ED 239 003 UD 023 306

AUTHOR Wortman, Paul M.
TITLE School Desegregation and Black Achievement: An Integrative Review.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE 18 Feb 83
GRANT NIE-G-79-0128; NIE-P-82-0070
NOTE 56p.; Paper prepared for the National Institute of Education Panel on the Effects of School Desegregation. For related documents, see UD 023 302-308.
PUB TYPE Information Analyses (070) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Achievement Gains; *Black Students; *Desegregation Effects; Effect Size; Elementary Secondary Education; *Evaluation Criteria; Mathematics Achievement; *Meta Analysis; Outcomes of Education; Program Effectiveness; Program Evaluation; Reading Achievement; *Research Methodology; Research Reports; School Desegregation

ABSTRACT

The focus of this paper is on the methodology used in analyzing studies on the effects of school desegregation on black achievement. The paper also addresses a number of substantive issues, including the overall effectiveness of school desegregation, the impact of type of achievement (math or reading), and time of desegregation (early or later grades). In the first section, a discussion of the advantages and disadvantages of meta-analysis is presented. The complications are explored of using the method to analyze the research literature on the effectiveness of school desegregation when that literature is almost totally composed of quasi-experimental or weaker research designs. In the second section, the adjustments that were made in the meta-analysis method for the purposes of the National Institute of Education (NIE) panel's study on the effects of school desegregation are described. The procedures and criteria used for including the 31 studies finally chosen for analysis by the author are described in the third section, as well as the criteria used by NIE in narrowing the selection down to 19 studies for consideration by the panel. In section 4, the results of analyses of both sets of studies are given. An appendix lists the studies used in both analyses. (CMG)

SCHOOL DESEGREGATION AND BLACK ACHIEVEMENT: AN INTEGRATIVE REVIEW

Paul M. Wortman, PhD

University of Michigan

February 18, 1983

This paper is the final report to the National Institute of Education Panel on the Effects of School Desegregation. Work on this paper was supported by funds from the National Institute of Education, (NIE-G-79-0128 and NIE-P-82-0070). The author gratefully acknowledges the assistance of Dr. Fred Bryant. The views expressed in this paper are those of the author and do not necessarily represent the views of the National Institute of Education.

TABLE OF CONTENTS

TABLES AND FIGURES

1.0 PROBLEM

Race relations between blacks and whites have played a significant role in the history of the United States. Social science theory and data, in particular, have figured prominently in the controversies that have constantly surrounded major events in this history. For example, the two landmark U.S. Supreme Court decisions dealing with desegregation, Plessy v. Ferguson in 1896, and Brown v. Board of Education in 1954 (Kluger, 1975) were both based in part on current social science evidence. More recently, the so-called Coleman Report or the Equality of Educational Opportunity Survey (Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld & York, 1966) was used by the Johnson administration to accelerate the desegregation process (Grant, 1973). The Coleman Report claimed that black student achievement increased in more integrated environments (i.e., with a greater proportion of white students). This study and finding not only led to a number of reanalyses by social scientists, but also to an increasing number of systematic studies using before and after measures (i.e., pretests and posttests) of achievement and control or comparison groups of segregated blacks. These studies aimed at eliminating the methodological weaknesses of cross-sectional surveys such as the Coleman Report and testing some of its hypotheses and those of other social scientists.

By the mid-1970's there had accumulated a sufficient body of scientific studies that a number of careful reviews appeared. Two of the most notable of these reviews were conducted by Bradley and Bradley (1977) and St. John (1975). The Bradleys examined 29 studies of the effects of desegregation on black achievement while St. John reviewed [illegible] (including 12 cross-sectional studies). Both found the evidence inconclusive. The Bradleys concluded that the evidence on the effectiveness of desegregation on black achievement was "inconsistent and inadequate" while St. John similarly acknowledged, "More than a decade of considerable research effort has produced no definitive positive findings." St. John went on to quote Light and Smith (1971) that "'progress will only come when we are able to pool, in a systematic manner, the original data from the studies.'" Such methods for synthesizing the results of scientific studies have recently gained widespread popularity largely due to Glass' seminal work on "meta-analysis" (1976, 1977).

Meta-analysis offers a number of advantages over previous methods for aggregating the findings of different studies (Light & Smith, 1971; Glass, 1977). In Table 1 we have listed some of the positive and negative characteristics of this technique. The major positive qualities are a single, precise, quantitative measure of the average magnitude of program impact. It is applicable to most social science research and provides an important result that is easy to grasp. Meta-analysis also allows one to consider sample size and design quality. This technique also has its "disadvantages" especially when extended to studies with methodological problems such as quasi-experiments (i.e., studies lacking random assignment).

Standard meta-analytic methods have already been applied to this literature (Crain & Mahard, 1982; Krol, 1978). The meta-analyses performed by Krol and Crain and Mahard both found small positive benefits for desegregation on black achievement. (.16 and .08 standard deviations, respectively). Both are flawed in our opinion. Krol's study illustrates the inappropriate application of Glass' method. For

Table 1

Advantages and Disadvantages of Meta-analysis for Quasi-experiments[1]

Definition	Advantages	Disadvantages
<u>Meta-analysis Method</u> The average effect size of a hypothesis tested in many studies. The term connotes "the analysis of analyses, i.e., the statistical analysis of the findings of many individual analyses."	o Precise determination of effects o Systematic, statistical approach o Design quality can be examined o Can examine effect of sample size o Includes some descriptive information	o Susceptible to publication bias o Requires a control group o Requires statistical information o Assumes a "common metric" for measure o Assumes the "strategic combination argument"

[1] Adapted from Krol (1978)

example, Glass (1977, p. 356) does recommend using pre-experimental designs lacking controls "*if* the treated group members' pretreatment status is a good estimate of their hypothetical posttreatment in the absence of treatment." As we will demonstrate in the next section, this suggestion may be unwarranted and ill-advised. Crain and Mahard (1982) in a very recent meta-analysis have taken a traditional Glassian approach and included all studies in their analysis. As we shall indicate below, we feel this approach is inappropriate. Many studies have so many methodological weaknesses that they should not be included. Moreover, some studies such as those using a cross-sectional survey cannot yield the necessary statistical information (since they lack both a pre-desegregation or pretest measure as well as a control group), but were included by Crain and Mahard. Other studies used white control groups or national test norms to generate effect sizes -- both are inappropriate comparisons as will be discussed below. Such studies account for half of those included in Crain and Mahard's meta-analysis. Most importantly, however both Krol and Crain and Mahard paid insufficient attention to the threats to validity that could confound and bias the results of their meta- analyses.

The school desegregation-achievement literature poses some special problems for the meta-analysis method. It is almost entirely quasi-experimental in composition and thus susceptible to other interpretations (i.e., so-called "plausible rival hypotheses"). Meta-analysis of such studies assumes that either appropriate statistical adjustments can be made for the various "threats to validity" or that the "strategic combination argument" (Staines, 1974) holds (see "disadvantages" in Table 1). This latter term stands for the belief that flawed studies can be combined because the "weaknesses cancel each other out." It is just this argument that Glass (1977) used in recommending meta-analysis of "weak"studies. While Glass was initially confident that his method could be used with quasi-experiments, his views have gradually changed (cf. Glass & Smith, 1979). The examination of the desegregation quasi-experimental studies presented in the following sections indicates that selection is a persistent "plausible rival hypothesis." That is, it is not cancelled out. Therefore, a number of steps have been taken to deal with this. First, an adjustment was developed for reducing the bias due to selection. Second, studies that were judged a priori not to have selection problems were compared with those requiring adjustment.

The focus of this paper is on the effect of school desegregation on black achievement. While interest in these data is primarily methodological and stems from earlier work by the author on the secondary analysis of the Riverside School Study (RSS) of desegregation (Linsenmeier & Wortman, 1978; Moskowitz & Wortman, 1981), a number of substantive issues are addressed. In addition to estimating the overall effectiveness of desegregation, such issues as the impact of type of achievement (math or verbal) and time of desegregation (early or later grades) are also discussed. This latter, substantive focus qualifies this study as an "integrative review" (Jackson, 1980). In the next section, the meta-analytic method used in this study is described. As the "disadvantages" column in Table 1 indicates not all studies are suitable for meta-analysis. Those with numerous or severe methodological flaws, inadequate reporting of statistical information, or insufficient control data were not included. In the third section, the procedure for including studies in the analysis is described. The results and conclusiõns are presented in the last two sections.

2.0 METHODOLOGY

To apply meta-analysis to quasi-experimental data one needs to obtain a measure of "effect size" (ES). The basic equation adopted from Cohen (1969) is:

$$ES = \frac{(\bar{X}_E - \bar{X}_C)}{S_C} \qquad (1)$$

where,

$\bar{X}_E$, $\bar{X}_C$ = the means for the treatment
(i.e., desegregation) or experimental (E)
and the control (C) or untreated (i.e., segregated groups
S_C = the standard deviation of the control group[1]

In the quasi-experimental case we have the following:

$$ES = \frac{(\bar{X}_{E_2} - \bar{X}_{C_2})}{S_{C_2}} - \frac{(\bar{X}_{E_1} - \bar{X}_{C_1})}{S_{C_1}} \qquad (2)$$

where,

1,2 indicate time 1 (pretest) and time 2 (posttest)

In a randomized experiment, $\bar{X}_{E_1}$, $\bar{X}_{C_1}$, yielding Equation 1. However, this assumption is not guaranteed in a quasi-experiment. In this situation it is likely that the groups will differ initially. That is, selection is a major threat to validity that is represented in this model.

Meta-analysis involves summing of the effect size estimates from all studies. We define it as:

$$\Sigma ES = \Sigma_i \left[\frac{\left(\bar{X}_{E_{2i}} - \bar{X}_{C_{2i}}\right)}{s_{2i}} - \frac{\left(\bar{X}_{E_{1i}} - \bar{X}_{C_{1i}}\right)}{s_{1i}} \right]$$

where,

> $\bar{X}$ is the sample mean of the experimental or control group at time 1 and 2 for the i^{th} study and s is the control group standard deviation.

The average effect size, Δ, is usually presented. This average can be computed in a number of ways. For example, all ESs can be summed and averaged. Since many ESs may be derived from a single study, this introduces bias due to nonindependent measures. It was largely for this reason that Landman and Dawes (1982) reanalyzed Smith and Glass' (1977) meta-analysis of the effectiveness of psychotherapy.

The desegregation literature is largely composed of quasi-experiments or even more poorly designed studies. As such, it is susceptible to a variety of threats to internal validity (i.e., the ability to infer causality). It is risky to assume that these potential sources of bias can be treated as random errors that are self-cancelling. Two threats in particular, have been much discussed in reviews of this literature. They are "selection" and "differential growth" or "maturation". These are considered in the next paragraphs; other threats to validity are discussed in the next section.

Selection

Campbell and his associates (Campbell & Erlebacher, 1967; Campbell & Boruch, 1975; Campbell & Stanley, 1966; Cook & Campbell, 1979) have been concerned with the recurrent problem in estimating program effects when various selection procedures are used. In particular, they have discussed selection of those students with extreme (pretest) scores and/or matching experimental and control subjects by (pretest) score. Both of these selection procedures are subject to substantial "regression artifacts" resulting from the unreliability of the measures used. While there is no agreed-upon procedure for adjusting for these selection effects, a number of methods have been developed (cf. Wortman, Reichardt, & St. Pierre, 1978). These methods require both student-level data and test reliabilities in order to be applied. That information is generally not reported in the studies of desegregation and would require reanalysis of individual studies if available. Instead, the pretest adjustment procedure described in Equations 2 and 3 will be employed. Since matching was rarely used, this method should adjust for the selection or "subject equivalence" problem that Bradley and Bradley (1977) and St. John (1975) found to be the major methodological weakness in the better or "well designed" studies. Neither Crain and Mahard (1982) nor Krol (1978) attempted to correct or adjust for bias introduced by initial subject nonequivalence.

Differential Growth

It is well-known that blacks and whites show different rates of intellectual growth. Thus differential growth or "maturation" may be considered an important source of bias in synthesizing the data from the desegregation literature. This problem is dealt with in three ways:

conceptually, empirically and analytically. First, only studies using black controls were examined. This is the comparison recommended by St. John (1975) and should reduce or eliminate the problem. Such controls avoid problems (or confounds) caused by race and socioeconomic status. They also allow examination of the major policy question being addressed: the effect of continued racial isolation or segregation. Fortunately, most studies used such a control group (i.e.. segregated blacks). As noted above, both Crain and Mahard (1982) and Krol (1978) included studies that used white controls.

Second, the results of the pretest adjustment are compared to those studies not requiring such corrections (i.e., no pretest differences) to determine if other differences or sources of bias remain. As will be noted, "differential regression to the mean" (Cook & Campbell, 1979) may account for the residual difference. And third, the analytic method is examined to determine its robustness to this source of bias. It may be recognized that Equation 2 is identical to the model for differential growth rates labelled by Campbell the "fan spread hypothesis" (Campbell & Erlebacher, 1970; Cook & Campbell, 1979). In fact, if differential growth is the only cause of change from time 1 to time 2, then according to the fan spread model:

$$\frac{\bar{X}_{E_1} - \bar{X}_{C_1}}{S_1} = \frac{\bar{X}_{E_2} - \bar{X}_{C_2}}{S_2}$$

This hypothesis implies that an increase in the mean is accompanied by a proportional increase in the within-group variance. Thus, ES = 0 when this "threat to validity" (i.e., differential growth) is present. This means that selection-maturation interaction will not bias the estimate of effect size for quasi-experiments of this type (i.e., the nonequivalent control group design or NECGD) that are pretest-adjusted. This is exactly the model proposed by Campbell (1971) and described by Kenny (1975). As Campbell and Boruch (1975) note, standardizing scores will eliminate this problem. The effect size measure as defined above in Equation 1 is a standardized score.

Practical Limitations

There are a number of problems in translating this small analytic model into an actual meta-analysis. First, the NECGD requires the means and standard deviations for the experimental and control groups on both the pretest and posttest. Often these essential data are not furnished especially in those cases where statistically non-significant results were obtained. The reliability of the tests used is even less likely to be reported. In order to deal with this situation, a variety of indirect approaches have been proposed (cf. Glass, 1977).

Using Significance Results. Reports often provide only information on sample size, significance level, and the value of the test statistic. In these cases the effect site can be obtained using indirect methods. In the case of the *t*-test, it is:

$$ES = t\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\text{from } t = \frac{\bar{X}_E - \bar{X}_C}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $n_1 = n_2$ and thus about half of the degrees of freedom (df), then according to Rosenthal (1978):

$$ES = \frac{2t}{\sqrt{df}}$$

This indirect estimate will be conservative when the exact significance level is not reported, and the $\underline{t}$ value is not given. Typically, the .05 or .01 significance levels are used in social science research. If the results are not significant, little if any information is usually provided. In this case, a .50 significance level will be used as Cooper (1979) has suggested. This is the expected mean value of the distribution of non-significant studies. Similar indirect computations can be derived from other test statistics such as $\underline{F}$ (see Appendix 7 in Smith, Glass, & Miller, 1980).

Gain Scores. Another common form of reporting results is the gain score. This is the change in each group from pretest to posttest. In Figure 1 this would be:

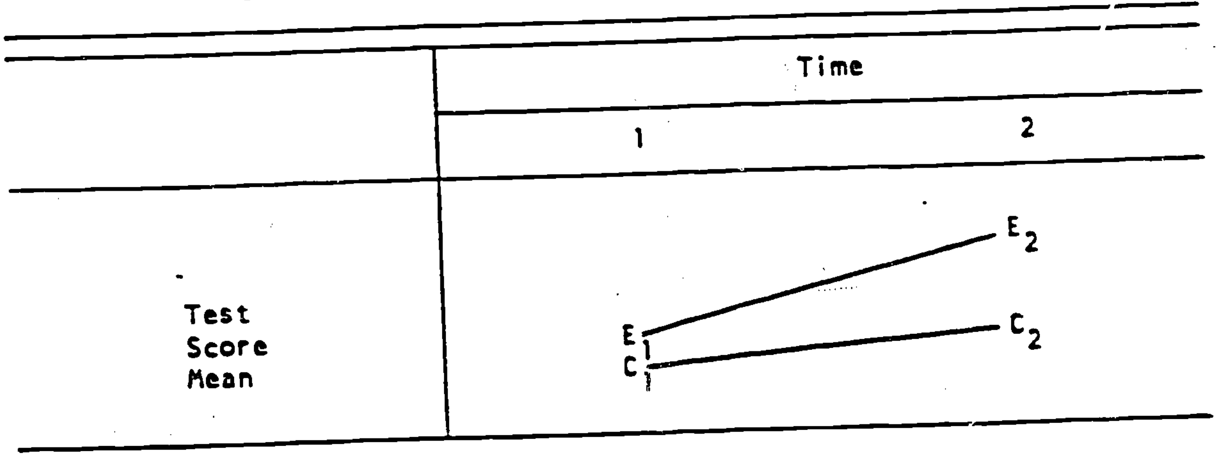
$$\text{gain}=E_2-E_1 \text{ and } C_2-C_1,$$

for experimental and control groups, respectively. A simple algebraic manipulation reveals that the difference in the two gain scores is equivalent to the numerator in the basic equation to estimate the effect size for quasi-experiments (Eq. 2). Thus if $s_1=s_2$, gain scores can be used to derive $\underline{d}$ for the NECGD quasi-experiment.

Other Quasi-experimental Designs. Other quasi-experimental designs are often encountered and it is important to consider them as well. The most frequently reported is the case study or in Campbell and

Figure 1

Hypothetical Results From a Study
Using a Nonequivalent Control Group Design (NECGD)



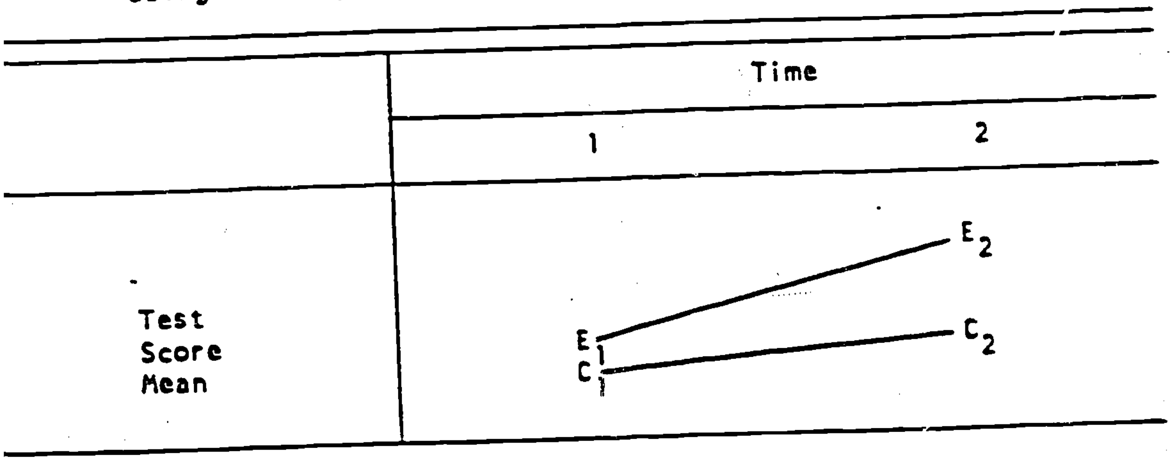
E=Experimental Group
C=Control Group

Stanley's terminology, the One-Group Pretest-Posttest (OGPP) Design. This is the NECGD without the control group. Krol (1979) suggests that an effect size estimate can be obtained by using the pretest mean and standard deviation as the control group. This is a risky assumption in our opinion, and one that is likely to lead to an overestimate of ES. As can be readily seen in Figure 1, the use of the standardized gain score (E_2-E_1) contains a pseudo-effect equal to C_2-C_1. Moreover, if strict selection criteria are used as they often are in compensatory education or competency testing remediation programs, then regression effects will also be incorrectly included. Thus we feel such case study data should only be used when the proper adjustments can be made. In order to examine design effects in meta-analysis, a number of these case studies were included in some of the analyses.

Control group data are frequently difficult to obtain for political and practical reasons. Programs may be designed to serve all

Figure 1

Hypothetical Results From a Study
Using a Nonequivalent Control Group Design (NECGD)



Experimental Group
Control Group

anley's terminology, the One-Group Pretest-Posttest (OGPP) Design. is is the NECGD without the control group. Krol (1979) suggests that effect size estimate can be obtained by using the pretest mean and andard deviation as the control group. This is a risky assumption in ir opinion, and one that is likely to lead to an overestimate of ES. can be readily seen in Figure 1, the use of the standardized gain ore (E_2-E_1) contains a pseudo-effect equal to C_2-C_1. Moreover, if trict selection criteria are used as they often are in compensatory ducation or competency testing remediation programs, then regression ffects will also be incorrectly included. Thus we feel such case study ata should only be used when the proper adjustments can be made. In rder to examine design effects in meta-analysis, a number of these case tudies were included in some of the analyses.

Control group data are frequently difficult to obtain for political and practical reasons. Programs may be designed to serve all

True Experiments. Although our focus has been on quasi-experiments, "true" or randomized studies would be useful. Just as we were concerned about the biased estimates produced by pre-experimental design (i.e., case) studies when compared to the NECGD quasi-experiments, it is important to determine the bias resulting from the latter designs. This information can be obtained if effect size estimates are available from randomized studies. Not all data sets have this mixture of designs, especially in education where there has been a strong tendency for applied. field problems to be approached quasi-experimentally while laboratory, theoretical issues have been investigated using randomized studies. There have been a few randomized studies or true experiments in the school desegregation area. Those that have been conducted such as Project Concern (Iwanicki & Gable, 1978) often report their results in such a way as to make it impossible to derive effect size estimates.

Crain (1983) identified five randomized studies among his top 20, three of which were based on data from Project Concern. Three of these studies (Rock et al., 1968; Samuels, 1971; Zdep, 1971 -- see Appendix A) were included among the 31 found acceptable in the present analysis. A more recent report from Project Concern (Iwanicki & Gable, 1978) was included in place of the two earlier reports used by Crain.[2]

Design Quality

Although the focus is on the NECGD, the quality of the studies using this design varies. Moreover, as noted above, there are often other designs employed. A number of approaches to assessing quality have been developed. The most well-known is the validity approach developed by Campbell and Stanley (1966) and recently further refined by

Cook and Campbell (1979). Essentially, the threats to validity indicate quality. Others (Boruch & Gomez, 1977; Sechrest & Yeaton, 1981) have stressed the "implementation" or "integrity" of the treatment. This is an important concept although one that is difficult to measure. The assessment of research quality is a new area and one that is critical in the synthesis of scientific studies. There has been much discussion of this issue (Mansfield & Busse, 1977; Eysenck, 1978; Glass, 1977, 1978) and the debate still continues (cf., Wortman, 1983). As the following section indicates, design quality is viewed as significant in selecting, coding, and analyzing the data in a research synthesis.

3.0 PROCEDURE

The meta-analysis approach first requires the retrieval of relevant scientific information. The importance of a thoroughly documented procedure at this point has been stressed by both Cooper (1982) and Jackson (1980). To that end, we obtained the cooperation of the authors of the two major studies systematically synthesizing the literature on the effects of school desegregation on black achievement (Crain & Mahard, 1978; Krol, 1978). Both Robert Crain and Ronald Krol generously provided copies of the articles and the coding schemes used in their analyses. We then extended and updated this data base through literature searches including ERIC, dissertation abstracts, references in the articles and books (especially, St. John, 1975), and dozens of letters to authors and school district offices. We developed a coding scheme and list of studies to be included in our analyses. These are described below. As we progressed with our initial coding effort, we realized that there were many studies that would have to be rejected. We felt it imperative to describe these studies and our reasons for rejecting them from the analysis. We did this for two reasons: (a) this is perhaps the most important, but judgmental, step in data synthesis, and (b) it is important to determine whether there are unique characteristics of excluded studies. All studies were read and coded by two independent reviewers. All discrepancies were resolved so that perfect agreement was reached. A more detailed description of this procedure and the studies excluded can be found in an earlier technical report (Wortman, King & Bryant, 1982). In the next three sections we discuss both of these concerns.

Exclusion Criteria. The decision to exclude a particular study from the analyses was based on assessments of the various threats to the study's validity. The number and magnitude of the flaws in the study were the deciding factor for inclusion or exclusion. The observed threats to validity fall into one or more of four basic classifications that have been developed by Campbell and his associates (Campbell & Stanley, 1963; Cook & Campbell, 1979). Thus, the criteria used to reject studies (see Table 2) represent specific instances or threats to internal, external, construct, or statistical conclusion validity.

Internal validity is broadly concerned with whether the treatment (i.e., school desegregation) in fact affected the outcome (i.e., academic achievement of black students). Threats to internal validity may be posed by uncontrolled variables representing effects of history, maturation, and the like as originally described by Campbell and Stanley (1963). Most of the factors listed in the table as threats to validity do not require further explication. However, the rationale behind a few may not be so apparent. For instance, studies utilizing cross-sectional survey designs (criterion 4a) were rejected from the analyses because they typically do not control for extraneous variables in local school settings that may affect achievement above and beyond the effects of desegregation. That is, they are usually observations at one point in time lacking both pretests and adequate controls.

Studies were also rejected that failed to describe their sampling procedures (criterion 4b) and thus make it impossible to rule out potentially confounding biases in the selection of comparison groups. Finally, the use of different tests for segregated and desegregated students at either pretest or posttest may pose "instrumentation"

Table 2

Criteria for Selecting Studies for Meta-analysis

Criteria for Rejection	Threats to Validity			
	Internal	External	Construct	Statistical
1) Type of Study:				
*a) Non-empirical			X	
*b) Summary report: insufficient detail for coding			X	
2) Location:				
*a) Outside U.S.A.		X		
*b) Geographically non-specific		X		
3) Comparisons:				
*a) Not study of achievement of desegregated Blacks			X	
*b) Multi-ethnic data combined			X	
*c) Comparisons across ethnicities only			X	X
*d) Heterogeneous proportion minority in desegregated condition			X	
*e) No control or pre-desegregation data	X			
*f) Control measures not contemporaneous	X			
g) Multiple treatment interference		X		
h) Excessive attrition	X			
*i) Majority black in desegregated condition[1]			X	
*j) Varied exposure to desegregation[1]	X			
k) Groups initially non-comparable	X			
4) Study Design:				
*a) Cross-sectional survey	X			
*b) Sampling procedure unknown	X			
*c) Separate non-comparable samples at each observation	X			
d) Grade levels grossly combined				X
e) Inadequate sample size				X
5) Measures:				
*a) Unreliable and/or unstandardized instruments	X			
*b) Test content unknown			X	
*c) Dates of administration unknown		X		
*d) Different tests used at pretest and posttest	X			
*e) Test of IQ or verbal ability			X	
6) Data Analysis:				
*a) No pretest means				X
*b) No posttest means				X
*c) No pretest standard deviations[2]				X
*d) No posttest standard deviations[2]				X
*e) No significance tests				X
*f) No data reported				X
*g) N's not discernable				X
h) Inappropriate statistics				X

*Criteria used to select NIF Core Studies

[1] For the NIE Core Studies these criteria were relaxed to allow studies that provided "specific justification" for this.

[2] For the NIE Core Studies these criteria were combined into a single criterion, unable to calculate effect sizes.

problems stemming from differential test reliability and low inter-test reliability. These problems may either produce spurious treatment effects or mask real effects. Each of these specific threats may confound the observed association between desegregation and achievement.

External validity refers to limitations in the generalizability of the study with regard to populations, settings, as well as treatment and measurement variables. One obvious reason for exclusion was studies conducted outside of the United States. Another common threat to external validity involved the confounding effect of compensatory equalization of treatment (e.g., extra teachers for segregated controls), or other kinds of multiple treatment interference (criterion 3g). These may disguise or distort findings indicating how desegregation affects achievement. Moreover, when the dates of test administration are not described (criterion 5c), problems arise in adjusting the effect-size estimates to a proper time interval as well as determining whether the pretest actually occurred prior to desegregation.

Construct validity refers to the appropriateness of the theoretical constructs, variables, and measures used. If the study did not really deal with desegregation and/or achievement, it was not included. Other studies were rejected on these grounds, but for less obvious reasons. These include those that at first appear to measure academic achievement of desegregated blacks, but which, in fact, measure a different construct such as I.Q. (an ability measure); those that measure a different treatment, such as bus transportation; or a different population such as whites or Chicanos (see criterion 3a).

Statistical conclusion validity is concerned with the appropriateness of the statistical analyses. This includes not only the analyses employed but also the sufficiency of the data reported for calculating effect sizes. For example, a study may improperly use ANOVA in the analysis of a non-equivalent control group design (i.e., criterion 6h) that violates assumptions of homogeneity of variance and of heteroscedasticity. Other studies may correctly employ statistical procedures where there is inadequate statistical power from sample sizes too small to reject the null hypothesis. Finally, studies which grossly combine achievement results of different grade levels must be rejected because the rate of achievement gain tends to increase more slowly with advancing grade level and thus grade-equivalent scores are really not comparable (as they are normed within each grade separately). Combining scores from various tests across grade levels further threatens internal validity insofar as instrumentation effects arise from variations in test reliability and other test characteristics (e.g., item difficulty and content).

Applying the criteria listed in Table 2 resulted in the exclusion of 74 studies. Most suffered from more than one problem. A number of these criteria are sufficient in themselves (i.e., "fatal flaws") to eliminate a study. All but three studies had such flaws. Overall, we have had to exclude the majority of studies examined including a number used in the previous meta-analyses performed (Crain & Mahard, 1978; Krol, 1978). A comparison of studies included and excluded is provided in Table 3. With the exception of Crain and Mahard (1978), we included only about half of the studies used in other major reviews. The 31 studies included in our analyses are listed in Appendix A. The studies were decomposed into effect size data for each grade and for reading and mathematics achievement, and thus yielded 106 separate "cases". The overall analyses, however, used the study as the unit of analysis by averaging the results within each study and combining these average effect sizes.

Table 3

Comparison With Previous Research Syntheses

PRESENT CASES	% of PRESENT CASES USED BY PAST INVESTIGATORS			
	KROL	CRAIN & MAHARD	WEINBERG	ST. JOHN
REJECTED (n=229)	13%	60%	25%	26%
ACCEPTED (n=106)	36%	87%	51%	57%

A considerable amount of effort was spent in documenting this aspect of the research synthesis. It represents an important, but often overlooked, part of formal data synthesis procedures, and one that can produce differing results. While meta-analysis, itself, is a formal, quantitative method, the selection of the sample to include in the analysis is not. Without appropriate, documented selection criteria, the results can be as subjective and biased as the literature reviews they seek to replace, (cf. Jackson, 1980)

One "disadvantage" of meta-analysis (see Table 1) is its susceptibility to publication bias. It is assumed that the research literature contains only studies showing positive, statistically significant results (i.e., publishable studies). The 31 studies found "acceptable" contained only two published articles. Desegregation research is largely (and perhaps appropriately), a fugitive literature. We feel that the retrieval strategy described abo[illegible]s captured the "target population" of studies (Cooper, 1982).

The NIE Core Studies

After this screening process had been performed and the 31 resulting studies analyzed, the NIE Desegregation Studies Team convened an expert panel to select the best studies in this area. The panel of six scholars including this author was supposedly balanced in their attitudes and published work on desegregation -- two pro, two con, and two neutral.*. The panel met in July, 1982 and initiated discussion of the most appropriate studies to be included in reviewing the literature. The criteria listed in Table 2 were examined by the panel and after some discussion a subset of them was used to select the highest quality studies available. In general these were NECGD studies comparing verbal and/or math achievement of desegregated and segregated blacks. The criteria actually used are starred in the table.

These criteria were entered into the computerized data base and 18 studies were found that satisfied these requirements. These studies are starred in Appendix A. One new study by Walberg (1971) was added at the request of some of the panel members. This study had been "rejected" in the original analyses since it suffered from an extremely high rate of attrition (criterion 3h) that differed for segregated and desegregated students (i.e., 27 and 48 percent, respectively). The number of students in the desegregated control group was quite small, ranging from 14 to 53. Moreover, grade levels were combined (criterion 4d). The Walberg study added eight "cases" to the data base. Moreover, one of the panelists wrote to one of the author of another study (Sheehan, 1979) to obtain missing means and standard deviations. This allowed the inclusion of two additional cases.

These studies differ substantially from those used in most previous reviews. With the exception of Crain and Mahard (1978) where all, but one, study was included, fewer than half were included in prior reviews. for example, .Bradley and Bradley (1977) included only five of these studies while St. John (1975) reviewed only nine of them.

4.0 RESULTS

The Glass effect sizes (ESs) for the 31 studies considered methodologically acceptable for performing a meta-analysis are presented in Table 4. The fourth row labelled "Grand" presents the overall effects averaged by study (i.e., the average of the average effect sizes for each study) and the ESs by three major research designs. In addition, these four categories are broken down by grade in the bottom twelve rows. The ESs for reading and mathematics are combined in this initial analysis to provide a single measure of overall effectiveness. Since some reviewers have noted greater gains for mathematics than verbal achievement (St. John, 1975; Krol, 1978), ESs for these two areas of achievement were also examined and are reported below.

Table 4

Glass Effect-Sizes For Each Grade Level

GRADE LEVEL AT POSTTEST	POOLED TOTAL OF "ACCEPTED" SAMPLE		GLASS EFFECT-SIZE X TYPE OF RESEARCH DESIGN					
			One Group Pretest-Posttest: 0 X 0		Nonequivalent Control Group: 0 X 0 - - - - - - - - - - 0 0		Static Group Comparison: X 0 - - - - - - - - - - 0	
	No. of Obs.[a]	Mean ES & (')	No. of Obs.	Mean ES & (')	No. of Obs.	Mean ES & (')	No. of Obs.	Mean ES & (')
1-6	74	0.43 (0.65)	8	1.75 (2.73)	46	0.28 (0.19)	16	0.24 (0.22)
7-9	11	*1.06(1.11)	4	1.99 (0.20)	4	*0.94(1.11)	3	-0.03 (0.23)
10-12	11	0.05 (0.04)	6	*0.01(0.05)	4	0.17 (0.01)	1	-0.18
GRAND	96	0.45[b] (0.60)	18	1.22[c] (1.96)	54	0.32 (0.26)	20	0.18 (0.20)
		F(2,95)=4.65, p < .02		F(2,17)=5.05, p < .03		F(2,53)=3.68. p < .04		F(2,19)=0.80, n.s.
1	2	-0.19 (0.01)	0	- -	1	-0.24	1	-0.14
2	10	0.17 (0.11)	1	0.01	5	0.09 (0.07)	2	0.08 (0.29)
3	8	0.39 (0.71)	1	2.15	5	0.28 (0.25)	0	- -
4	17	0.44 (0.54)	2	2.03 (1.20)	9	0.39 (0.10)	6	-0.03 (0.07)
5	22	0.51 (0.89)	3	1.54 (6.22)	16	0.38 (0.17)	3	0.17 (0.00)
6	15	0.56 (0.86)	1	3.15	10	0.18 (0.33)	4	*0.87(0.16)
7	4	*1.98(0.19)	2	2.18 (0.11)	2	*1.79(0.30)	0	- -
8	2	*1.80(0.34)	2	1.80 (0.34)	0	- -	0	- -
9	5	0.02 (0.07)	0	- -	2	0.10 (0.20)	3	-0.03 (0.02)
10	4	0.13 (0.03)	2	0.00 (0.29)	2	0.25 (0.01)	0	- -
11	4	0.12 (0.05)	2	0.15 (0.13)	2	0.09 (0.01)	0	- -
12	3	-0.15 (0.01)	2	-0.13 (0.00)	0	- -	1	-0.18
		F(11,95)=2.91, p < .005		F(9,17)=1.19, n.s.		F(9,53)=3.24, p < .01		F(6,19)=4.82, p < .01

*Significantly different from non-starred means within given column at beyond the .05 level by Scheffé test.

[a] Number of observations refers to the number of discrete cases present. Each study could furnish more than one case, since data were coded by grade level and type of posttest. There were 31 "accepted" studies, which yielded 106 observations (x = 3.42 observations per study).

[b] Overall, unweighted, mean effect-size. Weighting effect-size by size of sample within each study yields a mean effect-size of 0.42.

[c] Mean effect-size for one group pretest-posttest design is significantly greater than that for other designs at beyond the .0001 level by Schiffé test (overall F = 11.47, df=2,91, p 9< .0001).

The overall ES for the 31 studies is .45 standard deviations. The ES is relatively unaffected by various weighting schemes. This figure is considerably larger than those reported by Crain and Mahard (1982) and Krol (1978). However, the ESs for the more well-designed quasi-experiments are considerably smaller (i.e., .32 and .18). It is clear that the studies using the weaker OGPP design are inflating the estimate of the ES (i.e., 1.22). As was noted earlier, this latter design confounds maturation and initial differences in student selection with the effect of desegregation. Such design effects resulting from differences in study quality are commonly reported (cf. Wortman, 1983). In practically all such cases the weaker designs produce larger estimates of effects. Thus design quality must be considered in conducting an integrative review. As Jackson (1980) notes, "The results of the analysis may be misleading if there is not at least a modest number of studies with good overall design."

The bottom twelve rows of the table present the results by grade. The general pattern is for an increase in ES for grades 1-8 followed by a decline for the later grades. This finding contradicts those reported by Crain and Mahard (1978) and St. John (1975). The Glass ES for grades K-6 was slightly, but not statistically, lower than the ES for grades 7-12 (.43 and .55, respectively). Given the varying duration of these studies, Stephan (1982) calculated the ES per month for the NIE Core Studies. He found a pattern consistent with Crain and Mahard (1982) and St. John (1975).

All of these estimates of ES are susceptible to bias due to selection or absence of initial subject equivalence. The result for those studies where it was possible to employ the pretest adjustment to remove initial differences between segregated and desegregated groups are presented in Table 5. These studies used the non-equivalent control group design and reported sufficient pretest information to calculate ESs.

Table 5

Adjusted and unadjusted methods for the meta-analysis cf quasi-experiments

Computation Method	Overall Mean ES	Selection Problems[a]	No Selection Problems
Unadjusted	0.42 (n=32)	0.57 (n=20)	0.20 (n=10)
Pretest Adjusted	0.16 (n=32)	0.16 (n=20)	0.20 (n=10)
Pairwise t-value	$t_{62}=2.73$, $p < .02$	$t_{38}=2.94$, $p< .01$	$t_{18}=0$, n.s.

[a] In two cases it was not possible to determine whether or not there were selection problems.

The first column of the table indicates a sizeable and statistically significant difference between the "overall" unadjusted, Glass effect-size estimate and the pretest adjusted estimate (.42 and .16, respectively). The Glass estimate is similar to that reported above in Table 4. All studies were initially coded along a number of dimensions including most of Cook and Campbell's threats to validity before any effect sizes were actually calculated. The second and third columns compare studies with and without selection problems. The Glass ES estimate is higher for those studies with "selection problems" than the overall ES while the pretest-adjusted estimate remains the same as before (.57 and .16, respectively). Again, the two estimates are significantly different by statistical criteria. On the other hand, where selection was not considered a problem, the two estimates of ES are exactly the same (.20). This number is slightly higher for the pretest-adjusted estimates since two cases were omitted where it was not possible to determine a priori whether selection was a problem.

The difference between the pretest-adjusted ES and the ES for studies without selection problems may result from differential regression. Since the students involved in these studies generally score below the mean for their grade, their scores will regress to the higher mean at post-test solely due to the measurement error in the tests. Moreover, with an initial difference of .26 standard deviations, the control, segregated students will regress more. This implies that the pretest correction overadjusts slightly. Assuming a reliable test reliability of 0.8 to 0.9 for these students will account for the .04 difference.

The pretest-adjustment method thus appears to remove the initial differences due to subject nonequivalence. It is the author's opinion that this provides a fairly accurate estimate of the overall actual benefit of desegregation on minority, black achievement. According to Glass et al. (1981, p. 103), each .1 ES is equal to .1 grade equivalents or one month of educational gain. Thus desegregated students may be gaining about two months due to attending an integrated environment. The analysis indicates only a slight, but statistically non-significant, gain for the few cases where results greater than one school year were reported. Similarly, there were only a very few cases where the percentage black was reported. When the difference between percentage black in the control (i.e., segregated) and treatment (i.e., desegregated) groups was calculated, it revealed that most of the effects were obtained in those studies where the difference ranged from 76 to 85 percent. That is, students moving from almost completely segregated environments to predominantly white schools showed a sizeable (1.06 ES using the Glass method) effect. This finding is consistent with the Coleman Report.

Finally, the Glass effect size estimates for reading and mathematics were examined separately. These results are presented in Table 6. As with the overall ES, both effects are positive indicating a benefit for desegregated students. Contrary to previous research (Krol, 1978; St. John, 1975) the ES for reading achievement was considerably larger than that for math (.57 and .33, respectively). This difference was not statistically significant, however. Thus a single overall estimate of achievement effects appears to be an appropriate measure of the impact of desegregation.

Table 6

Mean Effect-Size for Math Vs. Reading Achievement Measures

Achievement Measure	Mean Glass ES & (σ^2)	*F*
Math (n=37)	0.33 (0.38)	1.86, df=1,87, *p* < .18
Reading (n=51)	0.57 (0.94)	

Note--Krol found a tendency for math achievement to show a greater effect-size than reading achievement (t_{16}=1.90, p=.08).

The NIE Core Studies

A similar analysis was performed on the 19 studies selected by the NIE panel of experts. The results are presented in Table 7. The information is presented by study with overall effects presented at the end. The pattern of results is quite similar to those presented above. All ESs are again positive indicating a beneficial impact of desegregation on achievement. The ESs are slightly lower partly due to the inclusion of the negative ESs for the Sheehan (1979) and Walberg (1971) studies.

The overall mean unadjusted Glass ES is .25. The unadjusted ES estimate is comparable to the .23 reported by Crain and Mahard (1982) and, more recently, the .24 by Crain (1983) for the best designed studies. It is only slightly less than the .28 ES that Crain and Mahard (1982) claim for "the estimated treatment assuming the best possible research design." However, all of those estimates ignore the bias introduced by the initial nonequivalence of the students. When adjusted for pretest differences, the ES is reduced to .14. Compared to the original 31 studies, the decrease for the Glass ES is .17, but it is only .02 for the pretest adjusted ES. The reason for this is that negative ESs have been added by the panel to the core studies which largely, but not entirely, reflect pre-existing differences among segregated and desegregated students. In these cases, however, the differences favored the segregated students. In fact, there is a large correlation between pretest and posttest effects sizes ($\underline{r}$ = .76) indicating that pre-existing differences largely remain at the posttest. Thus subject equivalence is a persistent source of bias in these studies. It is for this reason that the pretest adjustment method was employed. This adjusted ES provides a less biased estimate of the overall effectiveness of desegregation. The adjustment is equally successful for studies with large ESs (greater than 1.0) such as Rentsch (1967).

As with the larger set of 31 studies, the core studies show the effects for reading achievement to be modestly larger than those for mathematics (.28 and .23, respectively). However, when these figures are decomposed by duration or length of desegregation, there is an interaction with mathematics showing larger effects for those studies longer than one year. While there are relatively few cases available, this may explain the difference between the overall results in this study and those reported by others. It may be that studies of longer duration comprised the majority of those reviewed by Krol (1978) and St. John (1975).

Table 7. Effect Sizes for NIE Core Studies

Name of Study	# of Cases	% Black		Grade Level		Achievement Effect Size		Pretest-Adjusted Effect Size
		Seg.	Deseg.	Pretest	Posttest	Reading	Math	
Anderson (1966)	2	NA¹	NA	2	4	.63	--	.95
		NA	NA	2	4	--	.59	.53
Baker (1967)	4	NA	NA	2	2	.14	--	.23
		NA	NA	2	2	--	-.24	-.02
		NA	NA	3	3	1.02	--	-.04
		NA	NA	3	3	--	.55	.59
Bowman (1973)	2	99	16	3	5	.58	--	.02
		99	16	3	5	--	.07	-.06
Carrigan (1969)	6	50	5	K	1	-.24	--	-.41
		50	5	1	2	.34	--	-.02
		50	5	2	3	-.23	--	.30
		50	5	3	4	.00	--	-.13
		50	5	4	5	-.14	--	.33
		50	5	5	6	.52	--	-.31
Clark (1971)	2	95	NA	6	6	.08	--	--
		95	NA	6	6	--	-.25	--
Evans (1973)	6	NA	22	3	3	.02	--	--
		NA	22	3	3	--	.03	--
		NA	22	4	4	.02	--	--
		NA	22	4	4	--	.03	--
		NA	22	5	5	.02	--	--
		NA	22	5	5	--	.03	--
Iwanicki & Gable (1978)	3	NA	8	2	3	--	--	--
		NA	8	4	5	--	--	--
		NA	8	6	7	--	--	--
Klein (1967)	2	100	NA	10	10	.20	--	--
		100	NA	10	10	--	.30	--
Laird & Weeks (1966)	6	NA	NA	3	4	.58	--	--
		NA	NA	3	4	--	.46	--
		NA	NA	4	5	.81	--	--
		NA	NA	4	5	--	.48	--
		NA	NA	5	6	-.37	--	--
		NA	NA	5	6	--	-.45	--

Name of Study	# of Cases	% Black		Grade Level		Achievement Effect Size		Pretest-Adjusted Effect Size
		Seg.	Deseg.	Pretest	Posttest	Reading	Math	
Rentsch (1967)	6	90	5	3	5	1.14	--	.19
		90	5	3	5	--	.95	.06
		90	5	4	6	1.27	--	.58
		90	5	4	6	--	.92	-.17
		90	5	5	7	2.17	--	.76
		90	5	5	7	--	1.40	-.22
Savage (1971)	2	100	NA	9	11	.01	--	.14
		100	NA	9	11	--	.17	-.09
Sheehan (1979)	2	98	30	4	5	-.29	--	-.16
		98	30	4	5	--	-.27	-.16
Stone (1968)	2	60	NA	4	5	.42	--	--
		60	NA	4	5	--	.49	--
Smith (1971)	2	100	42	6	9	-.22	--	-.05
		100	42	6	9	--	.42	.10
Syracuse School District (1979)	1	89	10	4	4	.75	--	--
Thompson & Smidchens (1979)	2	42	5	3	5	-.33	--	--
		42	5	3	5	--	.10	--
Van Every (1969)	6	95	20	4	5	.78	--	.59
		95	20	4	5	--	.28	.11
		95	20	4	5	--	--	--
		95	20	4	6	-.25	--	-.44
		95	20	4	6	--	.36	.53
		95	20	4	6	--	--	--
Walberg (1971)	4	NA	NA	NA	NA	-.13	-.29	.11
		NA	NA	NA	NA	.11	-.28	-.24
		NA	NA	NA	NA	.16	.36	.21
		NA	NA	NA	NA	.29	-.06	-.01
Zdep (1971)	2	NA	12	2	2	.34	--	.65
		NA	12	2	2	--	-.15	-.15
• 19	• 62							

Name of Study	# of Cases	% Black		Grade Level		Achievement Effect Size		Protest-Adjusted Effect Size
		Seg.	Deseg.	Pretest	Posttest	Reading	Math	
OVERALL MEAN'	(N= 62)	82.49	15.03	4.05	5.12	.28	.23	.14
MEAN FOR TREATMENTS LASTING ONE YEAR OR LESS'	(N= 20)	71.00	11.58	3.65	4.20	.30	.11	.13
MEAN FOR TREATMENTS LASTING MORE THAN ONE YEAR'	(N= 14)	95.31	17.90	4.00	5.81	.28	.39	.12

Note: 'NA = Not Ascertainable

'Mean effect sizes, weighted by study

5.0 SUMMARY

The synthesis of scientific research using formal statistical procedures such as Glass' meta-analysis presents special problems when studies are methodologically flawed. The research literature on the effectiveness of school desegregation on minority black achievement is almost totally comprised of quasi-experiments or weaker research designs. While Glass has recommended including all studies in a research synthesis, his work has largely dealt with studies that are "well designed." In those instances where "poorly designed" studies have been included, design effects have been found (Glass & Smith, 1979; Gilbert et al., 1977; Wortman, 1981) indicating major differences in estimates of effects between studies with strong and weak designs. The typical approach to this problem is to examine the higher-quality studies taking into account, where possible, the flaws or threats to validity. This was the approach taken in this study. Specific methodological criteria for including studies in the research synthesis were developed and applied to the school desegregation literature. All studies were found to have some serious flaws, but 31 were considered acceptable for analysis. Even within this set, there was variation in design quality and a considerable design effect. The NIE panel of experts decided to include only the highest quality studies and this further reduced the set to 18 studies. The study by Walberg (1971) was felt to be of sufficient quality to be added to this set although it had originally been "rejected" for a variety of methodological flaws.

The NIE Core Studies had an overall effect size of .25 standard deviations. This is almost identical to the effect size estimate reported by Crain and his associates for well-designed studies. Since most of these studies suffered from initial subject nonequivalence, an adjusted effect size was calculated by subtracting out the effect size at the pretest prior to desegregation. This resulted in an effect size of .14. Given differential statistical regression to the mean, this is probably a slight underestimate. This is similar to that found for the larger set of 31 studies and also to Krol's (1978) finding. In examining the results of the two analyses reported above, the best overall estimate of the effect of school desegregtion on black achievement appears to be about .2 of a standard deviation. This estimate is based on those cases not having selection problems and is comparable to the adjusted estimates.

Other subsidiary analyses comparing type of achievement, duration of desegregation, grade level, and difference in percent black for segregated and desegregated students were also examined. Reading was found to be slightly higher than math achievement although this may vary with length of desegregation. The larger set of studies revealed a curvilinear pattern of effects with an increase from grades K-7 and a decrease from 8-12. This result does not agree with other findings indicating larger benefits the earlier desegregation occurs. No effect was found for amount of desegregation (i.e., less than one year compared to more than one year). Some support was found for the finding of the Coleman Report that effects are greatest in the most integrated environments.

What do these findings mean? The effect size found in both analyses reported here indicates about a two month gain or benefit for desegregated students. The meaning attached to this finding represents a judgment. This is where social science ends and social policy begins.

However, we have examined the scientific literature on coronary-artery bypass graft surgery for comparative purposes. This is a widely accepted medical procedure that is currently performed on well over 100,000 persons annually at a cost of nearly $2 billion. Much of this expense is reimbursed by third-party payers including the federal government. A research synthesis of the higher-quality studies (i.e. randomized) found a benefit of .8 standard deviations representing only a 4.4 percent increase in survival rates (Wortman & Yeaton, in press). This is a modest increase at a considerable social cost when compared to school desegregation. Moreover, programs aimed at the young such as school desegregation typically are more cost effective than those for the elderly such as bypass surgery.

Although the methods developed above have been useful in dealing with problems of student equivalence, they cannot adjust for the second major problem noted by St. John (1975) of "equivalence of schools." The actual details of the educational programs involved in the desegregation studies are not reported. Thus it is not possible to determine effective from ineffective programs. The real problem as Gerard and Miller (1975) conclude is "to foster integration of the minority children into the classroom social structure and academic program." Recent studies have addressed this issue and developed procedures for improving educational practice in desegregated classrooms (Aronson & Bridgeman, 1979; Slavin & Madden, 1979). A number of the papers by members of the NIE expert panel focused on these procedures. Such research based on sound social science theory is likely to lead to increased educational benefits for oesegregated students.

The political reality confronting the achievement of school desegregation today is the need to allow students in highly segregated urban inner cities access to schools in the surrounding white collar suburbs. Such "metropolitan plans" have been found to achieve desegregtion without white flight. They are also quite controversial and typically require cross-district busing. The results in St. Louis are encouraging. Here voluntary cross-district busing combined with inner city magnet schools have produced two-way desegregation with some whites returning to the city schools. It should be noted that the plan is an alternative to court-ordered mandatory metropolitan desegregation. Moreover, it should be added that such plans resemble the early voluntary plans in the Northeast. As a social policy, these plans -- capitalizing on good suburban schools, a cooperative environment, and motivated volunteers -- produced the largest effects of the studies examined.

6.0 FOOTNOTES

[1] Cohen's estimate of effect size, d, is nearly identical. The denominator includes information from both treatment and control groups, the pooled-within standard deviation. Hedges (1982) maintains that this produces a less biased estimate of effect. However, this estimator ignores problems caused by the effect of the treatment on the experimental (i.e., desegregated) group standard deviation.

[2] Unfortunately, it was not possible to calculate effect sizes from this study either since standard deviations were not reported. Similar problems plague the earlier reports as well.

[3] In fact, one of the "neutral" members had testified numerous times against desegregation in court cases.

7.0 REFERENCES

Aronson, E., & Bridgeman, D. Jigsaw groups and the desegregated classrooms: In pursuit of common goals. *Personality and Social Psychology Bulletin*, 1979, *54*, 438-446.

Boruch, R. F., & Gomez, H. Sensitivity, bias, and theory in impact evaluations. *Professional Psychology*, 1977, *8*, 411-434.

Bradley, L.A. & Bradley, G.W. The academic achievement of black students in desegregated schools: A critical review. *Review of Educational Research*, 1977, *47*, 399-449.

Bryant, F.B. & Wortman, P.M. Secondary analysis: The case for data archives. *American Psychologist*, 1978, *33*, 381-37.

Campbell, D. T. Temporal changes in treatment-effect correlations: A quasi-experimental model for institutional records and longitudinal studies. In G. V. Glass (Ed.), *Proceedings of the 1970 Invitational Conference on Testing Problems: The Promise and Perils of Educational Information Systems*. New York: Educational Testing Service, 1971.

Campbell, D. T., & Boruch, R. F. Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations tend to underestimate effects. In C. A. Bennett & A. A. Lumsdaine (Eds.), *Evaluation and Experiment: Some Critical Issues in Assessing Social Programs*. New York: Academic Press, 1975.

Campbell, D. T., & Erlebacher, A. E. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), *Compensatory Education: A National Debate* (Vol. 3). *Disadvantaged Child*. New York: Brunner/Mazel, 1970.

Campbell, D. T., & Stanley, J. C. *Experimental and Quasi-experimental Designs for Research*. Chicago: Rand McNally, 1966.

Cohen, J. *Statistical Power for the Behavioral Sciences*. New York: Academic Press, 1969.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D. & York, R.L. *Equality of educational opportunity*. Washington, D.C.: U.S. Government Printing Office, 1966.

Cook, T. D., & Campbell, D. T. *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Chicago: Houghton Mifflin, 1979.

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

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

Crain, R. L. Is nineteen really better than ninety-three? (Technical Report). Washington, D. C.: National Institute of Education, 1983 (forthcoming).

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

Crain, R.L. & Mahard, R.E. Desegregation plans that raise black achievement: A review of the research. Santa Monica, CA: The Rand Corporation (N-1844-NIE), June 1982.

Director, S. M. Underadjustment bias in the evaluation of manpower training. Evaluation Quarterly, 1979, 3, 190-218.

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

Gehan, E. A., & Freireich, E. J. Non-randomized controls in cancer clinical trials. The New England Journal of Medicine, 1974, 290, 198-203.

Gerard, H. B., & Miller, N. (eds.). School desegregation. New York: Plenum, 1975, 1975.

Gilbert, J.P. McPeek, B., & Mosteller, F. Progress in surgery and anesthesia: Benefits and risks of innovative therapy. In J. P. Bunker, B. A. Barnes, & F. Mosteller (Eds.), Costs, risks, and benefits of surgery. New York: Oxford, 1977.

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

Glass, G. V. Integrating findings: The meta-analysis of research. In L. S. Shulman (Ed.), Review of Research in Education, Vol. 5. Itasca, Ill.: Peacock, 1977. Pp. 351-379.

Glass, G.V. Reply to Mansfield and Busse. Educational Research, 1978, 7, 3.

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

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

Grant, G. Shaping social policy: The politics of the Coleman Report. Teachers College Record, 1975, 75, 17-54.

Hedges, L. V. Estimation of effect size from a series of independent experiments. Psychological Bulletin, 1982, 92, 490-499.

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

Kenny, D. A. A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. Psychological Bulletin, 1975, 82, 345-362.

Kluger, R. Simple justice. New York: Random House, 1975.

Krol, R. A. A meta analysis of comparative research on the effects of desegregation on academic achievement. Unpublished dissertation, 1978. Ann Arbor, Mich.: University Microfilms (#7907962), 1979.

Landman, J. T. & Dawes, R. M. Psychotherapy outcome: Smith and Glass' conclusions stand up under scrutiny. American Psychologist, 1982, 37, 504-516.

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

Linsenmeier, J. A. W., & Wortman, P. M. The Riverside School Study of desegregation: A re-examination. Research Review of Equal Education, 1978, 2(2), 1-40.

Mansfield, R.S. & Busse, T.V. Meta-analysis of research: A rejoinder to Glass. Educational Research, 1979, 6, 3.

Moskowitz, J. M., & Wortman, P. M. A secondary analysis of the Riverside School Study of desegregation. In R. F. Boruch, P. M. Wortman, & D. S. Cordray (Eds.), Secondary Analysis in Applied Social Research. San Francisco: Jossey-Bass, 1981.

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

Sacks, H., Chalmers, T.C. & Smith, H. Randomized versus historical controls for clinical trials. American Journal of Medicine, 1982, 72, 233-240.

Sechrest, L., & Yeaton, W. Empirical bases for estimating effect size. In R. F. Boruch, P. M. Wortman, & D. S. Cordray (Eds.), Secondary Analysis in Applied Social Research. San Francisco: Jossey-Bass, 1981.

Slavin, R. E., & Madden, N.A. School practices that improve race relations. American Educational Research Journal, 1979, 16, 169-180.

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

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

Staines, G. L. The strategic combination argument. In W. Leinfellner & E. Kohler (Eds.), Developments in the Methodology of Social Science. Dordecht, Holland: Reidel, 1974.

Stephan, W. G. Blacks and Brown: The effects of school desegregation on Black students. (Technical report). Washington, D.C.: National Institute of Education, 1982.

St. John, N.H. School desegregation outcomes for children. New York: John Wiley & Sons, 1975.

Teele, J. E. Evaluating School Busing: A Case Study of Boston's Operation Exodus. New York: Praeger, 1973.

*Walberg, H.J. An evaluation of an urban-suburban school bussing program: Student achievement and perception of class learning environments. Paper presented at the annual meeting of the American Educational Research Association, New York: 1971.

Weinberg, M. Minority students: A research appraisal. Washington, D.C.: U.S. DHEW, National Institute of Education, 1977.

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

Wortman, P. M. Randomized clinical trials. In P. M. Wortman (Ed.), Methods for revaluating health services. Beverly Hill, CA: Sage, 1981.

Wortman, P.M., King, C. & Bryant, F.B. Meta-analysis of quasi-experiments: School desegregation and black achievement. Part I - Retrieval and coding. Ann Arbor, MI: Institute for Social Research, 1982.

Wortman, P. M., Reichardt, C. S., & St. Pierre, R. G. The first year of the Education Voucher Demonstration: A secondary analysis of student achievement test scores. Evaluation Quarterly, 1978, 2, 193-214.

Wortman, P. M., & Yeaton, W. H. Synthesis of results in controlled trials of coronary artery bypass surgery. In R. J. L:gnt (Ed.), Evaluation Studies Review Annual, Volume 8. Beverly Hills, CA : Sage, in press.

Appendix A

Bibliography of Accepted Studies

Aberdeen, Frank D. *Adjustment to desegregation: A description of some differences among Negro elementary school pupils*. Unpublished doctoral dissertation, University of Michigan, 1969.

*Anderson, Louis V. *The effect of desegregation on the achievement and personality patterns of negro children*. Unpublished doctoral dissertation, George Peabody College for Teachers, 1966. (University Microfilm 66-11, 237)

*Beker, Jerome. A study of integration in racially imbalanced urban public school. Syracuse, New York: Syracuse University Youth Development Center, *Final Report*, May 1967.

*Bowman, Orrin H. *Scholastic development of disadvantaged Negro pupils: A study of pupils in selected segregated and desegregated elementary classrooms*. Unpublished doctoral dissertation, University of New York at Buffalo, 1973.

Bryant, James C. *Some effect of racial integration of high school students on standardized achievement test scores: Teacher grades and drop-out rates in Angleton, Texas*. Unpublished doctoral dissertation, University of Houston, 1968.

*Carrigan, Patricia M. *School desegregation via compulsory pupil transfer: Early effects on elementary school children*. Ann Arbor, Michigan: Ann Arbor Public Schools, 1969.

Clark County School District. *Desegregation Report*. Las Vegas, Nevada: Clark County School District, 1975. (ERIC No. ED 106 397)

*Clark, El Nadel. *Analysis of the differences between pre- and posttest scores (change scores) on measures of self-concept, academic aptitude, and reading achievement earned by sixth grade students attending segregated and desegregated schools*. Unpublished doctoral dissertation, Duke University, 1971.

Clinton, Ronald R. *A study of the improvement in achievement of basic skills of children bused from urban to suburban school environments*. Unpublished masters thesis, South Connecticut State College, 1969.

*Evans, Charles L. *Integration evaluation: Desegregation study II -- academic effects on bused black and receiving white students, 1972-73*. Fort Worth, Texas: Fort Worth Independent School District, 1973. (ERIC No. ED 094 087)

Hampton, C. *The effects of desegregation on the scholastic achievement of relatively advantaged Negro children*. Unpublished doctoral dissertation. University of Southern California, Los Angeles, California, 1970.

Hsia, Jayjia. Integration in Evanston, 1967-1971. Princeton, New Jersey: Educational Testing Service, 1971. (ERIC No. ED 054 292, UD 011 812)

*Iwanicki, E. F., & Gable, R. K. A quasi-experimental evaluation of the effects of a voluntary urban/suburban busing program on student achievement. Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, Canada, March 1978.

*Klein, Robert Stanley. A comparative study of the academic achievement of negro tenth grade high school students attending segregated and recently integrated schools in a metropolitan area in the south. Unpublished doctoral dissertation, University of South Carolina, 1967.

*Laird, M. A., & Weeks, G. The effect of busing on achievement in reading and arithmetic in three Philadelphia schools. Philadelphia, Pennsylvania: The School District of Philadelphia, Division of Research, 1966.

Laurent, James A. Effects of race and racial balance of school on academic performance. Unpublished doctoral dissertation, University of Oregon, 1969. (ERIC No. ED 048 393 UD 011 305)

Levy, Marilyn. A study of Project Concern in Cheshire, Connecticut: September, 1968 through June, 1970. Cheshire, Connecticut: Department of Education, 1970.

Lockwood, Jane D. An examination of scholastic achievement, attitudes and home background factors of 6th grade negro students in balanced and unbalanced schools. Unpublished doctoral dissertation, University of Michigan, 1966.

Moreno, Marguerite C. The effect of integration on the aptitude, achievement, attitudes to school and class, and social acceptance of negro and white pupils in a small urban school system. Unpublished doctoral dissertation, Fordham University, 1971.

*Rentsch, George J. Open-enrollment: An appraisal. Unpublished doctoral dissertation, State University of New York, Buffalo, 1967.

Rock, William C., et al. A report on a cooperative program between a city school district and a suburban school district. Rochester, New York: , 1968.

Samuels, Joseph M. A comparison of projects representative of compensatory: busing: and non-compensatory programs for inner-city students. Unpublished doctoral dissertation. University of Connecticut, 1971.

*Savage, L. W. Academic achievement of black students transferring from a segregated junior high school to an integrated high school. Unpublished masters thesis, Virginia State College, 1971.

*Sheehan, Daniel S. Black achievement in a desegregated school district. Journal of Social Psychology, 1979, 107, 185-192.

*Slone, Irene W. The effects of one school pairing on pupil achievement, anxieties and attitudes. Unpublished doctoral dissertation, New York University, 1968.

*Smith, Lee Rand. A comparative study of the achievement of negro students attending segregated junior high schools and negro students attending desegregated junior high schools in the city of Tulsa. Unpublished doctoral dissertation, University of Tulsa, 1971.

*Syracuse City School District. Study of the effect of integration -- Washington Irving and Host pupils. Hearing held in Rochester, New York, September 16-17, U. S. Commission on Civil Rights, 1966, pp. 323-236.

*Thompson, E. W., & Smidchens, U. Longitudinal effects of school racial/ethnic composition upon student achievement. Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, California, April, 1979).

*Van Every, D. F. Effect of desegregation on public school groups of sixth graders in terms of achievement levels and attitudes toward school. Doctoral dissertation, Wayne State University, 1969. Dissertation Abstracts International, 1969. (University Microfilms No. 70-19074)

Williams, Frank E. An analysis of some differences between negro high school seniors from a segregated high school and a non-segregated high school in Brevard County, Florida. Unpublished doctoral dissertation, University of Florida, 1968.

*Zdep, Stanley M. Educating disadvantaged urban children in suburban schools: An evaluation. Journal of Applied Social Psychology, 1971, 1. (ERIC No. ED 053 186 TM 00716)

*Article included in NIE Core Studies.