EFFECTS OF DESSEGREGATION ON ACADEMIC
ACHIEVEMENT.

WESTERN MICHIGAN UNIVERSITY, ED.D., 1973

C 1979

RONALD ALLAN KROL

ALL RIGHTS RESERVED

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
## TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td></td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Statement of the Problem</td>
<td>4</td>
</tr>
<tr>
<td>Definitions</td>
<td>7</td>
</tr>
<tr>
<td>Significance of the Problem</td>
<td>11</td>
</tr>
<tr>
<td>Purpose of the Study</td>
<td>15</td>
</tr>
<tr>
<td>Limitations and Assumptions</td>
<td>16</td>
</tr>
<tr>
<td>II</td>
<td></td>
</tr>
<tr>
<td>REVIEW OF THE LITERATURE</td>
<td>20</td>
</tr>
<tr>
<td>Introduction</td>
<td>20</td>
</tr>
<tr>
<td>Techniques and Methods of Aggregation</td>
<td>20</td>
</tr>
<tr>
<td>Meta Analysis of Research</td>
<td>39</td>
</tr>
<tr>
<td>The History of Desegregation - Integration in Education</td>
<td>43</td>
</tr>
<tr>
<td>Research on Desegregation in Education</td>
<td>49</td>
</tr>
<tr>
<td>Summary</td>
<td>67</td>
</tr>
<tr>
<td>III</td>
<td></td>
</tr>
<tr>
<td>DESIGN AND METHODOLOGY</td>
<td>69</td>
</tr>
<tr>
<td>Hypotheses to be Tested</td>
<td>69</td>
</tr>
<tr>
<td>Variables of Interest</td>
<td>75</td>
</tr>
<tr>
<td>Data Analysis Procedures and Steps</td>
<td>94</td>
</tr>
<tr>
<td>Summary</td>
<td>100</td>
</tr>
<tr>
<td>IV</td>
<td></td>
</tr>
<tr>
<td>RESULTS</td>
<td>102</td>
</tr>
<tr>
<td>Characteristics of the Data</td>
<td>102</td>
</tr>
<tr>
<td>Data Analysis: The Sample</td>
<td>102</td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
# TABLE OF CONTENTS (continued)

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Effects of Desegregation on Achievement</td>
<td>104</td>
</tr>
<tr>
<td>Effects of Desegregation on Achievement When Grade Level is Considered</td>
<td>105</td>
</tr>
<tr>
<td>Effects of Desegregation on Math and Verbal Skill Gains</td>
<td>109</td>
</tr>
<tr>
<td>Effects of Length of Desegregation on Achievement</td>
<td>112</td>
</tr>
<tr>
<td>A Comparison of the Voting Method and Meta Analysis Techniques</td>
<td>114</td>
</tr>
<tr>
<td>Effect Size of Designs With a Control Group vs. Effect Size for Designs Without a Control Group</td>
<td>116</td>
</tr>
<tr>
<td>Summary</td>
<td>117</td>
</tr>
<tr>
<td><strong>V DISCUSSION</strong></td>
<td></td>
</tr>
<tr>
<td>Review of the Problem and Procedures</td>
<td>119</td>
</tr>
<tr>
<td>Contrasting the Results With the Literature</td>
<td>123</td>
</tr>
<tr>
<td>Recommendations for Future Research</td>
<td>132</td>
</tr>
<tr>
<td>Implications for Educational Practice and Policy Making</td>
<td>137</td>
</tr>
<tr>
<td>How to Evaluate the Results of a Meta Analysis</td>
<td>139</td>
</tr>
<tr>
<td>Closing Statement</td>
<td>145</td>
</tr>
<tr>
<td>BIBLIOGRAPHY</td>
<td>148</td>
</tr>
<tr>
<td>APPENDICES</td>
<td>166</td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
# LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35</td>
</tr>
<tr>
<td>Advantages and Disadvantages</td>
<td></td>
</tr>
<tr>
<td>Different Methods of Accruing</td>
<td></td>
</tr>
<tr>
<td>Evidence from Research Studies</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>105</td>
</tr>
<tr>
<td>Effect Size for All Studies</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>107</td>
</tr>
<tr>
<td>Effect Sizes for Studies Involving</td>
<td></td>
</tr>
<tr>
<td>Students in Grades K-6 and Grades 7-12</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>108</td>
</tr>
<tr>
<td>Test of Means for Studies Using</td>
<td></td>
</tr>
<tr>
<td>Students in Grades K-6 Compared</td>
<td></td>
</tr>
<tr>
<td>to Studies Using Students in Grades 7-12</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>111</td>
</tr>
<tr>
<td>Effect Size(s) for Studies Which</td>
<td></td>
</tr>
<tr>
<td>Used Mathematics or Verbal</td>
<td></td>
</tr>
<tr>
<td>Achievement as a Dependent</td>
<td></td>
</tr>
<tr>
<td>Variable</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>111</td>
</tr>
<tr>
<td>Test of Means for Dependent</td>
<td></td>
</tr>
<tr>
<td>Samples Mathematics vs. Verbal</td>
<td></td>
</tr>
<tr>
<td>Achievement</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>112</td>
</tr>
<tr>
<td>Effect Size(s) for Studies Involving</td>
<td></td>
</tr>
<tr>
<td>Students Who Were in a Desegregated Environment</td>
<td></td>
</tr>
<tr>
<td>for One Year and Greater</td>
<td></td>
</tr>
<tr>
<td>Than One Year</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>113</td>
</tr>
<tr>
<td>Test of Means for Studies Involving</td>
<td></td>
</tr>
<tr>
<td>Students Who Were in Desegregated Environments</td>
<td></td>
</tr>
<tr>
<td>for One Year Compared</td>
<td></td>
</tr>
<tr>
<td>to Students Who Were in Desegregated Environments</td>
<td></td>
</tr>
<tr>
<td>for More Than One Year</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>116</td>
</tr>
<tr>
<td>Effect Size(s) for Studies With and</td>
<td></td>
</tr>
<tr>
<td>Without a Control Group</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>117</td>
</tr>
<tr>
<td>Test of Means - Control Group Designs</td>
<td></td>
</tr>
<tr>
<td>vs. Designs Without Control Groups</td>
<td></td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>DESCRIPTION</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Overview of the Cluster Approach</td>
<td>33</td>
</tr>
<tr>
<td>2</td>
<td>A Four-Celled Pre-Post Research Design</td>
<td>85</td>
</tr>
<tr>
<td>3</td>
<td>Example of Longitudinal Pre-Post Research Designs</td>
<td>86</td>
</tr>
<tr>
<td>4</td>
<td>Distribution of Effect Sizes</td>
<td>106</td>
</tr>
</tbody>
</table>
CHAPTER I

INTRODUCTION

Over the last 25 years this nation has spent tremendous sums of money to provide equal educational opportunity to people of all races. This money has been spent on court cases, desegregation plans, busing schemes and research in the area of equal opportunity, but we still do not have definitive answers as to the results of these efforts and expenditures. This research is an effort to provide more definitive guidelines as to the impact of efforts in the area of desegregation on academic achievement.

Many of the court decisions which have led to federal laws requiring integration were based on the assumption that desegregated schools were not equal. The underlying assumption was that black children would not be provided the opportunity to develop to their full potential since the facilities, staff, materials, books, and equipment were inferior to those provided in all-white schools. It was further assumed that if blacks were allowed access to similar facilities that they would experience intellectual growth which would be superior to that which they could achieve in all-black schools (Brown vs. Board of Education of Topeka, 1955; Bolling vs. Sharpe, 1954). 

1

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
However, the research relating academic achievement to integration has been contradictory. This is evident in a 1971 address to the American Educational Research Association by Walter Mondale on the subject of school integration.

What I have not learned is what we should do about these problems. I had hoped to find research to support or to conclusively oppose my belief that quality integrated education is the most promising approach. But I have found very little conclusive evidence. For every study, statistical or theoretical, that contains a proposed solution or recommendation, there is always another, equally well documented, challenging the assumptions or conclusions of the first. No one seems to agree with anyone else's approach. But more distressing: no one seems to know what works. As a result I must confess, I stand with my colleagues confused and often disheartened (AERA, February, 1971, New York City).

Recently, St. John (1975) reviewed the desegregation literature to determine the outcomes for children. She concluded the following about desegregation and academic achievement:

In sum adequate data have not yet been gathered to determine a causal relation between school racial composition and academic achievement. More than a decade of considerable research effort has produced no definitive positive findings. In view of the political, moral, and technical difficulties of investigation on this question, it is doubtful that all the canons of the scientific method will ever be met or a causal relationship ever established. Suggestive trends have been uncovered, however, as has one important negative finding: desegregation has rarely lowered academic achievement for either black or white children (St. John, 1975, p. 36).

She went on to say that if progress were going to be made in this effort to determine the causal relationship
between academic achievement and desegregation, we needed to pool, in some systematic manner, the original data from the studies and not rely on the words in the conclusions of the studies (St. John, 1975).

Recent methodological developments have provided a potential means to address some of the problems noted by Vice-President Mondale and Nancy St. John. In the *Educational Researcher* (November, 1976) article, Glass outlined a process that he labeled the "Meta Analysis of Research" which was intended to be a rigorous method for extracting knowledge from accumulated studies by integrating the results of many studies. In the same article, Glass described the differences between primary, secondary, and meta analysis of research.

To understand the goal of meta analysis one can think of the analysis of data at three distinct levels. The first level is "primary analysis." It is the analysis of data in a research study using conventional statistical methods. The next level is often referred to as "secondary analysis," which is the re-analysis of data for the purpose of answering the original research question with a different (or better) statistical technique or answering new questions from the old data. Examples of this type of activity are: the Mosteller-Moynihan secondary analysis of the Coleman study (1972); the Campbell-Erlebacher analysis of the
Ohio-Westinghouse Headstart evaluation (1970); and the Elashoff-Snow secondary analysis of Pygmalion in the Classroom (1971) to name three. Cook (1974) at Northwestern University has also been deeply involved in this type of research.

The third level of data analysis, meta analysis, refers to the analysis of analyses. It has specifically been used as a statistical analysis of a large collection of reported statistics from individual studies for the purpose of integrating the findings.

In his article Glass used meta analysis techniques to answer the question, "Does psychotherapy benefit those who receive such treatment?" This was done by analyzing 400 controlled studies and statistically creating composite control and experimental (treatment) groups. The two groups were compared by calculating a mean effect size for the control and the experimental group. He found that the treatment group mean effect size was about two-thirds of a standard deviation above the control group mean on the outcome variable.

Statement of the Problem

Because there are many opposing opinions on the question of the effects of integration on academic achievement, this study is directed toward a search for a causal connection between integration and academic achievement of those
minority children who were desegregated. Specifically it is directed toward answering the question, "Does desegregation lead to improved academic achievement of those minority children who were desegregated?" This investigation used the meta analysis technique as outlined by Glass (1976) for integrating reported research findings. In addition to the question of the general effect on academic achievement, there are also many concerns associated with specific differential effects of desegregation and achievement.

The first specific concern for this research was, "Are there differential effects of desegregation which are dependent on the educational level (grade level) of affected students?" This question is of particular relevance for if it can be established that students who experience desegregated education in their initial years of education benefited more than those exposed in the later years of their education, there are clear policy implications. Critics of previous research (Pinkney, 1968, 1970; Mayer et al, 1974) have charged that the reason "non significant differences" are often shown in research in this area is because of the difficulty of impacting on students during the latter years of their education. The critics contend that studies involving students in grades K-6 often show
greater effects (Anderson, 1966; St. John, 1975). Meta analysis could help to provide an answer to their conjectures.

The second specific concern is "Does desegregation lead to greater gains in certain subject areas than in others?" In the past research designed to answer this question has not been definitive although mathematics gains are reported more often than verbal or reading gains.

The third specific concern is "Is the amount of time in a desegregated educational environment related to achievement gains?" Here again, critics have charged that it is unfair to expect overnight changes in academic achievement based on only one year of desegregation. They contend that longer exposure to desegregated education will produce greater gains. This research is also designed to provide an answer to these critics.

In summary, this research will use the meta analysis technique to answer the question of the general effect of desegregation on academic achievement as well as questions of desegregation and achievement that are concerned with specific academic subjects, grade levels, and length of desegregation. It is expected that answers to these questions will be of interest to policymakers, legislators, educators, educational researchers, and scholars in the area of desegregation.
Definitions

Many of the concepts and phrases that will be used in the remainder of this document are used in different contexts by different authors. These concepts and phrases are defined at this point so that the reader will understand how these terms are used by the present author. The following is a discussion of the key words most often encountered in the literature and in this document.

Segregation

Segregation is probably the least confusing of the terms since it is used in everyday conversation in a context that is closely related to its use in much of the research in education. According to Dentler, et al (1967) "Sociologically, 'segregation' is a condition that tends to exclude or minimize association between groups, retarding the relationships, that do occur to subordinate-superordinate roles." (Dentler, 1967, p. 6). Although everyone would agree that a school which is 100 percent black, white, Puerto Rican, etc., was segregated, less unanimity exists for lower proportions. The criterion for a segregated school that will be adopted for the purposes of research in this study was suggested by Hauser (1964) and has been known as the 90-10 ratio. It holds that:
Schools whose student body is less than 10 percent Negro will be classified 'white' segregated; if less than 10 percent of the student body is white, the school will be considered 'Negro' segregated schools; the remainder will be referred to as integrated schools. (Hauser Report, 1964, p. 17)

Other definitions of segregation suggest that the determination of segregation be made based on the ethnic composition of the nation, state, or local community. In other words, if there are 50 percent black children in a given community, the composition of any given school should be 50 percent black. This approach is rejected as too cumbersome and unworkable for this research since it would be very difficult to determine the racial composition of the communities in which the research was performed. Therefore, the definition suggested by Hauser (1964) will be used in this study.

**Desegregation**

The term "desegregation" is often used interchangeably with integration although the most common meaning is that desegregation refers to the court or policy decision that requires that the schools which are segregated must integrate. In this study desegregation was interpreted to mean any purposive decision to integrate, whether it was court ordered or voluntary.

Therefore, "desegregation" refers to a policy decision while "integration" refers to the process of achieving the
goals of that decision. Integration is thus the process that automatically follows the decision to desegregate. Two questions that this research cannot adequately address are: "Does integration occur when school-wide racial composition of the minority group exceeds the 10 percent level? or "Does it occur only if the classrooms within the schools exceed the 10 percent level?" For purposes of this study integration was assumed not to have occurred until children were given the opportunity to meet and interact in the classroom situation. However, in much of the research on integration the degree of segregation within an integrated school is not clear.

Terms for Impetus of Integration

Other terms which arise in the literature are "induced integration," "forced integration," and "planned integration." Since most researchers did not define these terms, the definitions that appear below will be adopted for the current study.

Induced integration is a broad term used to describe integration which occurs because of court or legal action or threats to withdraw funds if schools are not integrated. Thus, schools are integrated due to an outside force which "threatens" some action if integration is not achieved.

In contrast, forced integration seems to indicate that a school is ordered to integrate on the basis of some legal
action that they opposed in court. That is, integration comes about by order of a specific court decision and many times is tied to a specific plan as outlined by that decision. This type of integration could also be called court-ordered desegregation.

**Planned integration** is a term used to describe the condition where a community or school board sees the need for integration of schools within the system and devises a plan to achieve the desired level of integration. This type of activity is different from the other two in that the motivation for desegregation and the ensuing integration comes from within the system and not from outside forces.

**Voluntary integration** is a term that can be applied to integration that is not the result of a court order. Therefore both induced and planned integration are variations of a voluntary integration scheme.

**Academic achievement.** Improved academic achievement is defined as gains on some specific commercially available achievement test such as the Iowa Test of Basic Skills (ITBS) or the Metropolitan Achievement Test (MAT). Such tests have traditionally been used as success criteria for academic achievement and therefore they are appropriate for the assessment of success in this research. These
tests are widely accepted by educators and legislators as indicators of what education should accomplish. These tests have also been the instruments used to measure achievement for most of the research in desegregation.

**Significance of the Problem**

One of the problems that has perplexed consumers of social science research is how to compare the findings of two or more existing studies that have conflicting results. This problem often arises because the studies have dissimilar designs, vary in overall research quality, use different instruments or methods to measure the dependent variables, or are conducted at different times.

It is also true that the consumer of these studies is not able to resolve the issue of conflicting results even with studies of similar design, overall research quality, etc. In other words, readers have no way of determining the "truthfulness" of the results of opposing studies based on quality, design, instrumentation, etc. Therefore, when readers are faced with the dilemma of resolving conflicts with regard to the results of two or more studies, they often revert to the system of values and beliefs they possess about the relationship among the variables in question. This type of behavior may be appropriate for the general public or even theory builders within the
social sciences, since their beliefs can be challenged and debated before a consensus arises with regard to the relationship of the variables debated.

However, as the research in question moves away from the theoretical end of the research continuum toward the applied research end, "truth" should not be simply a restatement of accepted values and assumptions. If we accept the wrong study as true, the consequences of intervention can have devastating effects on resource allocation and on the lives of persons in the programs. When individuals' lives will be directly affected by the applied research, we must not rely on personal belief systems and value judgments but should employ a thorough rational process of causal inference.

When we are using the results of research to formulate policies that will determine the design of social programs, we need to know the actual or most probable relationship between the variables if we are to design an appropriate intervention. If we are creating programs that are designed in the belief that X causes Y, and, in fact, Z causes Y, we are simply wasting the taxpayers' money since the program is likely to have no effect.

It is evident that mistakes can be made in basic and applied research and in evaluation and that these mistakes can be costly in both human and monetary terms. Improved
methods of improving the accuracy of decisions that are based on research and evaluation are thus needed.

The meta analysis technique has a great potential for bringing some rationality to decision areas that are particularly controversial and inherently political such as desegregation. While it is well recognized that meta analysis, or any other research technique, cannot or will not replace the current system of making such decisions, it is hoped that it can help put claims and counterclaims into proper perspective.

The policy decisions which could be impacted by a meta analysis of the studies on desegregation include the following:

1. How should money for compensatory education be spent most effectively? For example, if integrated black children make good gains in math but not in verbal skills, it would seem probable that more federal money should be spent on delivering compensatory math than on compensatory programs designed to improve verbal skills and that additional money should be invested in research and development in the area of reading and writing.

2. At what grade level(s) should efforts to integrate education be particularly emphasized? If children who experience integrated education in earlier grades seem to benefit more than those children who experience integration
in the later years of their education, this could indicate a direction for future federal efforts to implement integrated education in the lower grade levels.

3. What are the long-term effects of desegregated education on black and white children? Although this meta analysis research is not designed to investigate the long-term effects of desegregation, it might be possible to track the educational progress of those children who experience integrated education to find out if their educational career goals or life goals differ from those who experience segregated education. In other words, there may be long-term and worthwhile effects of desegregated education that have policy implications.

4. Finally, if the effects of desegregation on academic achievement seem to be minimal or nil, federal policymakers should ask WHY? The assumption has been that black children will benefit by an integrated educational experience; if they do not, it might indicate that research should go to the next lower level of the educational process. That is, it might not be sufficient to examine the effects of desegregation at the district or school level. Further efforts might be directed at finding out what happens in the classroom and in the particular interactions between teacher and integrated child. This meta analysis effort is at the macro level and cannot answer the "why" aspect of the outcomes.
Purpose of the Study

The primary question of this study is "Does desegregation lead to improved academic achievement? (under all circumstances? with all subject matter?)"

The specific objectives of this study are as follows:

1. To determine the overall effect of desegregation on the academic achievement of the desegregated group.

2. To determine if there were differential effects for desegregation which were dependent on the educational level of affected pupils.

3. To determine whether the achievement gains in mathematics due to desegregation were greater than in verbal and reading achievement gains. Some of the past research seems to indicate a trend toward improved achievement in math but not in verbal skills.

4. To investigate the effects of desegregation on academic achievement when length of exposure to a desegregation environment was taken into account. The critics have charged that the reason desegregation research has often showed no significant gains in achievement is that the research was conducted over a one-year period. The implication is that the effects of integration are slow acting and that the results cannot be observed or measured after only one year. This research compared the effect of
desegregation on academic achievement after length of desegregation was taken into account.

In addition to these primary objectives, the potential of using meta analysis to address major policy issues is being assessed in this study. Specifically, meta analysis techniques were compared to the more common voting method of aggregation, and designs with and without control groups were compared to determine their relative precision in identifying the effects of desegregation.

Limitations and Assumptions

The reader should be aware of several assumptions and limitations that bear on the adequacy of this study for addressing the question of desegregation on academic achievement.

Assumptions

The following assumptions were made after a thorough review of the literature and the practical problems of completing a study such as this in a reasonable amount of time:

1. The study focuses on the achievement gains of black children only since this is one important group that is served by desegregation programs.
2. There may be beneficial reasons for desegregated education that are not measured by traditional measures of academic achievement. These could include improved self-concept of black children, improved relations between white and black children, as well as the long-term impact of desegregated education on black children's educational goals and plans. These and others were not investigated because of the need to keep this study to realistic levels.

3. It is assumed that the method used to bring about integration has little to do with the impact on academic achievement. There is little research evidence to suggest that the method used to bring about integration will have a significant impact on the outcomes of integration. In fact, the only study which considered the method of bringing about desegregation was done by Hsia (1971) in Evanston, Illinois. In this study he compared the achievement of "walking" and "busing" in a desegregation plan and found a slight advantage in favor of "busers" although the advantage was not statistically significant.

Limitations

In addition to these assumptions there are several other limitations imposed by the meta analysis method of integrating studies:
1. Since this research is limited to published studies only, it is likely that there is some bias in favor of positive effects of segregation. Most researchers and scholars would contend that studies that show statistically significant results are more likely to be published than those which show no significant differences between groups. This author has not been able to locate a study to substantiate that assumption although it does seem tenable.

2. A second limitation is that many studies are eliminated on the basis of the sorting criteria needed to accomplish a meta analysis in a given topical area. Therefore, studies which did not provide the necessary statistics, were of designs that were excluded, or were descriptive in nature were excluded when in fact they made significant contributions to the literature.

3. A third problem when doing meta analysis is deciding whether or not a set of studies ought to be considered a universe or a sample. Most people, including Glass (1977), have treated the studies assembled as a universe but then treated them as a sample in their statistical treatment of the data. This study treats the set of studies as a sample since the formula used for analysis was adopted from Glass.

4. A fourth problem when doing meta analysis is the lack of common metrics for the measures used in the various...
primary studies on the topic. This is a problem because
different constructs are sometimes studied under one topic
area. In addition, the statistics which are used to assess
a relationship between two or more variables may not be
directly comparable across studies. This is a problem
because the viability of Glass's (1977) transformations
are not well established. This was not a significant
problem for this study since the dependent variables were
measured by similar commercially available tests.

5. Finally, the validity of using various methods to
salvage as many studies as possible has not been estab­
lished as of this date. This study will use three different
methods suggested by Glass (1977), but the validity of
these methods has not been tested.
CHAPTER II

REVIEW OF THE LITERATURE

Introduction

This section is divided into two major parts. The first part is a review of previous literature on techniques for accumulating and integrating evidence from research studies. The second part is a review of the history of desegregation and the subsequent studies on integration and its effects on children after desegregation.

Techniques and Methods of Aggregation

Probably the most common method of extracting information for policy decisions is a review of the previous literature. Vice-President Mondale indicated that this is one input used by decision makers at the national level.

According to Light and Smith (1971), a review of the literature involves three steps.

First, all the relevant empirical studies are gathered together. Second, studies with inadequate sampling procedures, measurement and instrumentation, or methods of analysis are identified and discarded. Third, the conclusions from the remaining studies are assembled and compared in an effort to find consistent results. The third step often encounters contradictions; similar studies frequently produce contrary results. (Light and Smith, 1971, p. 430)

20

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
If the review of the literature is strictly theoretical or an exercise to acquaint oneself with a body of knowledge, these inconsistencies can be confusing and disconcerting. When the purpose of the review is to acquire knowledge to be used to develop specific policies and programs, these inconsistencies can paralyze attempts by (policy level) decision makers.

Glass (1976) pointed out two major problems with this method of extracting answers from research. First, this approach takes design and analysis too seriously. That is, it is debatable if we should discard the results of half the studies on these grounds. The point is that by discarding many of the studies because of design flaws we are probably also discarding valid results, which could contribute to our overall analysis. Glass concluded:

At any rate, I believe the differences to be so small that to integrate research results by eliminating the 'poorly done' studies is to discard a vast amount of important data. (Glass, 1976, p. 4)

The second point raised by Glass was that it is not realistic to expect a researcher to read five hundred studies and extract a simple summary from them. He contends that the composite meaning cannot be prepared by narrative, discursive review processes without the aid of specific techniques for organizing, depicting and inter-relating data. This would be like trying to understand the results of 500 test scores by merely scanning the data
for trends. The point is that the meaning cannot be fully grasped without the aid of techniques for organizing, depicting, and interrelating the data for these 500 test scores.

In addition to a systematic review of the literature, there have been at least five other methods used to combine research results. Light and Smith (1971) outlined four previous methods plus their recommendations for combining studies. These can be described as (a) the Listing Method, (b) the Consistency Approach, (c) the Averaging Method, (d) the Voting Method, and (e) the Cluster Approach. For a summary of the advantages and disadvantages of these methods refer to Table 1.

The Listing Method

This method involves a simple listing of factors that have been shown to have an effect on a dependent variable in at least one of a group of studies.

A study that used the Listing Method was completed in 1974 (Jamison, Suppes and Wells, 1974). This study investigated the effectiveness of instructional television and alternative media. This study also investigated the effectiveness of instructional media, programmed instruction, and computer-assisted instruction. The effectiveness of these media were examined from an achievement standpoint
as well as the affective impact of the various types of instruction.

This study employed variables which have shown to have an effect on traditional classroom instruction. They concluded that teacher verbal ability seems to be an important variable and that small classes seem to improve the cognitive and affective performance of young children.

The conclusions with regard to media instruction effectiveness were based on a review of the literature process and only studies that were considered to be of high quality were accorded a vote. By this process Jamison et al. (1974) concluded that radio and TV instruction were as effective as traditional methods, that computer assisted instruction was effective when used to improve achievement scores of disadvantaged students and the researchers concluded that with regard to programmed instruction that findings of no significant difference dominate the literature.

Another study (Clark, 1971) that employed the listing method attempted to answer the question, "To what extent can the large body of experimental research in concept attainment be used to suggest an empirically derived set of prescriptions for teaching certain types of classroom concepts?"

To answer this question Clark used a complicated classification system for the studies and then derived
axioms and proportions based on the listing method, although he did use the voting method to compile his list. In other words, to make the list of teaching prescriptions, more studies had to show effectiveness than showed no differences. As a technique, method, etc., was declared a winner by this process, it was added to the list of statements about effective teaching.

Clark's work is a thorough review and condensation of the research on teaching and the classification system used is particularly noteworthy. However, the applicability of the 61 prescriptions derived by this process is not so evident. At best, they should serve as research hypotheses for future research in the area of teaching effectiveness.

The major problem with the listing approach is that it produces cumbersome long lists and it does not often differentiate those factors which have been identified many times from those which were found important in only one study, and some of the factors that make the list are likely to contradict other statements on the same list.

The Consistency Approach

This approach could be conceived of as the opposite of the first since this method uses only those studies that meet predetermined criteria while the listing methods
use all or most of the studies in a particular area. When the researcher uses this approach he creates consistency by establishing theoretical, qualitative and congruence criteria which will often eliminate all but a few of the studies that link the variables of concern. Thus the major weakness of this approach is that too much information is discarded in the name of consistency.

It is difficult to cite a particular study that relied exclusively on this method however. Many studies use a combination of this method and the listing or voting method to arrive at a conclusion.

In addition, it is often difficult to determine that such a method was employed since the author seldom fully describes the method or the criteria used to achieve consistency.

The Averaging Method

The averaging method attempts to make the integration process more systematic than the first two. This method consists of computing overall averages for relevant statistics across a complete set of studies. The summary measure can be a mean, but more often it is a median or some other measure which offers some protection against extreme values. An example of the use of this approach would be to take five simple correlations from five studies between the same
two variables and to report the median correlation value. As Light and Smith (1971) pointed out, this is a needed attempt at quantification, but this estimate throws away precisely the information we want most. That is, by averaging correlations we are not getting at a true estimate of the treatment effect for we may have some high positive correlations, some high negative correlations, and the median may not really be representative of an average expected treatment effect and thus not useful for the policymakers.

An example of the use of this method is provided by Light and Smith (1971).

For example, a researcher might take five values of the single correlation between a student's family income and his college aspirations, found in five different studies, and report the median value." (Light and Smith, 1971, p. 199).

Other examples of the use of the median as a summary measure can be found in Fielder (1971).

The Voting Method

This method is considered to be the best and most systematic of the first four methods. Light and Smith (1971) described the voting method in these words:

All studies which have data on a dependent and a specific independent variable of interest are examined. Three possible outcomes are defined. The relationship between the independent variable and the dependent variable is either significantly positive, significantly negative, or there is no
significant relationship in either direction. The number of studies falling into each of these three categories is then simply tallied. If a plurality of studies falls into one of these three categories with fewer falling into the other two the modal categorization is then assumed to give the best estimate of the direction of the true relationship between the independent variable and the dependent variable. (Light and Smith, 1971, p. 433)

A particularly ambitious study using the voting method was done by Schramm (1962) comparing the relative effectiveness of conventional classroom teaching and instructional television. In this aggregation study he used 393 studies in which instructional television had been compared with other classroom teaching methods. His goal was to address the effectiveness of instructional television as compared to conventional teaching methods in the areas of (a) motivation, (b) subject matter differential effects, (c) grade level differences in effectiveness, and (d) attitude of students and teachers.

An interesting and thorough study was done by Jackson (1975) on the effects of grade retention. The voting method was used to aggregate the differences in achievement and adjustment in order to arrive at a measure of effectiveness. On the basis of this process, Jackson concluded:

There is no reliable body of evidence to indicate that grade retention is more beneficial than grade promotion for students with serious academic or adjustment difficulties. This conclusion should not be interpreted to mean that promotion is better than retention, but rather, that the accumulated
research evidence is so poor that valid inferences cannot be drawn concerning the relative benefits of these two options. (Jackson, 1975, p. 627)

There are several problems with the voting method. First of all, the voting method disregards sample size. Large samples produce more statistically significant findings than small samples.

A second deficiency of the voting method is that it discards valuable descriptive information. This method does not tell one if method A is better than method B by a wide margin or by a nose and it does not make allowances for interpreting the strength of experimental effects. Listings of statistical significance or insignificance tells little about the strength or importance of a relationship.

Another major problem with the voting method has been termed Simpson's Paradox (Blyth, 1972), and it can lead to misleading or even incorrect conclusions about an effect size or direction. Basically, it states that it is possible to arrive at opposite conclusions by using the voting method and by combining the raw data. Perhaps an example given by Glass (1977) in an unpublished manuscript will illustrate the point.

Imagine that Researcher A is conducting a study of the effect of amphetamines on hyperactivity in sixth-grade children. In A's study, 110 hyperactive children receive the amphetamine and 70 receive a placebo. After six weeks treatment, each child is rated as either improved or worse. The following findings are obtained: (Glass, 1977, p. 17)
STUDY A

<table>
<thead>
<tr>
<th></th>
<th>Amphetamine</th>
<th>Placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improved</td>
<td>50</td>
<td>30</td>
</tr>
<tr>
<td>Worse</td>
<td>60</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td><strong>110</strong></td>
<td><strong>70</strong></td>
</tr>
</tbody>
</table>

The improvement rate for the amphetamines exceeds that for the placebo: .45 vs. .43

Suppose that Researcher B is studying the same problem at a different site and obtains the following results:

(Glass, 1977, p. 17)

STUDY B

<table>
<thead>
<tr>
<th></th>
<th>Amphetamine</th>
<th>Placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improved</td>
<td>60</td>
<td>90</td>
</tr>
<tr>
<td>Worse</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td><strong>90</strong></td>
<td><strong>140</strong></td>
</tr>
</tbody>
</table>

Again the improvement rate for amphetamines is superior to that for the placebo: .67 vs. .64. By the "voting method" of aggregation, the score would be 2-0 in favor of amphetamines. However, an aggregation of raw data produces the opposite conclusion:

STUDIES A & B COMBINED

<table>
<thead>
<tr>
<th></th>
<th>Amphetamine</th>
<th>Placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improved</td>
<td>110</td>
<td>120</td>
</tr>
<tr>
<td>Worse</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td><strong>200</strong></td>
<td><strong>210</strong></td>
</tr>
</tbody>
</table>

The improvement rates for placebo now exceeds that for amphetamines: .55 for amphetamines vs. .57 for placebo. (Glass, 1977, p. 18)

As one can see, there is a problem of interpretation. Which method of aggregation is correct? Obviously they...
cannot both be correct since they lead to contradictory conclusions. As one investigates this paradox further, it can be shown that (a) the paradox has nothing to do with statistical significance, and (b) the basic problem is related to the problems of unbalanced experimental designs.

It is well to point out that Simpson's Paradox is more than a fascinating theoretical problem, for the practical consequences of this paradox are quite serious. If the researcher is not aware of this problem, he will likely formulate incorrect decisions or policies if he uses the voting method. (See Bickel, Hammel, and O'Connel, 1975, and Gardner, 1976.)

The Cluster Approach

The cluster approach is a technique for aggregating the results of studies which is derived from a sampling technique called cluster sampling. Basically, it assumes that populations can be broken into small, identifiable sub-populations known as clusters. These clusters are not random samples from the population, but they are natural groupings within the population which differ in broad systematic ways. For example, a neighborhood group may be a cluster sample out of a larger sample of the city population or classrooms within a larger study could be taken as cluster samples if they tend to differ in some
identifiable and systematic manner. The following quotation should help to illustrate how clusters could be selected.

For example, when S. S. Grote research questions are directed towards the effects of tracking, each individual tracked classroom can be taken as an independent cluster. On the other hand, when early reading instruction is the focus of research, the reading groups within classrooms are the logical choice for a clustering unit. The crucial point is that whatever unit is chosen to be a cluster, it should be the natural focal point or molar unit of whatever educational process we are investigating. (Light and Smith, 1971, p. 211)

A cluster may not be a complete study but only a part of a larger study. This differs from the voting method where the complete study is always the unit of analysis. Thus, a study will often contain several clusters, and the cluster is the unit of analysis.

In order to aggregate research results using the cluster approach, one looks for natural differences which occur within the population of interest and then treats those clusters as individual populations and looks for differences in treatment effects between those clusters. At this point it would be well to outline the cluster approach as described by Light and Smith (1971).

1. The cluster approach requires access to the original data from the studies of interest, so the availability of these data must be determined prior to beginning the procedure.

2. The quality of the studies to be used in the cluster aggregation approach must be reviewed prior to beginning. The studies to be used should meet the following criteria.
a. Subjects must have been selected from a known and precisely definable population.

b. The dependent variables and those independent variables which are measured must be measured in the same way or subject to suitable conversion methods.

c. Overall quality of the studies selected must be comparable.

3. The cluster approach attempts to answer the question in what ways do the clusters differ from one another and of the ways in which they do differ, which can be identified with differences among the clusters?

4. The next step is to classify the clusters into the ways they can differ systematically. A group of clusters can differ in at least five ways:

a. The means of their variables

b. The variances of their variables

c. In the relation between the independent and dependent variable

d. In the treatment by trait interactions

e. In the subject by group interactions or in the ways a subject is effected by the group they are part of often known as contextual effects.

An abbreviated overview best captures the logic of the cluster approach. Figure 1 (Light and Smith (1971) is a flow diagram of the logic and methods for a cluster approach to the aggregation of studies.

From the discussion and the diagram of the overview of the cluster approach, it can be seen there are several problems and restrictions in using this method for integrating research results.
OVERVIEW OF THE CLUSTER APPROACH

Test for differences among clusters in:
- a) Means
- b) Variances
- c) Covariate Relations
- d) Subject-Treatment Interactions
- e) Contextual Effects

One or More Differences Found

Search for explanation:
- a) Selection Effects
- b) Amplification Effects
- c) Sensitization
- d) Different Proportions of Types of Subjects in Clusters
- e) Contextual Differences
- f) Unmeasured Variables
- g) Other

No Explanation Found

Explanation Found

Adjust Away Explained Cluster Differences

Cannot Combine Data From Clusters

Combine Data From Clusters

(Source: Light and Smith, 1971; Figure 4, page 463)
First of all, the cluster approach requires access to the raw data. This is often overly restrictive if not an impossible requirement for this type of activity as will be discussed under the methodology section. The restriction of using raw data limits the utility of combining results to such a level that the original goal of aggregation will be lost.

Secondly, this procedure imposes many statistical and methodological restrictions that severely limit the number of studies that will become data for a cluster approach to aggregation.

Finally, as can be seen from the diagram, there are many studies that cannot be included in a cluster analysis or worse yet, all of the studies could be excluded because clusters differ and suitable explanations for differences cannot be located.

Light and Smith (1971) concluded that the cluster approach has great merit when compared with the other four methods of aggregation. However, the "Meta Analysis of Research" method advocated by Glass and by this study has much greater utility, and thus is superior to the cluster approach for most applications. This is especially true due to the necessity of raw data and statistical assumptions imposed by the cluster approach.

A more thorough discussion of meta analysis techniques will serve to demonstrate the superiority of this method.
<table>
<thead>
<tr>
<th>Definition</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Review of Previous Literature</strong></td>
<td>o Researcher becomes conversant with previous work in the area</td>
<td>o Difficult to reach a conclusion by just reading a great number of studies</td>
</tr>
<tr>
<td>A systematic and thorough method of comparing results of previous research by utilizing the conclusions of those studies to arrive at an overall general conclusion</td>
<td>o Probably the most economical method</td>
<td>o Researcher's value system or preconceived notions likely to intervene on final conclusion</td>
</tr>
<tr>
<td></td>
<td>o Poorly done studies can be systematically excluded</td>
<td>o Many relevant studies are often discarded for quality reasons</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o Can lead to confusion and distrust of previous research</td>
</tr>
<tr>
<td>Definition</td>
<td>Advantages</td>
<td>Disadvantages</td>
</tr>
<tr>
<td>--------------------------</td>
<td>----------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>The Listing Method</strong></td>
<td>o Can produce long lists of possible variables for future research</td>
<td>o Produces long and often cumbersome lists that yield little insight for researchers</td>
</tr>
<tr>
<td></td>
<td>o Can be accomplished at same time as review of literature</td>
<td>o Does not differentiate factors which have been identified once from those identified many times</td>
</tr>
<tr>
<td></td>
<td>o Can be used in conjunction with another integrating method</td>
<td></td>
</tr>
<tr>
<td><strong>Consistency Approach</strong></td>
<td>o Can provide some quality control so that studies combined are often methodologically sound</td>
<td>o Too much information is discarded to achieve consistency</td>
</tr>
<tr>
<td></td>
<td>o Requires that researcher read all of previous studies in order to apply criteria</td>
<td>o Researcher biases and values are likely to influence which studies are discarded</td>
</tr>
<tr>
<td>Definition</td>
<td>Advantages</td>
<td>Disadvantages</td>
</tr>
<tr>
<td>---------------------</td>
<td>----------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Averaging Method</strong></td>
<td>o Provides a quantitative measure</td>
<td>o Computed measure can produce misleading results if data are based on related but conceptually different outcome measures</td>
</tr>
<tr>
<td></td>
<td>o Utilizes all studies</td>
<td>o Concept of averaging is probably not appropriate for the purpose of delivering a final answer</td>
</tr>
<tr>
<td></td>
<td>o More systematic than previous methods, including review of literature</td>
<td>o Disregards sample size</td>
</tr>
<tr>
<td></td>
<td>o Can provide direction for new research or replication of previous research</td>
<td>o Disregards quality of studies</td>
</tr>
<tr>
<td></td>
<td>o Can be used in combination with other methods</td>
<td>o Often discards valuable descriptive information</td>
</tr>
<tr>
<td><strong>Voting Method</strong></td>
<td>o Can be used in combination with other methods</td>
<td>o Does not allow for interpreting the strength of relationships</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o Can lead to misleading or incorrect conclusions (Simpson's Paradox)</td>
</tr>
</tbody>
</table>

Table 1 (cont'd)
<table>
<thead>
<tr>
<th>Definition</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster Approach</td>
<td>o Uses a quality measure as a criteria to decide which studies are to be</td>
<td>o Requires access to original data</td>
</tr>
<tr>
<td></td>
<td>considered</td>
<td>o May be hard to identify and describe known clusters</td>
</tr>
<tr>
<td></td>
<td>o Most systematic and statistical of the six methods</td>
<td>o Imposes many statistical and methodological restrictions</td>
</tr>
<tr>
<td></td>
<td>o Has great relevance for researcher involved in original research. Can</td>
<td>o Information is lost due to sorting criteria</td>
</tr>
<tr>
<td></td>
<td>attempt to determine treatment effects for subgroups of population</td>
<td>o Lengthy and precise process</td>
</tr>
<tr>
<td></td>
<td>o Has great potential for statistical analysis and precise determination of</td>
<td></td>
</tr>
<tr>
<td></td>
<td>effects</td>
<td></td>
</tr>
</tbody>
</table>
Meta Analysis of Research

There is not a great deal of literature or research dealing with meta analysis of research since Glass (1976) first coined the concept. However, a similar aggregation technique was also discovered simultaneously and independently by Rosenthal (1976) in his analysis of the experimenter effect in behavioral research.

In his research Rosenthal (1976) relied on the statistic "d" (Cohen, 1969) as a common measure of treatment effect. Rosenthal (1976) defined "d" as the difference between the means of the two groups being compared, divided by the standard deviation common to the populations. This statistic was chosen because:

1. It permits the researcher to compare the magnitudes of effects for a large variety of measures.
2. It is free from the particular scale of measurement and allows one to speak of effects measured in standard deviation units.
3. It is simple to understand.
4. It lends itself to such a large proportion of studies in social science which involve the comparison of an experimental group mean with a control group mean. (Rosenthal, 1976, p. 444-445)

A comparison of Rosenthal's (1976) methods with those of Glass and Smith (1976) indicates that both drew heavily on the work of Cohen (1969) as treated in Statistical Power Analysis for the Behavioral Sciences since both used the
statistic "d" as a measure of treatment effectiveness or effect size.

The general goal of meta analysis of research is a quantitative aggregation of findings without a loss of information which results from the quality control techniques of selection common with other methods of integration. Meta analysis also attempts to describe the relationship among findings and characteristics of the studies. The Glass and Smith (1976) procedure for meta analysis of psychotherapy outcome studies was explained as follows:

1. First a common measure of treatment effectiveness was defined. This measure, called the "effect size" was defined as the difference between the means of the psychotherapy and the control group divided by the control group standard deviation. (Of course all studies without control groups were excluded prior to this time.)

2. An extensive review of the literature was performed which uncovered almost 400 control group studies of psychotherapy.

3. Each study was described in quantitative or quasi-quantitative terms in several different ways. This included the mean difference between treated and untreated groups by tabulating various crosscuts of the data by type of outcome and by graphing effects of therapy over time.
By using these techniques, Glass and Smith (1976) concluded that the average effect size was 0.68 standard deviations and that the effect sizes of four different types of therapy were not greatly different.

Certainly there are problems inherent in meta analysis as there are in any other aggregation technique, but the point is that by using this method, one can retrieve a great deal more information out of past research than by any other method.

The problems which must be faced in any meta analysis can be grouped under the following categories:

1. Inaccessibility of a great deal of relevant studies.

2. Lack of standardization of methods and reporting formats in research studies.

3. Poor quality of some journal articles with regard to statistical information; often the means and variances are not even reported.

4. Difficulty in deciding when studies are similar enough to aggregate.

5. Determining whether the treatments are equal. For instance, if a person gets therapy one hour per week and another group receives the same treatment four hours per week, what can be done to integrate these studies?

These problems were mentioned at this time lest the reader feel that meta analysis is without its problems.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
They will be discussed in much greater detail in the design and methods section.

Meta analysis as a technique of interpreting research findings is still in its formative stages. Therefore, each meta analysis effort must be approached differently depending upon the type of data available (correlational, significance level only) and the type of designs involved. At this point, the responsibility for selecting appropriate methods and making intelligent decisions about how and what to integrate is still left up to the researcher. There are no definitive texts on the methodology of meta analysis, although Glass's (1977) manuscript is a significant contribution in that area.

At this juncture it would be worthwhile to turn to a review of the literature with regard to the history of desegregation and attempts at the integration of literature in this area so that the importance of this study can be shown. It is also hoped that the reader will gain a greater insight into the relevance of trying to integrate research and evaluation studies for the purpose of making policy decisions at the national level. This is particularly apparent in the area of desegregation, but it is equally true of other difficult decisions which are made at the national level, such as ways to attack alcohol and drug abuse, crime, and highway safety decisions.
The History of Desegregation-Integration in Education

There have been few Supreme Court decisions that have shaped the constitution as much as the ruling requiring the states to provide desegregated education. Those who have lived through this decision will remember the marches, the confrontations, the violence, and the bitterness of this controversy more vividly than any other institutional change in their lifetime. This controversy literally shook this nation from north to south. It was so far-reaching that it had implications for all levels of education; it was so pervasive that it has forced northern and southern communities to institute integration plans; it was so monumental that case after case was paraded before the Supreme Court with regard to desegregation plans and to challenges to those plans. The main significance of the Supreme Court decisions on desegregation in education is that they have had more impact on this nation than any other decision during this century. Since this was such a tumultuous, emotional time in our history, it is best to review the history of the Supreme Court decisions and desegregation so that it can be placed in the context of its time and so one can see why this controversy surrounding this decision is very much alive today. (Witness: Events in Boston, Massachusetts, 1975-1977.)
The most famous and notable decision regarding school desegregation is the Brown vs. Board of Education, Topeka, Kansas, 1954 and 1955. In the 1954 decision, the court quoted the language of the lower court in the Kansas case.

Segregation of white and colored children in public schools has a detrimental effect upon the colored children. The impact is greater when it has the sanction of the laws; for the policy of separating the races is usually interpreted as denoting the inferiority of the Negro group.

A sense of inferiority affects the motivation of a child to learn. Segregation with the sanction of the law, therefore, has a tendency to retard the educational and mental development of Negro children and to deprive them of some of the benefits they would receive in a racially integrated school system. (Brown vs. Board of Education 347, U. S. 483 at 494 from Bolner/Brown, 1974)

The court went on to say that "in the field of education, the doctrine of separate but equal has no place. Separate educational facilities are inherently unequal."

This decision declared unconstitutional the laws or statutes requiring or permitting separate public schools for white and minority children.

The second decision in the Brown vs. Board of Education came in the proceedings of the same court in 1955. At that time the court expanded the previous decision by placing the primary responsibility for abolishing segregated schools with local authorities. It also said that the changes necessary to correct dual school systems were to be made "at the earliest practical date" and "with all deliberate speed." This second case approached for the
first time the manner in which dual school systems were to be ended. It did not advocate a simple line of attack so much as a practical flexibility (Mills, 1973).

In spite of the Supreme Court's decree of 1955, the process of school desegregation did not come about very quickly. In 1964 the Supreme Court, disappointed with the lack of progress, announced in its Griffin decision (Griffin vs. County School Board of Edward County, 84 S. Ct. 1226) that the "time for mere deliberate speed has run out." (Webster, 1977)

However, despite this proclamation most school systems did not make significant progress until after the 1964 Civil Rights Act was passed. This was a far-ranging act which prohibited discrimination in employment, housing, and education, and permitted the United States Department of Health, Education and Welfare to withhold federal funds from segregated school systems. The impact of this provision was to stimulate desegregation plans in many parts of the country. Most of these plans were based on freedom of choice, or a variation known as free transfer.

Another outcome of this Act was the famous Coleman Report (1966) which was commissioned by the United States Office of Education under the 1964 Civil Rights Act. Coleman conducted a nationwide survey to determine the extent of educational opportunity and inequality within
public institutions in the United States. The report, which was a result of this commissioned study, contained evidence of the extent of school segregation in all parts of the country. This report had significant impact on the thinking of the courts, civil rights leaders and it has been the subject of much debate within research circles. This report suggested the following state of affairs:

1. Ten years after the Brown decision, there had been little progress in desegregating schools.

2. There were significant achievement differences between white and minority students in reading, verbal, and mathematics performance.

3. Schools attended by blacks were inferior with respect to physical facilities, teacher qualifications, class size and extracurricular activities.

4. The gap in achievement between blacks and whites got larger as children progressed through the school system.

Coleman also suggested that integration would help the achievement of black students without hurting the achievement of white students. Although the methodology and conclusions of the study have been much debated since that time (St. John, 1970) there is no denying its impact on policy decisions which were made as a result of the Coleman study. The decision to use busing as a means to
achieve racial integration was one direct result of the conclusions of this report.

"The Swan vs. Charlotte-Mecklenburg Board of Education, 402 U. S. 1, 18 (1971) decision represents the fullest statement to date by the Supreme Court on the use of busing to eliminate the effects of a dual school system." (Mills, 1973, p. 47) This case related to the Charlotte, North Carolina Public Schools and its failure to present an acceptable desegregation plan on the basis of earlier District Court decisions. After the District Court rejected all the plans submitted by the School Board of Charlotte, it formulated a plan which it believed conformed to earlier Supreme Court decisions and would result in a completely desegregated system.

The final plan adopted by the District Court stated the school district of Charlotte must use whatever strategies or methods that were necessary to achieve a totally desegregated school system. The School Board of Charlotte challenged this decision and appealed to the Supreme Court. On April 20, 1971, the Supreme Court upheld the lower court's plan to achieve desegregation in Charlotte and the stage was set for similar cases in other communities.

In 1973, the Court delivered a split decision concerning cross-district, cross-county busing in Richmond, Virginia. Due to the tie vote (4-4) the lower court's...
opinion was upheld. Recently, a similar decision was reached with regard to cross-district busing in Detroit.

From these very recent decisions, it is clear that busing is still a controversial means of achieving integration. There are still protests and boycotts by students and parents of the Boston Schools; James S. Coleman has recently stated that he believes he made a mistake in suggesting that busing was a viable method of achieving integration.

In addition, over the past ten years, vigorous differences of opinion on the relative merits of planned desegregation (and particularly forced busing) have been evident in the literature. These debates center on the effectiveness of busing and associated strategies as techniques for improving minority achievement, aspirations, self-concept, educational and occupational opportunities, and as techniques for facilitating race relations. As was pointed out earlier, a great deal of the Supreme Court's decision in 1954 was based on the premise that black students would not achieve as well as they could in segregated school environments. It is for that reason and so that the task of this study is within realistic limits that the review of previous literature on desegregation will concentrate on the relationship between induced desegregation and academic achievement.
The most striking aspect of the review of the previous research in this area is that it would be relatively easy to support either the pro or con side of this question by selecting the appropriate studies. Secondly, one soon discovers that most of the studies concerned with this topic could be discarded on the basis of inadequacies of design, methodology, sampling, instrumentation and almost every other reason research is criticized and summarily ignored. This is true because most, if not all, of the research concerned with achievement and desegregation was undertaken because a policy or legal decision was made to desegregate a school system. Therefore, most of the research can be classified as either applied research or evaluation, and as such, much of the control remained outside the province of the researcher. Research in any area that has applied policy implications and is conducted in a field setting will frequently have more methodological flaws and threats to internal and external validity (Campbell and Stanley, 1966) than research which is conducted in a laboratory.

Research on Desegregation in Education

The literature on the effects of desegregation on achievement clearly indicates that there has been a rather heated and continuing debate between certain researchers in this area. This review will concentrate on the most
frequently cited persons in this debate as well as the studies most frequently cited by them. There are several excellent reviews of the literature which have been performed independently from the aforementioned debates (St. John, 1975; Weinberg, 1970).

One of the most frequently encountered names in the literature is Armor (1972, 1973). In his original articles and in his reply to Pettigrew et al (1973), Armor seriously questioned the effectiveness of busing and other strategies as techniques for improving minority achievement. He presented evidence that busing in six different locations had not had a positive effect on minority student performance. Much of Armor's review of the literature concentrated on a comparison of the academic gains of bused and non-bused siblings. Armor concluded that there were not significant differences between academic gains between the bused group and the non-bused brothers and sisters.

Armor's conclusion was quickly challenged by Pettigrew et al (1973). Pettigrew contended that Armor's review of the literature was incomplete, biased, misleading, and that it lacked a thorough analysis of the studies selected. He singled out a study conducted by Armor (1972) in Boston as being unrepresentative and inadequately discussed and noted that a great deal of the evidence used by Armor to make his point came from the study. Pettigrew then went
on to present a list of eight studies which had not been reviewed by Armor that were purported to support the success of various desegregation programs. Pettigrew limited his discussion and analysis to these eight studies even though he had chastised Armor for performing an incomplete and inadequate review of the literature.

Armor (1973) replied to Pettigrew's reply by accusing Pettigrew et al of applying a "double double standard" to the research literature. He said the first double standard related to the burden of proof.

To initiate the action one can use any type of social science data whether or not it directly tests the policy in question and regardless of its technical accuracy. But once the integration policy is in full force, it cannot be questioned unless one can conclusively prove that school integration cannot have an effect on educational benefits. (Armor, 1973, p. 130)

The second double standard related to the appropriateness of scientific evidence in arguing the question of busing.

One willingly applies social science findings to public policy if they are in accordance with one's values but declares them irrelevant if they contradict one's values. (Armor, 1973, p. 130)

In addition to these accusations, Armor responded to Pettigrew's criticism of the METCO research and pointed out that many of the studies cited by Pettigrew were biased and inadequate.

In an attempt to reconcile the opposing "objective" evidence by Armor and Pettigrew, Wilson (1973) presented
a rather jaded observation. His observation can best be summed up in two laws concerning the evaluation of public policy programs.

First law: All policy interventions in social problems produce the intended effect if the research is carried out by those implementing the policy or their friends.

Second law: No policy intervention in social problems produces the intended effect if the research is carried out by independent third parties, especially those skeptical of the policy. (Wilson, 1973, p. 133)

Wilson's point is well taken; inconclusive evidence is open to many interpretations. Without deliberately distorting the results, a set of flawed data can be used to argue virtually any position.

St. John (1970, 1975), another author who has contributed a great deal in this area, offered another explanation for these opposing conclusions. In her earlier review of the literature concerning the effects of desegregation she concluded that the evidence on the effects of busing were marred by serious methodological shortcomings rendering it inconclusive.

In spite of her conclusion in 1970, St. John (1975) recently completed a text which is a comprehensive review of the literature on school desegregation and outcomes of children. It is truly a worthwhile addition to the literature in this area. It is well organized, thorough and objective. St. John's text provided many of the references that were used in this study.
St. John divided her review of research into categories based on the type of design used to accomplish each study she reviewed. Her categories were (a) cross-sectional studies that lack prior measurement of achievement, (b) longitudinal studies that lack a proper control group, and (c) studies with before and after data for both control and experimental groups.

As St. John pointed out, even before and after studies with control groups are quasi experiments rather than true experiments in that the two samples were not randomly drawn or randomly assigned to segregated or desegregated schools.

The most important research using cross-sectional data is the Equality of Educational Opportunity Survey (EEOS) commonly known as the Coleman Report (1966). This study has been widely debated and reanalyzed. The tentative conclusions of this study were that the proportion of whites in a school was positively related to individual performance, but the effect was accounted for more by other characteristics of the student body than the racial composition of the school. One relevant variable seemed to be socioeconomic status (SES). Reanalysis of the Coleman data by McPartland (1968) indicated that desegregation was associated with higher achievement for black pupils only if they were in predominantly white classrooms.
A reanalysis by Cohen, Pettigrew, and Riley (1972) substantiated the Coleman conclusion that racial composition of the school had little effect on the verbal achievement of blacks when school quality and the background of individuals and their peers were controlled.

While St. John concludes that great finesse of analysis went into the original work as well as the reanalysis, there are several major criticisms of the quality of the original data which may tarnish the conclusions as to the influence of racial composition on children. These criticisms can be summarized as follows:

1. There were no measures of prior academic performance thus it is difficult to state that segregated and integrated children were comparable in scholastic aptitude prior to their entry in such schools.

2. The measure used to control for home background did not measure or control for all facets of family influence.

3. School and classroom black-white percentages referred only to one point in time and they may not have been accurately reported.

4. Factors other than racial composition of the classmates were not adequately controlled (i.e. social class).
It is evident that each of these shortcomings could have a major effect on the relationship between achievement and desegregation so at best the conclusions of Coleman, Cohen, Pettigrew and Riley are tentative.

St. John also discussed cross-sectional studies in local school districts. She concentrated on a study in Pittsburgh which she conducted (St. John, 1973). School white percentages were used as the independent variable and Metropolitan Achievement Test scores in arithmetic, reading, language, and science were used as the dependent variables.

The results indicated that desegregation was associated with higher performance in arithmetic but not in other subjects. She conceded that without random assignment, matching, or a pretest, it was impossible to determine the reason for the differences in the results. She speculated that arithmetic, in contrast to reading, is a school-learned skill and was perhaps better taught in racially mixed schools. Another possible explanation was that segregated and desegregated students were not originally alike in their potential for learning arithmetic.

Due to the inconclusiveness of the previous cross-sectional studies there has been an increase in studies that used before and after measurement of achievement. However, many of these studies lack a genuine control
group. St. John used these studies as her second major group for review and discussion: longitudinal studies. She included fourteen such studies in her review. Most of these studies were done at the elementary level. The duration of the study was one year in five studies and two to five years in the others. Busing of black children was the most common method to achieve integration in these fourteen studies.

Only a few of the studies reported show clear gains for desegregated minority group children in comparison with their own previous growth in a segregated setting. The most clear-cut positive findings were those in Hoke County and Goldsboro, North Carolina, where blacks showed post-desegregation gains relative to national norms. In her summary of these studies St. John concluded:

As it stands neither the Riverside evidence nor that of the other longitudinal studies provides strong support for the hypothesis that desegregated schooling benefits minority group children. (St. John, 1975, p. 26)

The third group of studies she reviewed were considered to control for more of the threats to internal validity. Although these studies all have control groups, they also introduced the research problem of self-selection, since those who elected to be desegregated were likely to be different from those who did not elect to be desegregated. To overcome this bias, various attempts were made to match
subjects in the control group on the basis of SES, family background, and earlier achievement. Attempts in Boston and Hartford to handle the choice of a suitable control group were particularly noteworthy.

As mentioned earlier in this report, Armor (1972) formed a control group from the siblings of pupils bused to the suburbs. By this means he was able to insure a measure of comparability between experimental and control groups. In his study significant gains were made by students in the lower grades (K to 3) who were bused, while greater gains in achievement were made by those who remained in the city for grades 4 to 6.

St. John stated several conclusions about the effects of desegregation on academic achievement in her review of the literature:

1. There is some indication that younger children, especially those of kindergarten age, tend to benefit more than older children from integration.

2. The length of exposure to integration has not proved to be an important variable.

3. Gains in mathematics are reported more often than gains in verbal achievement.

4. Within school desegregation needs to be better measured and controlled if we are to find the "why" with regard to gains or no gains.

5. There has not been adequate research as to the method of achieving integration. In only one city, Evanston (Hsia, 1971), were the achievement gains of busers and walkers compared. In this case the busers gained more.
Researchers have not controlled for such variables as the level of community controversy over desegregation, the friendliness of white parents and students, flexibility or prejudice of the staff, the content of the curriculum, or the method of teaching.

Another major contribution to the desegregation literature was made by Weinberg (1970, 1968). His overall question of concern was "How has racial desegregation affected academic achievement?" (Weinberg, 1970, p. 31)

In his book Weinberg (1970) considered four major groups of studies. These were (a) studies which reported on academic achievement under racially segregated conditions, (b) studies in which the situations were bi-racial but which exhibited no special concern for stimulating desegregation, (c) studies which incorporated some type of control group in their design, and (d) studies which specifically dealt with busing as a method of integration.

First, let us examine those studies that were concerned with academic achievement under segregated conditions. Much of the research in this group is of historical interest for much of it was published around the time of the 1954 Supreme Court decision. It is also useful as a general baseline for black achievement at that time and now almost 25 years later.

In 1954 the Texas Association of School Administrators surveyed the academic achievement of eighty percent of the state's school children. It was reported that white
children were achieving very well when compared to national norms, however, black children were performing unsatisfactorily (Weinberg, 1970). Other surveys in the south showed that black children in Dade County, Florida, were two years behind national norms while white children were slightly ahead of these same norms (Weinberg, 1970). In Nashville black sixth graders were more than two years behind white students in overall achievement.

Even in the north, the situation was not satisfactory. Half of all students in New York City in 1959 were reading more than two years below grade level. A majority of these were blacks and Puerto Ricans. Landers (1967) reported that in 1966 "the typical central Harlem student in grade 5 was retarded one year and one month."

Other surveys dealt with differences in IQ scores between blacks and whites. In general, the mean of black students was one standard deviation below the mean of white students. Kennedy (1966) reported that in 1960 black children in the southeastern states had a mean IQ of 80.7 contrasted with an overall IQ mean of 100. Others (Deutsh and Brown, 1964; Scott, 1966; Schreiher, 1962) reported that as the black child progressed through the educational system his IQ declined rather than increased. This indicated to some that the educational system harmed rather than helped the intellectual development of black children.
although the cause of this phenomenon had not been established with research. Weinberg's remaining groups of studies used the research on segregated education as a baseline to address the question, "Does attendance at racially mixed schools bring about any changes in IQ and achievement?" (Weinberg, 1970, p. 35)

The first group of studies dealing with desegregation and achievement was concerned with learning in bi-racial situations that occurred by natural circumstances. That is, no special effort was made to stimulate the resulting integration. In Portland, Oregon, in 1960, high school students were divided into black and non-black by grade level and the racial composition (in percent) of the elementary schools they had attended. The achievement of black students was analyzed by comparing the achievement of blacks from schools with a high percentage of whites with the achievement of blacks from schools with a low percentage of whites. The study concluded that black achievement was benefited by attendance at schools with relatively more whites and middle class students.

However, Weinberg pointed out that these data should be treated cautiously since the conclusions were based on teachers' grades and not objective tests. Weinberg concluded that social class appeared to be a larger determinant of the differences than race.
As indicated earlier in this review of the literature, after five years of desegregation in Washington, D.C., black students performed better than other black students had during the five years preceding desegregation (Weinberg, 1970). In Austin, Texas, the median scores of two ninth grade classes in Austin remained unchanged after one year of desegregation (Morland, 1963). A study in Evansville, Indiana, concluded that the academic level of the student body of the integrated school had not been lowered by desegregation (Wiley, 1961). These surveys of desegregated education are possible indications of the effects of integration but they lack the sophistication and control of pre post designs which utilized control or comparison groups (St. John, 1975). Weinberg (1970) also discussed studies which used control or comparison groups.

Many of the studies cited by Weinberg provided data for this research since they allowed pre-post comparisons on desegregated and segregated groups.

Elliott and Badal (1965) attempted to determine the effects of racial composition on achievement if scholastic aptitude, as measured by IQ, is controlled. Their study was conducted in 1962 in Oakland, California, with 4,693 fifth graders. All in all, Elliott and Badal concluded that racial composition of the class had no important influence on achievement when scholastic aptitude was

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
controlled. Weinberg contended that this research did
not test desegregation effects, since the research con­
cerned changes in school atmosphere, but not in the
performance of individual children.

Stalling (1959) studied academic achievement before
and after desegregation in Louisville. After one year
achievement scores of blacks raised more than did those
of whites.

In 1958 Samuels (1958) conducted a study in New
Albany, Indiana, where he compared the achievement of
black and white children in an integrated junior high and
a segregated junior high. Samuels found that after two
years of desegregation the achievement gap between blacks
and whites narrowed significantly while achievement gap
in the segregated schools was not changed.

Fortenberry (1959) studied black achievement in
Oklahoma City under conditions of segregation and integra­
tion. Comparisons were made over a two-year period.
Findings showed that by eighth grade, students in inte­
grated classes had gained more in arithmetic than in
segregated conditions while there were no differences in
reading achievement. All differences in mathematics were
statistically significant leading Fortenberry to conclude
that "in general Negroes achieve better in mixed than in
non-mixed classes." (Fortenberry, 1954, p. 44)
In a study concerned with changes in IQ scores, as the result of integrated school experiences, Katzenmeyer (1962) followed 193 black and 1,061 white children for two years. The change in pre-post means of the black group of children was found to be statistically significant beyond the 0.001 level when compared to blacks in segregated schools while the mean gain of the white group was not significantly different than gains experienced in segregated schools. Katzenmeyer concluded that the changes were a result of the increased social interaction between black and white children.

Klein (1967) studied the effects of desegregation in South Carolina over an eight-month period. Klein concluded that "the integrated school setting is neither educationally deleterious nor educationally beneficial for Negro students at least over an eight month period."

In a study at the University of South Carolina, Geiger (1968) compared the achievement data after one year of desegregation and reported that there were no significant differences between percentage of blacks in a class and the mean achievement of that class. He arrived at this conclusion by comparing the achievement gains of classes with various percentages of blacks. No differences in achievement were found when the percentage of blacks was used as a dependent variable. He suggested that the fears
of detrimental effects on white students in desegregated schools were not based on fact.

In a study that involved pairing by schools in New York City, Slone (1968) showed that black students in the integrated schools scored significantly higher on arithmetic achievement than did blacks in a segregated school.

Another study that indicated improvement for black students in desegregated environments was done by Williams (1968) in Brevard County, Florida. While the desegregated black students continued to score lower than their white counterparts, in five out of six achievement measures they scored significantly higher than the blacks who remained in segregated high schools.

Another study in a suburb of New York by Denmark (1967) indicated that in the first year of desegregation, black children in grades three to five increased their rate of achievement on verbal ability and narrowed their achievement gap in comparison with their white classmates. Denmark concluded by saying that (a) earlier (gradewise) desegregation is better than later desegregation (gradewise), and (b) integration and compensatory education work best together.

The trend to show significant differences in mathematics achievement and not in verbal and reading scores was repeated in a study by Prichard (1969) in Chapel Hill,
North Carolina. In this study Prichard showed that the math achievement of fifth and seventh grade transfered students was significantly higher than those students who were not transfered after one year of desegregation.

Weinberg's review of studies that used busing as a method of integration showed the same type of conflicting results as other reviewers in this area with regard to the effects of desegregation on academic achievement.

In the East Harlem project, bused children showed "dramatic improvement in their school work and attendance." (Weinberg, p. 82) In contrast, Moorefield (1967) found no achievement effects on the bused group if Kansas City. The list could go on and on but the score card always looks the same. A great many of the studies seem to indicate increases in academic achievement as a result of desegregation, and an equal number of studies seem to indicate no change in academic achievement as a result of desegregation. The debate to determine which group of students is larger is dependent on the individual reviewer who determines which studies to include and which to exclude in their tally.

This notwithstanding, Weinberg concluded:

The evidence is strong that desegregation improves the academic achievement of Negro children. In a few cases, desegregation did not provide such stimulation; and in a rare case or two, Negro children's achievement fell. The evidence is even stronger that white children fail to suffer any
learning disadvantage from desegregation. These positive conclusions are supported in turn by the U. S. Commission on Civil Rights' Racial Isolation report. (Weinberg, 1970, p. 87-88)

Another author who made a significant attempt to resolve the question of academic achievement and desegregation is Irwin Katz (1964). In his article he pointed out that there were several problems with the data about black performance in desegregated schools. The problems were:

1. Many schools did not have separate data for blacks and whites because of policies of racial non-classification.

2. If total legal desegregation occurred it is often accompanied by major vigorous efforts to raise educational standards. Thus the effects of desegregation might be masked.

3. In many studies only a small number of highly selected black pupils were admitted to previously all-white schools. This is particularly true in Southern states. This selection process makes any results suspect.

4. Also, many studies did not permit before-after comparisons making it difficult to assess actual performance. (Katz, 1964, p. 383-384)

With these reservations in mind, Katz agrees with Weinberg, stating that the published data as presented gave a favorable picture of black academic achievement in racially mixed settings.
Summary

This concludes the review of previous literature and research for the purposes of this study. A summary of these studies on desegregation follows.

First, combining results of studies does not produce unequivocal evidence that desegregation, in any form, produces changes in academic performance for blacks. At the same time, one can say that there is little evidence to show detrimental effects on white pupils in desegregated settings.

Secondly, we note that the cited studies came about because of some need to integrate the school systems and not because the experimenter wanted to expand the body of knowledge in this area. Therefore, most of the designs encountered can be considered weak designs because they control for few of the threats to validity (Campbell and Stanley, 1963). Randomization was never used to create control or experimental groups and the method used to form the control groups was suspect in many studies.

There is some evidence that black students benefit more with regard to mathematics achievement scores although causal relationships were not established.

Finally, since there are still various scholarly opinions about the effects of desegregation on academic achievement, more research needs to be done using "tighter"
designs and better methodologies to help clarify what the effects actually are. This research is an effort to assess the effects of busing on academic achievement by using meta analysis techniques. Therefore, this study should contribute to the body of knowledge on the effects of desegregation in general on academic achievement.
CHAPTER III

DESIGN AND METHODOLOGY

Hypotheses to be Tested

The research hypotheses to be tested in the study are all concerned with the effects of desegregation on achievement. The type and quality of design used to accomplish the studies and the probability of positive outcomes are also examined. In other words it was assumed that as the quality of the design increased, the likelihood of positive results decreased. Meta analysis techniques are used to investigate this assumption.

The effects of desegregation on achievement and the effect of the quality of the design on reported outcomes will be assessed by the use of "effect size" (ES). According to Glass (1977) "effect size" is the difference between the means of the experimental and control groups, measured in standard deviations of difference between the experimental and control groups. In his study Glass (1976) concluded that the mean of the experimental group was 0.68 standard deviations above the mean of the control group. Effect size is considered to be a measure of treatment effectiveness. Treatments which produce large differences between experimental and control groups will produce
larger "effect sizes" than those which do not produce such marked differences.

Effect size can be calculated by taking the difference of the experimental group mean and the control group mean divided by the standard deviation of the control group. The formula can be shown as follows:

\[ ES = \frac{\bar{X}_E - \bar{X}_C}{S_X} \]

where:
- \( \bar{X}_E \) = mean of experimental group
- \( \bar{X}_C \) = mean of control group
- \( S_X \) = standard deviation of the control group
- \( ES \) = effect size

The following are the research hypotheses for this study. All of these hypotheses were derived from previous research and/or from the review of the literature by Weinberg (1970) and St. John (1975) and are based on the purposes of the study as stated in Chapter I. These hypotheses respond to those stated purposes in an effort to draw conclusions as to the effects of desegregation on academic achievement.

The first hypothesis can be considered the overall null hypothesis to be tested by the meta analysis technique. It is stated as follows:
H1: There is no difference in the academic achievement of the experimental group (desegregated) and the control group (segregated).

The second hypothesis is:

H2: There is a greater mean effect size for desegregated students at the elementary level than those at the secondary level.

It appears to be a direct contradiction of the first hypothesis which indicates no significant differences between groups. However, studies concerning desegregation with elementary (K-6) children have shown positive results more often than those involving secondary (7-12) children. Studies which showed significant achievement gains for the experimental group include those by Anderson (1966), Danahy (1971), Mahan (1968), Samuels (1971), Wolman (1964), and Wood (1969).

An effect like this should not be too alarming since educators have known for some time that it is much harder to make up achievement deficiencies after the seventh grade. Schramm's (1960) analysis of instructional television showed that gains were made in 33 percent of the studies in grades 3-9 and in only 13 percent of the studies concerned with instructional television at the high school level.
In addition, St. John (1975) concluded that her review of the literature "indicates that younger children, especially those of kindergarten age, tend to benefit more than older children from desegregation." (St. John, 1975, p. 37) This also should not be too alarming, for if integration is a positive influential force it should have greater impact before students have experienced many years of negative forces. It should be easier to affect the learning of children who have experienced fewer years of academic failure in an inferior educational environment.

The third hypothesis is:

H3: Students in desegregated environments will show greater gains in mathematics than verbal skills. This hypothesis is concerned with determining if students in the experimental group gained more in mathematics than students in the control group. These gains will also be compared to gains in reading skills between the two groups.

This hypothesis is indicated by previous research. St. John (1975) stated, "gains in mathematical rather than verbal achievement are reported frequently enough to deserve further study." (St. John, 1975, p. 37) Studies which tend to support this hypothesis are: Prichard (1969), Mayer (1970), Anderson (1966), Funtenberg (1959), Walker (1969).

The fourth hypothesis is:

H4: There is a greater mean effect size for students who have been in desegregated environments for more than one year when compared to those who have had only one year in such an environment.
This hypothesis draws the supposition that if one year of treatment is good, two years are better. This is thought to be true for a number of reasons:

1. Much of education is dependent on previous learning. Therefore, if the previous learning is of a better quality than that provided for the control group, it is likely the subsequent learning will benefit by the longer exposure to "better" education.

2. If desegregation has a positive impact on learning, it should have a greater effect the longer the child is exposed to this beneficial environment (treatment).

3. Also, there is no reason to expect desegregation to produce effects in the negative direction after one year's effects in the positive direction.

4. It is unrealistic to expect one year's exposure to counteract five or more previous years' exposure to inferior education. For example, one cannot expect the effects of 14 years of exposure to an unstable family life to be canceled out by one year's exposure in a stable foster home.

The issue here is treatment level. That is, one should expect larger or greater exposure to a treatment to produce a greater effect until saturation occurs or the law of diminishing return prevails.
St. John (1975) concluded that length of exposure to desegregation has not proved to be an important variable, perhaps because most experiments or studies have been of too short duration for adequate comparisons to be made. This research compared studies of one year's duration to those of more than one year.

The fifth hypothesis is:

H5: The analysis using the voting method (Light and Smith, 1971) will show a positive effect for desegregation that will not be substantiated by the meta analysis of research techniques.

Weinberg (1970) indicated that desegregation studies with positive results outnumbered those with negative results although St. John (1975) indicated that past research does not indicate a trend in either direction. This hypothesis was an attempt to respond to these differences of opinion.

The sixth hypothesis can be stated as follows:

H6: There is greater mean "effect size" shown when a meta analysis is performed on designs without control groups as compared to the designs with control groups. Both designs used a before and after measure of achievement.

This final hypothesis is concerned with "effect size" when the quality of the design used to compute the resulting effect size is taken into account.

It is assumed that designs which control for few sources of variance were likely to indicate greater effects than those designs which were better conceived and as such
controlled for more sources of variance. Glass (1977) indicated that he would expect soft or weak designs to show significant differences more often than designs which controlled for as many sources of rival explanations as possible. If the design used does not rule out all or most of the likely causes of a perceived effect, it is difficult, if not impossible, to attribute the effect solely to the treatment intervention. Weak designs control for fewer of the other likely causes. Therefore, observed effects may well be the result of some factor that was not accounted for in the design rather than the treatment. In Campbell and Stanley's (1963) terms, weak designs control for fewer sources of invalidity than strong designs and thus effects can be attributed to one of these sources of invalidity rather than the treatment. Stronger designs control for more sources of invalidity and are more likely to identify effects that were caused by the treatment.

Variables of Interest

This section contains definitions of the terms that will be used throughout the remainder of this study.

Dependent Variable

In this study academic achievement will be the dependent variable. In most of the studies it was the criterion
used to measure the success of some desegregation plan. Typical examples of the instruments used to measure the variable were: Iowa Test of Basic Skills, Metropolitan Achievement Test, Stanford Achievement Test, and, in some cases, teacher-made tests.

**Independent Variables**

**Desegregation.** Desegregation is the plan and/or decision designed to bring about integration. This is often called planned desegregation, forced desegregation, or induced desegregation. It can come about by busing, magnet schools, voluntary transfer, and any other technique used to bring about integrated education.

The review of the literature for this study found that the overwhelming method used to achieve integration was busing. Of the studies used in this research, 81 percent used busing as the method to achieve integration. The remaining studies used various voluntary transfer plans, redistricting plans, closing of segregated schools and open enrollment schemes.

**Population**

The population of studies for the purpose of this research includes studies on achievement and desegregation done during the years 1955-1977. However, it soon became
apparent that a further delineation of the population was needed. Therefore, Glass's (1977) criteria for sorting were adopted. Only studies with the measure of achievement before and after desegregation were considered for the population. This process eliminated many studies that were descriptive, correlational or cross-sectional in nature.

The problems encountered were not unlike those outlined by Glass (1977) in his manuscript. The first problem encountered was the access-availability problem. This problem is concerned with the fact that many of the studies in this topical area are not available through conventional library sources although they are often cited in other author's bibliographies. This problem will be discussed in greater detail at the conclusion of this section.

Secondly, many of the studies that were located in conventional libraries suffered from lack of standardized reporting methods. Because the would-be meta researcher needs certain statistics to accomplish his analysis, studies which do not include the pertinent statistics must be excluded. For example, if a study with statistically non-significant results does not report the exact "t" or "F" value obtained, the meta researcher will not be able to include this study in his population. Therefore studies which would otherwise meet all other sorting
criteria must be excluded simply because the author or editor chose not to report the necessary statistical value.

A third problem that has to be addressed when the meta researcher decides to carry out such a study is the sample-population distinction. This distinction is concerned with the question of whether a set of located studies on a topic ought to be considered a sample or a population. This is a difficult decision to make, but it is desirable to locate as many of the existing studies on the topic as possible. "Since there is no way of ascertaining whether the set of located studies is representative of the full set of existing studies on the topic, the best protection against an unrepresentative set is to locate as many of the existing studies as possible."

(Jackson, 1978, p. 7) If the number of studies located is greater than the number that can be carefully reviewed, the meta researcher can always devise a sampling plan after all the relevant studies have been located.

The population sample decision has a bearing on whether tests of statistical significance are appropriate, and the number of cases which are needed to use various statistical tests. Glass (1977) initially suggested that the located studies be considered a population, but subsequently he treated them as samples. This is true for at least four reasons:
1. Even a thorough search is quite likely to miss some of the relevant studies.

2. If a search did locate all of the completed and published studies, these studies should only be considered a sample of the phenomena being studies or a sample of all possible studies on the topic.

3. The analysis of an interpretive review is usually intended to make inferences about the phenomena investigated in the individual studies rather than about the particular studies in the phenomena.

4. Finally as Blalock (1972, p. 238-239) points out, even if one thinks he has a population or universe he might want to determine if it is likely that the results could have occurred by chance. The argument is not concerned with the question of generalizing to a larger population. The argument revolves around the processes that could have generated differences between groups that are not related to the variables being examined.

Because of these four reasons the population (universe) of published studies located for this meta analysis will be treated as a sample of the studies on the phenomenon although the author feels that the published studies located represent the universe of published studies on the topic.

The access-availability problem was particularly problematic for meta analysis on desegregation because
a large portion of the research studies on desegregation are evaluation studies done by research and evaluation departments of local education associations in order to satisfy the courts, the federal agencies or their own interests. Because of the reluctance of many school districts to share their reports and the time and difficulty that would be required to accumulate these reports, evaluation studies done by school districts or local education associations were eliminated from the population of studies used in this research. Because of the problems of access and availability, it was decided that the population of interest would be concerned with those studies that could be gathered from conventional library sources such as ERIC, dissertation abstracts, journal articles and texts on the subject. Therefore, the population of interest includes those studies that were published in 1955-1977 and thus were available in modern library and document retrieval systems.

A systematic review of the research literature on the effects of desegregation was completed in August of 1977. The following five sources were searched for appropriate journal articles, books, dissertations and other unpublished materials, such as ERIC documents.

1. Library of Congress Card Catalogue at Western Michigan University, University of Michigan, Michigan State University, and Notre Dame University.
Relevant studies were located in this source by using the card catalogue at these various sites under subject headings of: School Integration, Segregation in Education, Busing, Desegregation, and Race Relations.


Appropriate dissertations were located by using the subject index under the same headings as used in the search using the card catalogue.


Articles were located by using the subject index under the subject headings: Desegregation, Public Schools-Desegregation.


Related articles and studies were located by using the subject index under the following subject headings:

(a) Desegregation
(b) Segregation in Education
(c) Busing
(d) Desegregation in Education
(e) Education
(f) Negroes in the United States-Education
(g) Public Schools-Desegregation


This was accomplished by a computer search of all ERIC documents. The relevant independent variables for
this research were: desegregation, busing, integration, school desegregation, school integration, integration methods. The dependent variables of concern were: academic achievement, integration effects, program effectiveness, intellectual development, and integration outcomes.

In addition, two recent books that were found in the card catalogue search proved to be useful in providing leads for other studies for this research. These books were School Desegregation: Outcomes for Children by Nancy St. John, 1975, and Desegregation Research: An Appraisal by Meyer Weinberg, 1970. These texts were a valuable resource for this study. They were used to add studies to the population from journal articles cited in the texts, chapters from books cited in these texts, ERIC documents that were not uncovered in the ERIC search and dissertations which were not discovered in Dissertation Abstracts.

Since both of these books are thorough, comprehensive reviews of previous research and literature in the area of desegregation, it is felt that all studies that met the sorting criteria have been uncovered.

Based on this search process, the studies used for analysis procedures were distributed as follows:
this research were: desegregation, busing, integration, school desegregation, school integration, integration methods. The dependent variables of concern were: academic achievement, integration effects, program effectiveness, intellectual development, and integration outcomes.

In addition, two recent books that were found in the card catalogue search proved to be useful in providing leads for other studies for this research. These books were School Desegregation: Outcomes for Children by Nancy St. John, 1975, and Desegregation Research: An Appraisal by Meyer Weinberg, 1970. These texts were a valuable resource for this study. They were used to add studies to the population from journal articles cited in the texts, chapters from books cited in these texts, ERIC documents that were not uncovered in the ERIC search and dissertations which were not discovered in Dissertation Abstracts.

To insure that all the published studies on desegregation and academic achievement were located, a list of the identified studies was sent to six people who are recognized experts in the area of desegregation (see Appendix C and D). Three of the six experts replied. They generally agreed that they could not make any additions to the submitted list of studies.
Based on this search process, the studies used for analysis procedures were distributed as follows:

<table>
<thead>
<tr>
<th>Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dissertations</td>
<td>32</td>
</tr>
<tr>
<td>ERIC Documents</td>
<td>13</td>
</tr>
<tr>
<td>Journal Articles</td>
<td>6</td>
</tr>
<tr>
<td>Chapters of Books</td>
<td>4</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>55</strong></td>
</tr>
</tbody>
</table>

The following sorting criteria were used to determine which studies would be included in the analysis portion of this research:

1. Studies were not included unless they were of a before and after research nature. That is, for the purposes of meta analysis, it is necessary to have a measure of achievement prior to integration and a post-measure of achievement after integration in order to determine an effect size.

2. Studies were not included unless they utilized some quantitative measure of achievement for their dependent variable.

3. Studies were excluded if they did not reveal the number of subjects, the alpha level, and the test of significance or pre-post means and standard deviations. Without these data it is impossible to perform a meta analysis.

4. Studies which consisted exclusively of attitudinal data, teacher grades, and opinions were not included in the study population.
5. Cross-sectional studies were excluded because it was felt that studies of this type are correlational in nature and thus would require a separate analysis. They were also excluded because they were often racial composition studies and not studies based on desegregation as a treatment.

6. Studies which used I.Q. as the dependent variable were not included since I.Q. was considered to be a different concept than achievement.

These sorting criteria reduced the number of studies for the population as follows:

<table>
<thead>
<tr>
<th></th>
<th>Before Sorting</th>
<th>After Sorting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dissertations</td>
<td>61</td>
<td>32</td>
</tr>
<tr>
<td>ERIC Documents</td>
<td>122</td>
<td>13</td>
</tr>
<tr>
<td>Journal Articles</td>
<td>48</td>
<td>6</td>
</tr>
<tr>
<td>Chapters of Books</td>
<td>29</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>260</td>
<td>55</td>
</tr>
</tbody>
</table>

At this point it would be well to emphasize while there were 55 relevant studies, the total population of studies used in the final analysis is 129 because many of the authors of the 55 studies included separate analyses that could be treated as separate studies by meta analysis techniques. That is, many included smaller studies within their study by grade level or academic subject which could be analyzed separately by the use of meta analysis.
Procedures

The procedures used to accomplish the analysis were those outlined by Glass (1976, 1977) as well as a comparison to the voting method as outlined by Light and Smith (1971).

There were two major types of designs encountered in research in the area of desegregation. The two types require different techniques to arrive at a successful integration of the findings. The two major types were:

(a) The four-celled Pre-Post design with a control group (see Figure 2 below).

(b) The two-celled Pre-Post design without a control group. This design is identical to Figure 2 except the control blocks are eliminated. (Blocks 3 and 4)

The first type constitutes the strongest design encountered in research in this particular area. However, it has weaknesses, since randomization is not employed. According to Campbell and Stanley (1963) true experimental designs must employ randomization. The four-celled Pre-Post design with a control group can be diagrammed as follows:

Figure 2

A Four-Celled Pre-Post Research Design

<table>
<thead>
<tr>
<th></th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental Group</td>
<td>Segregated (1)</td>
<td>Desegregated (2)</td>
</tr>
<tr>
<td>Control Group</td>
<td>(3) Segregated</td>
<td>(4) Segregated</td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 3 is included to illustrate the longitudinal nature of the design.

**Figure 3**

**Example of Longitudinal Pre-Post Research Design**

<table>
<thead>
<tr>
<th></th>
<th>(Before)</th>
<th>(After)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre</td>
<td>Post</td>
</tr>
<tr>
<td><strong>Experimental:</strong></td>
<td>0</td>
<td>X</td>
</tr>
<tr>
<td><strong>Control:</strong></td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

By using this diagram one should keep in mind that a specific amount of time has elapsed between the pre and post testing periods, normally one academic year. The "X" indicates that some form of desegregation was instituted after the pre test. The "0"'s indicate the testing that was performed before and after desegregation. The control group was not exposed to such a process.

Findings of these studies can be aggregated as follows:

1. The common measure of treatment (desegregation) effectiveness will be defined as "effect size."

2. Effect size (ES) is the difference between the means of the desegregated group and the control group divided by the control group's standard deviation. The
control group standard deviation is used because it is the standard deviation unaffected by treatment (Glass, 1977, p. 30).

3. The average effect size is calculated by averaging the effect sizes from studies which are part of the group.

As Glass (1977) indicated, the summary statistic for this analysis will indicate that "average effect size is 'Y' standard deviations above or below that of the average control group member." In his example, the average effect size for the experimental group was 0.68 indicating:

that the average person receiving some form of psychotherapy was about two-thirds standard deviation more improved than the average control group member. (Glass, 1977, p. 23)

The procedures for calculating an effect size imposes restrictions that are worth noting. These limitations can best be explained by reviewing the formula for effect size. From Glass (1977):

\[ ES = \frac{\bar{X}_{Exp} - \bar{X}_{Control}}{S_{X_C}} \]

where: \( ES = \) Effect Size

\( \bar{X}_{Exp} = \) Mean of Experimental Group - POST TEST

\( \bar{X}_{Control} = \) Mean of Control Group - POST TEST

\( S_{X_C} = \) Standard Deviation Control Group - POST TEST

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
In order for the researcher to compute effect size for four-celled designs using the above formula, the standard deviation of the control group (post test) is required.

While it may seem self-evident that post test means for both control and experimental groups are required, the meta researcher may discover that he is furnished with mean gain scores rather than post test means and that the standard deviation of the gain scores is not included. Original paired data are almost never included so it makes it impossible to work backwards to determine the required means and standard deviation. If the data cannot be retrieved then the meta researcher must use one of the other methods to compute an effect size. These methods will be discussed under Effect Size and Significance Level in Chapter V.

The next group of designs considered were those designs which were longitudinal in nature but did not employ a control group for comparisons. These designs can be pictured by referring to the four-celled diagram and imagining that cells (3) and (4) are empty since there was no control group:

<table>
<thead>
<tr>
<th></th>
<th>Pre</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental:</td>
<td>0</td>
<td>X</td>
</tr>
</tbody>
</table>
"X" is the treatment (desegregated education) and the "0's" are the pre and post test measures of achievement.

The problems of attempting to analyze designs of this type were recently made manageable by Glass (1977) in his manuscript, *Integrating Findings: The Meta Analysis of Research*. According to Glass these types of studies can be analyzed as follows:

1. **Treatment effectiveness** is the mean difference between pre and post measures.

2. The pre measure is to be used as the simulated control group mean and the post measure is the experimental group mean.

3. "Effect Size" is defined as the difference in means divided by the standard deviation of the pre measure, since this is considered to be the control group.

Effect size can be calculated as follows:

\[
ES = \frac{X_{\text{pre}} - X_{\text{post}}}{S_X (\text{pre})}
\]

**Effect Size Based on Significance Level**

A commonly encountered problem results because of the way research reports are published. Reports often give sample sizes and an indication of whether a mean difference was statistically significant at a certain alpha level. As per Glass's (1977) suggestion, data of this character can be handled as follows for meta analysis purposes:
1. Suppose mean of group 1 (experimental) exceeded mean of group 2 (control) and \( N_1 = 120, N_2 = 120 \).

2. It was significant at .05 level of significance by use of a two tailed test with \( N_1 + N_2 - 2 \) degrees of freedom.

3. Therefore, 
   \[
   t = \frac{\bar{X}_E - \bar{X}_C}{\sqrt{S^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} = 1.98
   \]

4. From this one can get a conservative estimate of effect size by:
   \[
   ES = \frac{\bar{X}_E - \bar{X}_C}{S_X} = 1.98 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}
   \]

5. Thus given only a \( t \) value and an \( n_1 \) and \( n_2 \) one can get effect size as:
   \[
   ES = t \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}
   \]

One of the limitations with this method is that a meta researcher might expect every study to include the \( n \)'s for both groups, the alpha level and the actual \( t \) value. In fact, many studies only state that:

1. The mean of the experimental group exceeded the mean of the control group and it was statistically significant at the .05 level; OR

2. The mean of the experimental group did **not** exceed the mean of the control group and it was **not** statistically significant at the .05 level.

In the first case the method previously outlined can be used to give a conservative estimate of effect size.
because an appropriate t (based on sample sizes) can be inserted in the formula to give an estimate of effect size. Therefore, even if one does not know the exact t value, an estimate of effect size can be determined.

However, when the results are not statistically significant and the exact (actual) t value is not given, this method can not be used to arrive at an estimate of effect size. Obviously, if the results are not significant but the actual t value is included, this method can be used to determine an estimate of effect size.

The fact that statistically significant studies could be included in the study even if the exact t values for those studies were not published, while studies that were not statistically significant could only be included if the actual t value was stated led to a source of bias that was not anticipated prior to the start of this study.

This bias came about because many of the studies with statistically not significant results did not publish the exact obtained t value or the statistics required to calculate the exact t value. Because of this tendency, statistically not significant studies were excluded at a rate that was much higher than their representation in the total population of studies. In contrast, all the studies that reported statistically significant results, specific alpha levels, and sample sizes were included in the final group of studies.
For example, out of the total 129 studies that made up the population for this study, 68 were statistically not significant studies. Therefore, statistically not significant studies make up 53 percent of the initial population. However, over half (35) of these studies could not be included in the final calculations of effect size because the publisher did not state the actual obtained t value.

In contrast, to statistically not significant studies, the original population of 129 studies included 61 (47 percent) statistically significant studies but only 12 (20 percent) of these studies had to be eliminated from the final calculations of effect size because the publisher had omitted some required statistical information.

It is evident that the requirements for calculating an effect size using significance levels introduced a bias by excluding not statistically significant studies at a higher rate than was the case for the statistically significant studies. In fact, while statistically not significant studies initially made up 53 percent of the total population, they made up only 40 percent of the final sample due to the bias introduced by using significance levels for calculating an effect size. Conversely, while statistically significant studies made up 47 percent of the total population, only 20 percent of these studies
had to be eliminated; therefore, they made up 60 percent of the final sample.

Since it was evident that the exclusion of over half of the statistically not significant studies would artificially inflate the resulting mean effect size(s), several strategies were developed to minimize the effects of this bias problem. These strategies included the following:

1. An attempt was made to contact the authors of non-significant studies. Of the 35 original studies only 16 could be located. Efforts to locate authors included the use of alumni associations and use of various professional and faculty directories. Of these 16, 8 responded but they did not provide the statistical information needed to include their study in the calculations of effect size. Most of the authors who responded sent a copy of the published article or said they could not take the time to locate the original data. A copy of the letter sent to these authors and a list of these authors is included in Appendix C and D respectively.

2. It was decided to randomly remove 11 of the statistically significant studies so that the proportion of statistically significant studies in the final sample would be approximately equal to the proportion of significant studies in the population. This was accomplished by
coding each of the studies with a number and removing them by use of a random number generator in the computer which utilizes a table of random numbers and a five-digit value based on time to initiate the random number generating process. After this was done, effect size(s) for each hypothesis were re-calculated using the remaining studies.

It is recognized that these strategies which were used to compensate for a problem inherent in the meta analysis technique of using significance levels to calculate effect sizes did not totally solve the original problem. However, this study did identify an important problem for future meta analyses in other topical areas. An extensive effort was made to insure that this bias problem would not affect the final conclusions of this study or the conclusions drawn by readers of this study.

Data Analysis Procedures and Steps

This section deals with the specific procedures and methods required to address the hypotheses of interest and to bring some logical structure to this phase of the research.

Hypotheses that were tested for statistical significance were tested by means of a t test using independent
samples or a t test for dependent groups when the data was clearly not based on independent groups. The description of the independent t test is as follows:

The hypothesis tested is that the difference between the means of two populations, \( u_1 - u_2 \), is equal to zero, against the alternative hypothesis that it is different from zero:

\[
H_0: \quad u_1 - u_2 = 0 \\
H_1: \quad u_1 - u_2 \neq 0
\]

\( H_0 \) is tested against \( H_1 \) by means of the following test statistic:

\[
t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2 \over n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}
\]

where \( \bar{X}_1 \) and \( \bar{X}_2 \) are the means of the samples from populations 1 and 2 respectively, \( S_1 \) and \( S_2 \) are unbiased estimates from samples 1 and 2 of the common population variance and \( n_1 \) and \( n_2 \) are the sizes of samples 1 and 2. (Glass and Stanley, 1970, p. 295)

The t statistic for testing \( u_1 - u_2 = 0 \) with dependent groups is as follows:

\[
t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1^2 + S_2^2 - 2 r_{12} \frac{S_1 S_2}{n}}}
\]

where \( \bar{X}_1 \) and \( \bar{X}_2 \) are the means of the samples from the populations 1 and 2 respectively, \( S_1 \) and \( S_2 \) are unbiased estimates from samples 1 and 2 of the common population variance, \( r_{12} \) is the correlation between \( \bar{X}_1 \) and \( \bar{X}_2 \) and \( n \) is the size of groups. Since the data is based on dependent samples, it is in the form of \( n \) pairs of observations. (Glass and Stanley, 1970, p. 300)
The procedures and methods were as follows:

1. First, effect sizes for all studies were calculated using formulae on pages 89-90. This step was required to address hypothesis H1 concerning the overall effect of desegregation utilizing all studies in the population. The mean effect size was calculated by dividing the sum of the effect sizes for all studies by the total number of studies.

2. The next step was required to address hypothesis H4 concerning the effect size and length of desegregation. This was accomplished by constructing two populations of studies; those of one year or less of integration and those of greater than one year of integration. The mean effect size for each group was calculated separately using the formulas on pages 89-90. The effect sizes were compared and a t test was used to test the statistical significance of the difference.

3. To accomplish the next step separate effect sizes were calculated for achievement in mathematics and achievement in verbal skills. In each case effect size was calculated for within subject area gains as well as comparing the magnitude of the effect size for these two achievement gains. A dependent t test was used to test the statistical significance of gains in math compared to those in verbal skills.
4. With regard to achievement and desegregation, the final step was to separate the studies into two groups; those involving elementary children (grades K-6) and those involving junior high and high school students (grades 7-12). Then separate effect sizes were calculated for each group and compared. A t test was performed to check for statistical significance between the two grade levels of interest. This step was intended to respond to hypothesis H2.

The next portion of the data analysis procedures was related to comparing the results of a meta analysis to the more commonly used voting method procedures.

5. Therefore, the voting method was used to calculate a winner for:

(a) all studies together
(b) control group designs vs. no control group designs
(c) elementary and secondary grade levels
(d) math and verbal skills
(e) one year and greater than one year of segregation

These results were compared and contrasted to the results obtained from a meta analysis using the same dependent variables.

6. The final phase of the data analysis was to compare the effect sizes for control group designs and designs without control groups to determine if the type and quality of design is likely to have an effect on the effect size. This step is an effort to respond to hypothesis H5.
As one can see, in order to answer the hypotheses of interest, the data must be broken into various categories that separate analyses performed on each set or grouping of data. The data are organized as follows:

1. Effect sizes were calculated for all studies by separating the total group of studies into two groups; one group will consist of all the experimental groups from all studies, and the second group of studies will be made up of all the control groups from all the studies.

<table>
<thead>
<tr>
<th>All Experimental Groups</th>
<th>Pre</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Segregated</td>
<td>Desegregated</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>All Control Groups</th>
<th>Pre</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Segregated</td>
<td>Segregated</td>
</tr>
</tbody>
</table>

2. The studies were then separated into those involving one year of desegregation or less and a group of studies involving greater than one year of desegregation.

<table>
<thead>
<tr>
<th>One Year</th>
<th>Pre</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental</td>
<td>Segregated</td>
<td>Desegregated</td>
</tr>
<tr>
<td>Control</td>
<td>Segregated</td>
<td>Segregated</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Greater Than One Year</th>
<th>Pre</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental</td>
<td>Segregated</td>
<td>Desegregated</td>
</tr>
<tr>
<td>Control</td>
<td>Segregated</td>
<td>Segregated</td>
</tr>
</tbody>
</table>

Mean effect sizes for each group were calculated and compared by use of a test of statistical significance.
3. Studies were separated into those involving mathematics achievement and those involving verbal skills.

<table>
<thead>
<tr>
<th>Mathematics</th>
<th>Pre</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental</td>
<td>Segregated</td>
<td>Desegregated</td>
</tr>
<tr>
<td>Control</td>
<td>Segregated</td>
<td>Segregated</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Verbal Skills</th>
<th>Pre</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental</td>
<td>Segregated</td>
<td>Desegregated</td>
</tr>
<tr>
<td>Control</td>
<td>Segregated</td>
<td>Segregated</td>
</tr>
</tbody>
</table>

The effect sizes for each group were calculated and compared by use of a test of statistical significance.

4. Studies were separated into those involving desegregation grades K through 6 and those involving desegregation grades 7 through 12.

<table>
<thead>
<tr>
<th>Grades K-6</th>
<th>Pre</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental</td>
<td>Segregated</td>
<td>Desegregated</td>
</tr>
<tr>
<td>Control</td>
<td>Segregated</td>
<td>Segregated</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grades 7-12</th>
<th>Pre</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental</td>
<td>Segregated</td>
<td>Desegregated</td>
</tr>
<tr>
<td>Control</td>
<td>Segregated</td>
<td>Segregated</td>
</tr>
</tbody>
</table>

Effect sizes for each group were calculated and compared by use of a test of statistical significance.
5. Studies were separated into those using control group designs and into those employing designs without control groups.

<table>
<thead>
<tr>
<th>Control Group Designs</th>
<th>Pre</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental</td>
<td>Segregated</td>
<td>Desegregated</td>
</tr>
<tr>
<td>Control</td>
<td>Segregated</td>
<td>Segregated</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Designs Without Control Groups</th>
<th>Pre</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental</td>
<td>Segregated</td>
<td>Desegregated</td>
</tr>
</tbody>
</table>

Mean effect sizes for each group were calculated and compared by use of a test of statistical significance.

Summary

This proposal has shown that the previous literature and attempts to arrive at a conclusion concerning the effects of desegregation have been largely ineffective. As Light and Smith (1971) indicated, these have been efforts at arriving at a conclusion by reading a group of conclusions. The researcher using this method reviews a group of conclusions from existing research studies and then arrives at a summary conclusion based on a synthesis from the group of conclusions. This process has been unscientific and often distorted by the value perspective of the person reaching the conclusion. This has been particularly
true in a controversial area such as desegregation in education. It is assumed that meta analysis techniques can bring some scientific, rational thinking to bear on this subject.

It showed that previous attempts to extract knowledge from existing research have been inadequate and ineffective due to the nature and quality of research data available.

The study employed a technique known as meta analysis of research as outlined by Glass (1976, 1977) to test the results for statistical significance. Meta analysis is a rigorous alternative to the casual, narrative discussions of research studies which typify attempts to make sense of the research literature.

The design and procedures are outlined so that the reader can begin to understand the problems that are encountered when one attempts to do a meta analysis in any topical area. Most, if not all, procedures used in this research design were drawn from Glass's (1976, 1977) two papers in meta analysis.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
CHAPTER IV

ANALYSIS OF THE DATA (RESULTS)

Characteristics of the Data

This chapter contains an analysis of research evidence on the effects of desegregation on academic achievement.

Discussion in this chapter will focus on the results of the data analyses discussed in the preceding section, as well as unanticipated problems which were encountered in analyzing the data. It is assumed that problems in analyzing data from published research reports are not unique to desegregation data and are likely to affect meta analysis research in other topical areas.

Data Analysis: The Sample

After the appropriate sorting criteria were applied to the studies gathered for consideration, the sample consisted of 129 discrete studies which were drawn from 55 larger studies. This means there are 55 authors or citations that are listed in the bibliography but that most of the 55 studies referenced had data within their study that could be treated as a separate study. For example, if the larger study included data for different grade levels, each grade level was treated as a separate study;
or if the author(s) used math and verbal achievement gains as their dependent variable, these were treated as separate studies. After the studies for analysis were accumulated, there were 129 studies for meta analysis purposes. It was evident effect sizes (ES) would have to be calculated by several different methods since most of the studies did not include the standard deviation and several did not make use of a control or comparison group. Therefore, it was decided to calculate effect sizes using three distinct methods.

1. Effect size was calculated by subtracting the mean of the control group from the mean of the experimental group and dividing by the standard deviation of the control group. This method was used if a control group was used and the appropriate means and standard deviations were furnished by the author.

2. Effect size was calculated for studies without control groups by subtracting the pre mean from the post mean and dividing by the pre standard deviation. This method was used if there was no control group and the appropriate means and standard deviations were furnished by the author.

3. Finally, effect size was calculated by converting stated significance levels into effect size if the alpha
level used, t or F value and sizes of groups were given. This method could not be used for studies reporting not statistically significant differences unless the author stated the exact obtained t or F value.

In addition to the preceding methods, it was necessary to introduce several precautions because of the bias due to excluding studies reporting non-significant results but not reporting alpha levels. This "statistically significant bias" problem was discussed in Chapter III.

For each hypothesis a table will be furnished that indicates the effect size for the group of studies being analyzed after 11 statistically significant studies were removed by a random process. This step was taken as a method of minimizing the bias introduced by the previously discussed "statistically significant bias problem."

Overall Effects of Desegregation on Achievement

The first effect sizes to be calculated were for the total population of studies including all designs, all subjects, and all lengths of desegregation. Effect sizes were calculated by using one of the three methods as outlined by Glass (1977, 1976) and previously described (page 89). The results are shown in Table 2.
Table 2
Effect Size for All Studies

<table>
<thead>
<tr>
<th>Total Number of Studies</th>
<th>Minimum Effect Size</th>
<th>Maximum Effect Size</th>
<th>Mean Effect Size</th>
<th>Standard Deviation of Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>71</td>
<td>-1.17</td>
<td>1.13</td>
<td>.16</td>
<td>.41</td>
</tr>
</tbody>
</table>

The distribution of these data is fairly normal although it does have several discontinuities. For a graphic representation of these data see Figure 4. One can see by reviewing these graphs that the data are oriented toward positive effect sizes since only ten of the seventy-one effect sizes are negative values.

By reviewing Table 2 one can see that the experimental group benefited by 0.16 standard deviations. That is, the mean of the experimental group exceeded the mean of the control group by 0.16 standard deviations. One should also note that the effect sizes for the studies used to make up this population of 71 studies had a range that varied from -1.17 standard deviations below the mean of the control group to +1.13 standard deviations above the mean of the control group.

Effect Size on Achievement When Grade Level is Considered

In order to respond to hypothesis H2 and to test the assumption, in the literature, that students in lower
Figure 4

Distribution of Effect Sizes

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Elementary grades gain more than students in junior high and high school, the data were clustered into two groups by the grade levels of the students involved. One group consisted of studies involving students in grades K through 6 and the second group consisted of studies involving students in grades 7 through 12. Hypothesis H2 stated:

There is a greater effect size for desegregated students at the elementary level than those at the secondary level.

Table 3 is a summary of the results for studies involving lower grade levels (K-6) and for studies involving upper grade levels (7-12).

Table 3
Effect Sizes for Studies Involving Students in Grades K-6 and Grades 7-12

<table>
<thead>
<tr>
<th>Grade Levels</th>
<th>Total Number of Studies</th>
<th>Minimum Effect Size</th>
<th>Maximum Effect Size</th>
<th>Mean Effect Size</th>
<th>Standard Deviation of Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-6</td>
<td>51</td>
<td>-0.96</td>
<td>1.13</td>
<td>0.22</td>
<td>0.34</td>
</tr>
<tr>
<td>7-12</td>
<td>20</td>
<td>-1.17</td>
<td>0.59</td>
<td>0.01</td>
<td>0.53</td>
</tr>
</tbody>
</table>

The mean effect size for students in grades K-6 is quite similar to the mean effect size calculated for all studies, however the effect size for students in grades 7-12 seems to be considerably smaller than for all studies and for the group of studies for grades K-6.
By comparing the two grade levels the reader can see that the effect size for the lower grades is considerably larger than the mean effect size for grades 7-12. In order to test the hypothesis (H2) a t test was performed to determine if the difference in the means was statistically significant. A t test yielded the following results:

Table 4
Test of Means for Studies Using Students in Grades K-6 Compared to Studies Using Students in Grades 7-12

<table>
<thead>
<tr>
<th>Grade Level</th>
<th>N</th>
<th>Mean Effect Size</th>
<th>Variance</th>
<th>Degrees Freedom</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-6</td>
<td>51</td>
<td>.22</td>
<td>.10</td>
<td>1.93</td>
<td>69</td>
</tr>
<tr>
<td>7-12</td>
<td>20</td>
<td>.01</td>
<td>.29</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The results indicate that the effect size for lower grade levels (K-6) is not statistically significantly greater than the effect size for the upper grade levels (7-12). This finding indicates that for this study sample desegregation does not influence learning more in elementary levels than it does in secondary levels even though the magnitudes of the effect sizes appear to be quite different (0.22, 0.01).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
It is appropriate to point out that the process which was used to remove studies randomly was a simple random process based on the total 71 studies and not a stratified random approach to removing studies by grade level, subject area, length, etc. In this particular group of studies, this process removed ten studies from the K-6 group of studies and only one from the 7-12 group of studies. If a stratified (by grade level) random process had been used, seven studies would have removed from the K-6 group of studies and four from the 7-12 group of studies. This would have produced an N of 54 for K-6 group and an N of 17 for the 7-12 group of studies. Certainly, this could have affected the results, but it would be difficult to conclude to what extent.

Effect Size of Desegregation for Math and Verbal Achievement Gains

Another contention in the literature is that students in desegregated settings gain more in mathematics achievement than they do in reading achievement. Because of this assumption it was decided to test this contention by means of hypothesis H3 which stated:

H3: Students in desegregated environments will show greater gains in mathematics than in reading or verbal skills.

To test this hypothesis the studies were clustered into a population of studies concerned with mathematics
skills and a population of studies concerned with reading skill achievement. The process of separating studies into specific groups produced a problem with this group of studies that was not a problem with the other variables used to separate studies into categories for analysis. The problem is that when these studies were separated into a group of studies on verbal achievement and a group of studies on mathematics achievement it was evident that the test of independent means was not appropriate since most of the subjects in the verbal studies were the same as those who were involved in the mathematics studies. For example, of the total of the fifty studies which were appropriate for this analysis, thirty-four studies (seventeen pairs) were composed of the same subjects. Only three of the studies in the mathematics group of studies could not be matched with the same subjects in the verbal category of studies. Since the data were not based on independent groups it was appropriate to use the t statistic for dependent groups to test the difference in means between these two groups of studies (Glass and Stanley, 1970, p. 300). Table 5 is a summary of the effects for these two populations.
Table 5
Effect Size(s) for Studies Which Used Mathematics or Verbal Achievement as a Dependent Variable

<table>
<thead>
<tr>
<th>Achievement Measure</th>
<th>Total Number of Studies</th>
<th>Minimum Effect Size</th>
<th>Maximum Effect Size</th>
<th>Mean Effect Size</th>
<th>Standard Deviation of Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics</td>
<td>17</td>
<td>-.99</td>
<td>.54</td>
<td>.17</td>
<td>.38</td>
</tr>
<tr>
<td>Verbal</td>
<td>17</td>
<td>-1.17</td>
<td>.47</td>
<td>.07</td>
<td>.33</td>
</tr>
</tbody>
</table>

These data seem to indicate that students' gains in mathematics are greater than their gains in verbal skills. The t test of independent means was performed; the results were as follows:

Table 6
t Test of Means for Dependent Samples Mathematics vs Verbal Achievement

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>Mean Effect Size</th>
<th>Variance</th>
<th>Obtained t</th>
<th>Degrees Freedom</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics Achievement(1) (N=17)*</td>
<td>.17</td>
<td>.14</td>
<td>1.90</td>
<td>16</td>
<td>.08</td>
</tr>
<tr>
<td>Verbal Achievement(2) (N=17)*</td>
<td>.07</td>
<td>.11</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*r₁,₂=0.85
As one can see the difference in the means is not significant at the .05 level.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Effects of Length of Desegregation on Achievement

The next hypothesis to be tested concerned achievement gains for those students who were desegregated for one year as compared to those students who were desegregated for more than one year. This hypothesis stated:

H4: There is a greater mean "effect size" for students who have been in desegregated environments for more than one year when compared to those who have had only one year in such an environment.

Table 7 is a summary of the results for students desegregated for only one year compared to the results for students who were in a desegregated environment for more than one year.

Table 7

Effect Size(s) for Studies Involving Students Who Were in a Desegregated Environment for One Year and Greater Than One Year

<table>
<thead>
<tr>
<th>Length of Desegregation</th>
<th>Total Number of Studies</th>
<th>Minimum Effect Size</th>
<th>Maximum Effect Size</th>
<th>Mean Effect Size</th>
<th>Standard Deviation of Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>One Year or Less</td>
<td>37</td>
<td>-1.17</td>
<td>.59</td>
<td>.11</td>
<td>.48</td>
</tr>
<tr>
<td>Greater Than One Year</td>
<td>34</td>
<td>- .40</td>
<td>1.13</td>
<td>.21</td>
<td></td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
To test the significance of the difference in the mean effect sizes, a t test was performed. The results are shown in Table 8.

Table 8
Test of Means for Studies Involving Students Who Were in Desegregated Environments for One Year Compared to Students Who Were in Desegregated Environments for More Than One Year

<table>
<thead>
<tr>
<th>Length of Desegregation</th>
<th>Mean Effect Size</th>
<th>Variance</th>
<th>Degrees Freedom</th>
<th>Obtained t</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>One Year (N=37)</td>
<td>.11</td>
<td>.23</td>
<td>.96</td>
<td>69</td>
<td>.34</td>
</tr>
<tr>
<td>Greater Than One Year</td>
<td>.21</td>
<td>.10</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As one can see by reviewing Table 8, the length of exposure to desegregated education did not have a statistically significant effect on the resulting achievement. The results were not significant at the .05 level of significance, therefore the hypothesis was not substantiated. For this study sample longer exposure to desegregated education did not produce significant achievement gains.
A Comparison of the Voting Method and Meta Analysis Techniques

The fifth hypothesis was designed to test the assumption that if a simple voting method was used to test the benefits of desegregated education that more studies would show positive effects than would be the case if the meta analysis techniques were used. This hypothesis stated:

H5: The analysis using the voting method (Light and Smith, 1971) will show a positive effect for desegregation that will not be substantiated by the meta analysis of research techniques.

Of the 129 studies that became part of the studies for analysis, 61 were statistically significant and 68 were not statistically significant. This means that of the 129 studies located for this study, the "no significant effects" side was 53 percent (68/129) while the "statistically significant effect" side was 47 percent (61/129) of the total vote.

However, if one uses the 71 studies that were appropriate for meta analysis techniques, one would discover that the outcome would shift so that the new "winner" would be in favor of desegregated education. For example, of the 71 studies where an effect size could be calculated 43 studies were statistically significant while 28 studies were not statistically significant. Therefore, out of the total vote of 71 studies, results in favor of desegregation would gain 60 percent (43/71) of the total vote while

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
studies which show not statistically significant results would gain 40 percent (28/71) of the total vote.

This difference arises because of the requirements needed to calculate an effect size. The reason it came about with this particular set of studies is because studies that were not statistically significant and did not publish the exact t value and/or statistics required to calculate the exact t value could not be included while statistically significant studies could be included even if the exact t value was not published.

For purposes of this research it is concluded that the voting method did not produce a different outcome than the meta analysis method. That is, by using the voting technique or the meta analysis technique on the 71 studies used for this research one could conclude that desegregation produced slightly positive results. Hypothesis H5 was not substantiated since the results for the two methods were both slightly in favor of the desegregated group of studies.

A test of statistical significance was not performed on this hypothesis because previous attempts of aggregating studies by the voting method have not used tests of statistical significance. The voting method has been used in aggregating the results of research studies much as it has
been used in popular elections. A winner is required to receive 50 percent plus one vote of the total votes and the margin or extent of the difference is not tested by means of a statistical test.

Effect Size of Designs With a Control Group vs. Effect Size for Designs Without a Control Group

The final hypothesis stated:

H6: There is a greater mean "effect size" shown when a meta analysis is performed on designs without a control group as compared to designs with control groups.

The results of the meta analysis on these two groups is shown in Table 9.

Table 9
Effect Size(s) for Studies With and Without a Control Group

<table>
<thead>
<tr>
<th>Type of Design</th>
<th>Total Number of Studies</th>
<th>Minimum Effect Size</th>
<th>Maximum Effect Size</th>
<th>Mean Effect Size</th>
<th>Standard Deviation of Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>With Control Group</td>
<td>33</td>
<td>-1.17</td>
<td>.59</td>
<td>.10</td>
<td>.42</td>
</tr>
<tr>
<td>Without Control Group</td>
<td>38</td>
<td>-.96</td>
<td>1.13</td>
<td>.21</td>
<td>.40</td>
</tr>
</tbody>
</table>

It is interesting to note that studies which were considered to be of a weaker design (no control group)
did have a greater effect size than those studies which were considered to be stronger (with control group). This difference is in the direction indicated by the hypothesis. A t test was performed to determine if the difference was statistically significant. The results are shown in Table 10.

Table 10

Tests of Means
Control Group Designs vs. Designs Without Control Groups

<table>
<thead>
<tr>
<th>Design</th>
<th>Mean Effect Size</th>
<th>Variance</th>
<th>Obtained t</th>
<th>Degrees Freedom</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>With Control Group</td>
<td>.10</td>
<td>.18</td>
<td>.88</td>
<td>69</td>
<td>.35</td>
</tr>
<tr>
<td>(N=33)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Without Control Group</td>
<td>.21</td>
<td>.16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=38)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Summary

This analysis showed that:

1. The overall mean effect size for all studies is 0.16 standard deviations.

2. The mean effect size for grades K-6 is 0.22 standard deviations and the mean effect size for grades 7-12 is 0.01 standard deviations. This difference was not statistically significant at the .05 level of significance.
3. The mean effect size for math achievement gains is 0.17 and the mean effect size for verbal achievement is 0.07. This difference was not statistically significant at the .05 level of significance.

4. The mean effect size for studies of one year is 0.11 while the mean effect size for those studies that were greater than one year is 0.21. The difference was not statistically significant at the .05 level.

5. The voting method did not produce a different answer than the results of the meta analysis for all (N=71) the studies. The voting method indicated that 60 percent of the studies showed positive achievement effects for desegregation while 40 percent showed no effects. Meta analysis techniques produced a positive overall effect size of 0.16 standard deviations.

6. That the mean effect size for control (comparison) group designs is 0.10 standard deviations while the mean effect size for studies which did not use a control (comparison) group was 0.21 standard deviations. The difference was not statistically significant at the .05 level.
CHAPTER V

DISCUSSION

Review of Problem and Procedures

From the outset this study was designed to use meta analysis techniques to address a particular educational problem. Specifically, this study used meta analysis techniques to aggregate the results of the published studies of desegregation and academic achievement. Meta analysis using effect sizes was chosen because it allows one to compute a quantifiable answer to the questions under consideration and because meta analysis was thought to make use of more information than any other method of aggregation. Also, it is likely that less of the researcher's bias is introduced into the aggregation process since studies were included or excluded on the basis of pre-established criteria.

By use of the meta analysis process, this study investigated the question "Does desegregation lead to improved academic achievement for those children who are desegregated?" In addition, the same meta analysis techniques were used to answer more specific questions about achievement and the possible benefits of desegregation.
Meta analysis techniques were used to address the following questions:

1. Does desegregated education at the elementary level produce greater achievement gains than desegregated education at the secondary level?

2. Are gains in mathematics achievement greater than those in verbal achievement in desegregated environments?

3. Does the length of time in a desegregated educational environment have an effect on achievement?

In addition to these specific educational concerns, this study was also concerned with the viability of the meta analysis technique for addressing studies involved with action research oriented topics. This study compared the results achieved from the meta analysis procedure with results from the more commonly applied voting method technique to check for consistency and to find out if similar answers were provided by the two techniques.

Finally, studies were divided into groups composed of designs which were considered to have the fewest threats to validity and those considered to have more threats to validity. Separate meta analyses were performed on each group to determine if studies which used designs with more threats to validity would produce larger effect sizes than the better controlled studies. The supposition was that studies which used lower quality designs would produce
greater effect sizes than higher quality designs. According to Glass (1978), this supposition should always be tested in a meta analysis study although it may not always be true. If the quality of design is not a factor, it gives the researcher some justification for combining all studies of all designs. If the quality of the design is a factor, the researcher would be wise to perform separate analyses on the two groups of designs.

The procedures used to perform a meta analysis do not require access to the original data for a particular study, but meta analysis does require certain summary statistics. In addition, more than one method can be used to calculate the meta analysis summary statistic "Effect Size" (ES).

Review of Data Requirements and Methods of Meta Analysis

For this study three methods were used to calculate effect sizes. The method used to calculate the effect size for each study was determined by the amount of statistical information furnished by the study's author and the type of design used. For studies which used a pre and a post measure of achievement and a control group there were two distinct methods used. The method used was dependent on the statistical information provided. The two methods used for these designs were:
1. If the researcher provided post means and the standard deviation of the control groups' post tests, effect size could be calculated by subtracting the post mean of the control group from the post mean of the experimental group and dividing by the standard deviation of the control group's post test.

2. If post means and/or the standard deviations were not provided, the alpha level or "p" value transformations provided by Glass (1977) were utilized. This method was useful if significance values were provided and the size of the experimental and control groups were also given.

For designs without control groups, one of the preceding methods was used depending on the information provided in the original study. If the original author provided pre means, post means, pre standard deviations and post deviations, the study was handled much like designs with a control group except the pre test values were treated as if they were control group values and post test values were treated as if they were the experimental group values.

If designs without control groups did not provide the pre-post means and standard deviations but they did provide the critical alpha level used for tests of statistical significance or the calculated (actual) alpha value, they were handled by use of transformations provided by
Glass (1977). As with the former designs the number of subjects in the pre and post testing period was necessary in order to use the transformations suggested by Glass.

Contrasting the Results With the Literature

Several of the purposes and hypotheses that were established in the proposal stages of this study came about after an extensive review of the literature. The assumptions and contentions derived from the literature have been tested by means of the hypotheses which were established. The following discussion will address comparisons of the results obtained by means of a meta analysis with the assertions made in the literature.

The first comparison is concerned with the question of the overall impact of desegregation on academic achievement. With regard to this question, the majority of researchers have concluded that desegregation, as a process, improves the academic achievement of affected minority children. However, several authors (St. John, 1975; Bradley and Bradley, 1977) have concluded that almost all of the studies have severe methodological deficiencies. Hence, it is inappropriate to conclude that desegregation produces positive effects.

With these mixed opinions in mind, a meta analysis was performed on all published studies that provided suitable data and/or statistics for a meta analysis. The meta
analysis of the 71 eligible studies indicated that the effect size was 0.16. This finding is in the direction of beneficial effect for the desegregation process. More specifically, it indicates that the mean of the desegregated children was 0.16 standard deviations greater than the mean of the students who were not segregated. Based on these data, however, one cannot conclude that desegregation is an effective intervention strategy since the observed effect size was not tested for statistical significance. The statistical significance of the overall effect size was not tested by means of a statistical test because the statistical implications of an effect size have not been discussed in the literature. This study was concerned with applying the techniques proposed by Glass to an area of research that might have some practical or applied significance for educational policy making and not to answer specific questions about the use of meta analysis. Glass (1977) did not discuss the statistical significance of the effect size calculated in his research on psychotherapy, and this study will not discuss the statistical significance of the 0.16 standard deviation effect size calculated for this group of studies on desegregation. On the other hand, one cannot say based on this study that desegregation produces harmful effects.

However, the statistical significance of the differences of the mean effect sizes by subject matter, grade level and
length of desegregation were compared by use of an appropriate t test. This decision was considered to be appropriate since these significance tests are based on subjects involved in the desegregated treatment and not between treatment and no treatment groups. Therefore, it is a within group comparison which is not subject to the selection bias inherent in comparing voluntary desegregated education to segregated education.

In addition, the practical significance of an effect size of 0.16 has not been established. This study has not attempted to address the question of practical significance. However, if meta analysis is to be useful for policy decision making, efforts must be made to address the question "What is a practically significant effect size for an intervention such as this?"

In addition to the overall question of the effects of desegregation on academic achievement there are several other educational questions that have been raised in the literature with regard to the specific effects of desegregation when one considers the subject matter, grade level of the children involved, and the length of the desegregation process. Each of these concerns was addressed in the study by a hypothesis and a meta analysis of those specific groups of studies. The first specific educational question addressed in this study was designed to respond to the question "Do children in the lower grade levels (K-6) benefit
more than children in upper grade levels (7-12) as a result of desegregation?"

To test the assertion that desegregated education is more beneficial for the lower grade levels, a meta analysis was performed to determine if the effect size was greater for the lower than for the upper grade levels. This difference was tested by a t test of independent means.

By using meta analysis the effect size for grades K-6 was .22 and .01 for studies concerned with grades 7-12. This difference was not statistically significant with an obtained t value of 1.93. Therefore, it would appear that the difference in effect in favor of lower grade levels was not substantiated by use of the meta analysis technique.

Another educational question that has often been raised in the literature is "Do children in desegregated educational environments gain more in mathematics or in verbal or reading achievement? St. John stated "gains in mathematical rather than verbal achievement are reported frequently enough to deserve further study." (St. John, 1975, p. 37) Her assertion was drawn from a review of studies by Prichard, 1969; Mayer, 1970; Anderson, 1966; Fortenberry, 1959; and Walker, 1969.

In this study, the effect size for studies concerned with mathematics achievement was 0.17 while the effect size for studies concerned with verbal achievement was 0.07. This difference was not statistically significant.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
An assertion which is concerned with the educational benefits of desegregation concerns the length of time a student is exposed to a desegregated educational environment. The critics contend that the reason that most studies of desegregation and academic achievement do not show positive effects is because the child is not exposed to the favorable educational environment long enough to show an effect.

This study used a meta analysis technique to determine if studies concerned with more than one year of desegregation showed a greater effect size. The total group of studies was divided into two groups; one for studies of one year duration or less; and those concerned with more than one year of desegregation. This method produced an effect size of 0.11 for studies of one year and an effect size of 0.21 for studies of greater than one year. Therefore, studies concerned with desegregation did produce an effect size which was 0.10 standard deviations greater for studies of more than one year than studies of one year or less. While the difference was in the predicted direction, a t test produced a t of .96 which was not significant at the .05 level so one cannot contend that this analysis procedure lends validity to the claims made by those critics.
Another issue in the literature is one that is not related to desegregation but to the validity of effects from weaker designs as compared to the effects from stronger designs. Weak and strong are used to differentiate those studies which controlled for more of the threats to internal validity (strong) compared to those studies that controlled for fewer of the threats to internal validity (weak). In this study stronger designs were considered to be those which utilized a control (comparison) group and weaker designs were those studies which did not use a control or comparison group. There is a commonly held assumption that studies which use weaker designs are more likely to show positive effects than studies which control for more of the threats to validity. The argument is that the effects of weaker studies are just as likely to be an artifact of sampling, maturation, history, etc., as they are a "real" result of the treatment used in the study. It is assumed that since stronger designs control for more of the threats to internal validity, a positive effect is more likely to be a result of the treatment and not an artifact of a weakness in the study design.

In their book, Evaluating Social Programs, Rossi and Williams (1977) state that weaker designs are most likely to be used for assessing the outcomes of social action programs such as desegregation programs. They contend
that the potential consumers of such studies should ask themselves two questions about the results of such studies. These are:

1. How much credence can be placed in their results?

2. Under what circumstances can one apply the results of such studies? (Rossi and Williams, 1972, p. 461)

The guidelines offered by Rossi and Williams to potential decision makers state that weaker designs can be used as a sorting mechanism since they are much more economical to apply and because if weak designs do not show positive effects it is unlikely that stronger designs will show them either.

In a letter to the editor of *Educational Researcher* (January 1978, VII, No. 1) Glass replied to a letter from Mansfield and Busse (October 1977, VI, No. 9) on the subject of meta analysis and weak versus strong designs. Glass pointed out that one should oppose an a priori assumption that "weak designs" or "methodological inadequacies" produced the final effect size. He stated in his own investigation of "psychotherapy" (Glass and Smith, 1977) that the "quality of the design" accounted for only one percent of the variance in effect size. Conversely, he pointed out that one of his students (Hartley, 1977) found that "quality of design" did correlate with effects across a hundred or so experiments in mathematics learning.
Clearly weaknesses of design or method should be a genuine concern of a meta researcher.

This study investigated the differences in effect size for weak versus strong designs by calculating a separate effect size for each of the two groups. There were 38 studies without a control group and 33 which used a control or comparison group. The effect size for the weaker designs was 0.21 while the effect size for the stronger designs (with control group) was 0.10. The difference of 0.11 standard deviations was not significant at the .05 level and it seems safe to conclude that the quality of the design was not a factor although weaker designs did have a greater effect size.

This study compared the results of the voting method to the results of the meta analysis (using effect sizes) on the 71 studies that were suitable for meta analysis.

The voting method produced a winner (43 to 28) in favor of desegregation. It has not been established in this study if this margin in favor of desegregation is statistically significant because tests of statistical significance have not been previously applied to the results of the voting technique. Light and Smith (1971)
and Glass (1977) have described this technique in terms that make voting based on treatment or no treatment comparable to voting as it is used in political elections. Therefore, it has not been used to determine the strength of a win or the statistical significance of a win but to determine how many votes were received by the treatment or no treatment group. This deficiency is cited as one of the major weaknesses by Glass (1977) since it seems important to know if a particular treatment wins by a nose or a mile. In this study the desegregated group received sixty percent of the total vote although the statistical significance of this win was not tested.

The meta analysis technique, using effect size, produced an effect size of 0.16 standard deviations. This means that the experimental group's mean was 0.16 standard deviations above the mean of the control or comparison group. Glass (1976, 1977) has not established any decision rules to determine the statistical significance of an overall effect size nor did he report the statistical or practical significance of the overall effect size in his study on the effects of psychotherapy. Therefore, this study did not attempt to determine the statistical or practical significance of the 0.16 standard deviations effect size determined by this study.
The results of this study would indicate that desegregation produces positive achievement effects for affected children which seem to be substantiated by the agreement in the results produced by the voting and meta analysis technique. The comparison of the two methods, in this topical area, would lead one to the conclusion that the two methods would produce the same answer though this conclusion needs to be more firmly established by research in other topical areas which compares the results of these two methods of aggregating research evidence. With regard to this topical area, one could conclude that either method could have been used to determine the benefits of desegregation.

Recommendations for Future Research

Implications for future research in the area of desegregation are presented in this section. Also considered are directions for future research on meta analysis.

When one reviews the literature in a particular area he is likely to uncover a wide range of studies. The studies reviewed will probably range from those that can be considered exemplary with regard to methods, design and analysis, to those which can probably be discounted since the conclusions reached are not tenable due to methodological design, and/or analysis flaws. As such the
reviewer is likely to disregard studies with obvious design defects since there are many rival hypotheses that could explain the outcomes, as well as (or in addition to) the intended treatment. However, the literature in the area of desegregation and academic achievement does not contain the full range of a continuum, from exemplary to mediocre to very poor studies that might be expected. Studies on desegregation and academic achievement are concentrated in the half of the continuum from mediocre to very poor. There are no exemplary studies, and the best are only mediocre by any criteria for method, design and/or analysis (St. John, 1975; Bradley and Bradley, 1977).

These points were raised because these conditions give rise to a set of recommendations that are likely to be raised by anyone who reviews research in this area. In fact, several authors have already made similar points about research in this area (Armor, 1972; St. John, 1975; Webster and Mendro, 1973; Bradley and Bradley, 1977). The effects of poor design, methods, etc., are obstacles a literature reviewer must contend with but when one is using these same studies for purposes of meta analysis, the effects of these poor designs compound attempts to reach conclusions based on such studies. The point is that if the original researcher had performed studies with high quality designs and methods, it would be much
easier to accept the results of a meta analysis based on such studies. When one completes a meta analysis on a group of studies that have low quality designs, the researcher and the reader are left wondering what the results would have been if the designs had controlled for more sources of invalidity.

While this study did separate the accumulated studies into two groups of designs, the only criterion used to sort the studies was the use of a control group versus no control group. There are many other criteria that could be used to rate the quality of a design. It would be instructive in future studies to use other criteria to investigate this question.

With this preceding discussion in mind, this researcher would like to join the ranks of other researchers in this area who proclaim that we must attempt to perform research in this area that controls for more sources of internal validity and thus rules out more of the alternative rival hypotheses. Specifically, several of the design problems which were evident in the group of studies used for this meta analysis will be reviewed and discussed.

The first area of concern is selection bias. None of the studies used random assignment of subject to treatment
or control conditions. This may not be possible in a politically volatile area such as this but more could be done to control for the motivation and family background variables of those that select the desegregated treatment or to measure the values of each group prior to the initiation of desegregation so that family background variables could be used as covariates in the analysis, particularly since the equality of educational opportunity survey (Coleman et al, 1966) concluded that schools have little effect on students' achievement that is independent of their family background. In this group of studies, most researchers let the subjects or their parents determine if they received the treatment or not without assessing pre-existing motivation, values, or demographic variables. Typically, once the researcher received enough experimental subjects, a comparison group was established from those students that chose not to participate. Obviously, this method does not account for the parental differences between volunteers and non-volunteers in terms of educational values, parental help and support in the home, or conversely the negative effects on education by the parents who choose not to participate. It is not known what the educational growth of these volunteer subjects would have been if they had not been placed in a desegregated environment. While it may not be possible to randomly
assign subjects in the naturalistic, political context required of desegregation research, more can and must be done to insure that initially the two groups are comparable in achievement and potential for achievement growth or to use the techniques described by Kenny (1975) to assess treatment effects in non-equivalent control group designs.

Another topic that needs further research attention is concerned with what is the "treatment" actually being delivered inside the school and the classrooms which are the subject of the desegregation research. The treatment for all the studies in this meta analysis is simply the act of placing the minority children in a predominantly white school. None of the researchers attempted to find out the atmosphere inside these schools, or if the children were actually placed in integrated classes within the desegregated school. It might also be worthwhile to investigate teacher differences with regard to academic gains in desegregated environments. It has been stated (West and Anderson, 1976; Rosenthal and Jacobsen, 1968) that teacher expectations are likely to affect students' achievement in the classroom but little research has been conducted in the context of desegregated education and teachers' expectations.
Implications for Educational Practice and Policy Making

When one anticipates a potential applied use for the results of a particular meta analysis, he must be aware of the fact that the resulting effect size(s) will only be one of the factors that might affect a particular policy. It is a generally accepted premise that scientific evidence is only one of the information sources that can affect a change in practice or policy. It is also known that in a democratic pluralistic society such as ours, many other social and political factors must be considered by persons in policy making positions. This reality is of particular importance for policy decisions for public education in the United States. Desegregation in education has been an emotionally charged issue since its beginning in 1954 and it continues to be an educational issue that commands a large degree of attention from the press, politicians, and the general public. Generally, the issue may be described as the "busing problem" although it is not clear whether it is the method of desegregation which is at issue or the overall concept of desegregation in education that creates so much attention and emotional and political fervor. The point is most research which can be used to mold or modify public policy and/or practice must be considered as only one of the many potential contributors.
to public policy. Certainly, research data is a rational factor to the total input if that research is done well. However, it is worthwhile to recall that the decisions we make in our personal lives are based on emotion as well as fact. With these cautions in mind, this section will explore the potential uses of this meta analysis for educational policy making and practice in the area of desegregation in public education.

Three major emphases are indicated by the results of the meta analyses performed for this study.

The first emphasis is concerned with government's role in stimulating research in areas of policy concerns. This study indicated that most of the studies in this area were not considered to be high quality research. Therefore, it would be worthwhile for federal educational departments and agencies to consider the funding of research/evaluation projects which are designed to more definitively address the policy questions of concern.

The second emphasis is concerned with getting state and local practitioners involved as educational innovators. At the state and local level educational administrators and policy makers should consider designing educational programs with the goals of exploring all or most of the variables investigated by this study. These would include educational programs which compare early to later
desegregation, effects of desegregation by grade level, and the effects of desegregation on achievement in various subject areas. These programs could be implemented in conjunction with funded research on their effectiveness and/or impact.

How to Evaluate the Results of a Meta Analysis

Since the meta analysis approach to aggregating the results of individual studies is a relatively new technique, the experience gained in this study should help future meta researchers avoid or minimize some of the pitfalls and difficulties of this method. Some of the problems that arose in the course of this meta analysis and the techniques that were used to deal with those problems are discussed in this section. The discussion will also include the elaboration of problems or potential problems that were not addressed in this particular study. Strategies to address these problems will also be discussed so that future researchers might be aware of these problems prior to the start of their research.

The steps and procedures that are required to accomplish a high quality meta analysis are similar to those required for any research project. The differences are largely concerned with the use of published research.
studies as data versus data collected by the researcher with his own instruments. As such, those problems are related to problems with secondary analysis. When the researcher uses secondary data instead of primary data, he is not aware of all the compromises that were made in collecting the original data, therefore a good deal of faith must be placed in the original researcher's ability and motivation. The reader of a meta analysis is even one more step removed from the original data since he does not have access to the original data but only the summary statistics from the original studies. Because of this fact the procedures, decision rules and boundaries of a meta analysis should be clearly delineated in the meta analysis report. Therefore, the first criterion that should be used to evaluate the validity of a particular meta analysis report is as follows.

Did the meta researcher clearly and completely describe all the steps and procedures he used to delineate the topic under consideration, collect the studies which became data, sort the collected studies and specific methods and techniques used to calculate effect sizes? This criterion is important if the results of a meta analysis are to be replicated. Certainly, if meta analysis techniques are likely to be useful for policy making, they must be reproducible by a second meta researcher or
we would be back to the same predicament we were faced with in trying to use the results of the original study. Which study do we believe?

While it is strongly felt that this is the overriding or macro criterion that might be used to evaluate a meta analysis study, there are several specific criteria that could be used to determine if the meta researcher met this overall criterion.

The first decision that a meta analysis researcher is faced with is the selection of the problem. In the beginning, he may have only a general idea of the variables he is interested in; he will soon discover that the variables for a meta analysis must be as clearly delineated as they should be for any other research study. This will normally come about after he reviews the literature to determine what variables have been researched and by how many researchers. The point is that the independent and dependent variables must clearly be specified so that later sorting decisions are made on pre-established criteria and not for the convenience of the meta researcher. This stage of the research is critical because it sets the stage for what studies will be aggregated with what other studies. In summary, how well does the meta researcher define the problem and variables that will be investigated by his meta analysis? The key is that he must be specific.
but not so specific that he ends up with a very small number of studies to aggregate.

The second specific criterion is concerned with the validity of the data collection process. The validity of the data collection process can be estimated by determining the extent to which these questions are answered:

1. Who collected the data? If more than one person, how was inter-reviewer reliability assured? If it were only one person, how was validity assured?

2. Did the researcher describe all the sorting criteria and decision rules used to accomplish the data collection? Do the sorting criteria introduce any sampling bias problems?

3. Every meta researcher should incorporate some checks on the validity of the data collection process. Does the author describe them? Could they be replicated? Do they seem reasonable or should he have used another approach?

4. Did the author furnish a list of studies that became data for his meta analysis? Are there any obvious sources of bias in the list of studies?

Once the meta researcher has accumulated the studies that will become his data, he is faced with the problem of calculating effect sizes from these studies. Anyone who has attempted a meta analysis will attest to the fact that
the researcher is likely to encounter problems that he did not anticipate in this stage. (Glass (1977) has addressed some of these problems associated with calculating effect sizes. However, since meta analysis is a relatively new technique, other problems and calculation type decisions must be solved by the meta researcher. Several questions about the calculation of effect size(s) should be included in the report. For example:

1. How did he calculate effect size?
2. If more than one method were used, do the decision rules seem logical and consistent?
3. Did the method used introduce any bias? If bias was introduced by the calculation needs, did the author ignore the problem or did he attempt to solve it?
4. Did the methods used to address these bias problems seem complete and logical, or can you think of a better method?

Another concern that should surface during the calculation stage is "the quality of design problem;" that is "Does the quality of the design affect the resulting effect sizes?" As Jackson (1978) has pointed out, one or more methodological flaws in a study, even when serious ones, need not cause biased findings. They only create a threat to validity which may or may not cause a bias. It is imperative that the meta researcher show evidence
that he did check for differential effect sizes based on quality of design so that the reader may intelligently interpret the results. It is also important that the author of a meta analysis carefully and completely specify the criteria used to determine quality and justification for the selection of those particular criteria.

The fifth criterion concerns the conclusions reached by the meta researcher. While meta analysis has many advantages over other analytical approaches to integrative reviews, there are still some serious difficulties. Therefore, it is important that the conclusions drawn from the calculated effect size(s) be interpreted with caution. Since this is such a new approach to aggregating studies, the meta analysis should point out the most logical conclusions as well as conclusions that are likely due to known problems in the data, the calculations or the particular topical area. If meta analysis is going to have a future, it must not be accepted as a panacea or the credibility of this approach will fall into disrepute and it will not be developed to its fullest.

At the beginning of this section it was pointed out that the criteria that should be used to evaluate a particular meta analysis are similar to those criteria that should be used to evaluate a particular research study. In summary, the criteria discussed were as follows.
1. Can the meta analysis be completely replicated by a second meta researcher based on the assumptions given in the meta research report?

2. Are the independent and dependent variables carefully and specifically identified?

3. Can the validity of the data collection process be ascertained?

4. Are the formula, statistics, and methods used to calculate effect sizes described. Do they seem appropriate?

5. Did the author check for differences in effect size(s) due to quality of designs in original studies?

6. Did the author state a definitive conclusion as a result of interpreting the effect sizes or did he qualify the results as he should?

Since the research efforts in this area have just begun, it is anticipated that other meta researchers will discover other criteria that should be used to evaluate the results of a particular meta analysis but at this time it is felt that the six criteria listed above are the major standards by which a particular meta analysis should be judged.

Closing Statements

The experience of conducting this meta analysis has provided some evidence that the meta analysis approach may
be an important contribution to social science methodology if it can be applied and tested on a wide range of aggregation problems. It should not be accepted as a panacea to the aggregation problem since efficacy of meta analysis has yet to be fully established. The main thought that the reader should gain from this study is that meta analysis is a new technique that seems to have more advantages and less disadvantages than some of the other analytical approaches to doing integrative reviews. However, it is not a fully developed approach as of this time. Therefore, all results must be interpreted with care.

Several major concerns in this study could profitably be explored by other meta researchers to further the development of meta analysis techniques and contribute to the knowledge base in the area of desegregation and academic achievement. These include the following:

1. A meta analysis of unpublished studies in the area of desegregation and academic achievement should be accomplished and compared to the results of this study.

2. More work must be done to determine the practical significance of calculated effect sizes. If meta analysis has any hope of being useful to policy makers and administrators, some attempt must be made to introduce practical considerations into the summary statistics.
3. Research should be done to determine the validity of the various methods used to calculate effect size. Several methods had to be used in this study although the comparability of resulting effect sizes has not been established at this time.

4. More work must be done to determine the effect of using unequal group sizes to calculate a mean effect size. At the present time, meta analysis techniques using effect sizes are used to aggregate all studies without regard to the sizes of the groups used in the original individual studies. Therefore, the effect size for a study which used 50 subjects is calculated in the same manner as a study which used 1,000 subjects. It seems clear that a method should be developed to calculate effect sizes based on the sizes of the groups in the original studies.

5. Much more work must be done in the study of the relationship between desegregation and achievement. It is clear that desegregation in education is the law of the land yet we have little or no research evidence that indicates positive effects of desegregation on achievement. This is a call for high quality research to establish causal links between these two variables although it must be conceded that these links may not exist.


Berman, Daniel M. It is so ordered, the Supreme Court rules on school segregation. New York: W. W. Norton & Co., Inc., 1966.


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.


Falk, William W. *School desegregation and the status attainment process: Some results from rural schools.* Louisiana State University, August 1976.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.


Griffin vs. County Board of Prince Edward County, 377 U. S. 218 S. Ct. 1226, 12L Ed. 2nd 256, 1964.


Harris, Gary Reeves. *A study of the academic achievement of selected Negro and white fifth-grade pupils when educational ability is held constant*. Unpublished doctoral dissertation, University of North Carolina, Chapel Hill, 1967.


Harryman, Milburn E. *The effects of racial composition of five high schools in Kansas City upon achievement and citizenship ratings of selected students*. Unpublished doctoral dissertation, University of Kansas, Lawrence, 1967.


Higgins, Paul S. What happened to the kids after their segregated school closed? A nine year control group follow-up of elementary students. Saint Paul, Minnesota: Saint Paul Public Schools, March 1976.


Jackson, Gregg B. The research evidence on the effects of grade retention. Review of Educational Research, Fall 1975, 45 (4), 613-635.


Klein, R. S. A comparative study of the academic achievement of Negro 10th grade high school students attending segregated and recently integrated schools in a metropolitan area in the south. Unpublished doctoral dissertation, University of South Carolina, Columbia, South Carolina, 1967.


Patchen, Martin. *Inter-racial perceptions among high school students*. May 1974. (ERIC Document Reproduction Service No. ED 123 532)


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.


Smith, Annie D. & Johnson, Constance. The impact of desegregation on achievement test scores of black and white students in rural and an urban county: Implications for counseling. April 1976. (ERIC Document Reproduction Service No. ED 133 658)


Van Every, Donald F. Effect of desegregation in public school groups of sixth graders in terms of achievement levels and attitudes toward school. Unpublished doctoral dissertation, Wayne State University, Detroit, Michigan, 1969.


APPENDIX A

STUDIES USED FOR ANALYSIS PURPOSES


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.


Klein, R. S. A comparative study of the academic achievement of Negro 10th grade high school students attending segregated and recently integrated schools in a metropolitan area in the south. Unpublished doctoral dissertation, University of South Carolina, Columbia, South Carolina, 1967.


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.


Smith, Annie D. & Johnson, Constance. The impact of desegregation on achievement test scores of black and white students in rural and an urban county: Implications for counseling. April 1976. (ERIC Document Reproduction Service No. ED 133 658)


Van Every, Donald F. Effect of desegregation in public school groups of sixth graders in terms of achievement levels and attitudes toward school. Unpublished doctoral dissertation, Wayne State University, Detroit, Michigan, 1969.


APPENDIX B

LETTER TO DESEGREGATION EXPERTS

Dear

I am a doctoral candidate at Western Michigan University working on a dissertation in the area of desegregation and academic achievement.

Since you are a recognized expert and scholar in this area, I would like to draw upon your expertise to insure that I have located as many of the relevant studies as possible. I am particularly interested in studies that are concerned with the effects of desegregation on the academic achievement of those children who were integrated.

I have enclosed a list of studies that I have located. These include dissertations, journal articles, ERIC documents and books containing research in this area.

Would you please review this list for completeness and add any studies that you think might be relevant for my research. You are probably very busy at this time of year but I hope you will take the time to review my list and add published studies you are aware of. My committee has required that I seek experts to review my references.

I hope you can find the time to forward your written reply by August 15, 1978 or call me at 313-763-1153. If I do not hear from you by August 15, 1978, I would like to call you for your response. Thank you in advance for your consideration to this request.

Sincerely,

Ronald A. Krol
APPENDIX C

LIST OF DESEGREGATION EXPERTS

Dr. Nancy H. St. John
University of Massachusetts
100 Arlington Street
Boston, MA 02116

Dr. David J. Armor
Rand Corporation
1700 Main Street
Santa Monica, CA 90406

Dr. Lawrence A. Bradley
Assistant Professor of Psychology
Fordham University
Fordham Road
Bronx, NY 10458

Dr. Gifford W. Bradley
Department of Psychiatry
Columbia Presbyterian Medical Center
New York, NY 10032

Dr. Thomas F. Pettigrew
Associate Professor of Social Psychology
Department of Social Psychology
1330 Wm James Hall
Cambridge, MA 02138

Dr. Meyer Weinberg
Department of History
Loop College
64 East Lake Street
Chicago, IL 60601
APPENDIX D

LETTER TO ORIGINAL RESEARCHERS

Dear Dr.

I am a doctoral candidate who is completing a dissertation in the area of desegregation and academic achievement. As part of my literature review, I located your article, and found it to be worthwhile contribution to the literature.

However, in order to aid my undertaking of your research I would like some additional information that was not included as part of your published article. The information I need is indicated on the attached form. Please indicate grade level(s), pre/post means and standard deviations, size of each group and significance tests used.

To save time you may fill in the attached form and return it to me at your earliest convenience.

I sincerely hope you can comply with this request because I feel that these additional statistics would help me to more fully understand your research.

Thanks in advance for your consideration of this request.

Sincerely,

Ronald A. Krol

175

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
# APPENDIX E

## EXAMPLE OF FORM SENT TO ORIGINAL RESEARCHER

<table>
<thead>
<tr>
<th>Author</th>
<th>Study No.(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(please enter data in spaces marked with an "X" and return completed form)

<table>
<thead>
<tr>
<th>Grade Level(s)</th>
<th>PRE MEANS</th>
<th>PRE STD. DEV.(S)</th>
<th>POST MEANS</th>
<th>POST STD. DEV.(S)</th>
<th>Significance TEST(S) Used if any (.05,.01)</th>
<th>Exact Calculated Value(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N=</td>
<td></td>
<td>N=</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>N=</td>
<td></td>
<td>N=</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Please Return Completed Form To:

Ronald A. Krol
University of Michigan Medical School
G116 Towsley Center
Ann Arbor, Michigan 48109
APPENDIX F

LIST OF ORIGINAL RESEARCHERS

Dr. David J. Armor
Rand Corporation
1700 Main Street
Santa Monica, CA 90406

Dr. James Chester Bryant
909 Wiberly Street
Angleton, TX 77515

Dr. Kenneth E. Cypert
985 Mockingbird Lane
Stephenville, TX 76401

Dr. Ann Hechter Danahy
1457 Eleanor
St. Paul, MN 55116

Dr. Burleigh Gardner
857 Burr Avenue
Winnetka, IL 60093

Harold B. Gerard
University of California at Los Angeles
Department of Psychology
405 Hilgard
Los Angeles, CA 90024

Carl F. Hansen
Department of Education
University of Arkansas
Little Rock, AR 72201

Dr. Jayjiia Hsia
Director of Mid-Western Education Testing Service
960 Grove Street
Evanston, IL 60201

Dr. Edward F. Iwanicki
315 Mount Vernon Street
Dedham, MA 02026

Dr. Charles E. King
Department of Sociology
North Carolina Central University
Durham, NC 27707

Dr. Roger Brasel Marcum
615 Aberdeen
Cary, IL 60013

Dr. Paul Prichard
423 N. Greensboro Street
Carrboro, NC 27510

Dr. George Jacob Rentschler
Department of Administration and Supervision
Louisville Public Schools
2301 South 3rd Street
Louisville, KY 40208

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.