Meta-analytic techniques were used to explore overall conclusions and variables moderating treatment effects in the research literature on school desegregation and black achievement. Studies were classified on the basis of the threats to their validity as either accepted or rejected for the analysis. For the initial analysis quasi-experimental studies were accepted, yielding an average effect size of .45. The better-designed studies had an average effect size of .34, which was reduced to .16 when adjusted for pretest differences. The National Institute of Education (NIE) convened an expert panel that reviewed and reanalyzed these results. An average pretest-adjusted effect size of .14 was found for the 19 studies selected for analysis by the NIE panel. An average effect size of .20 was found for the better-designed studies that had no selection problems. This is equivalent to two months of educational gain. The largest effects occurred among students moving from highly segregated to predominantly white schools. Reading achievement gains were larger than those for mathematics, but the difference was not statistically significant.

School Desegregation and Black Achievement

An Integrative Review

PAUL M. WORTMAN
University of Michigan
FRED B. BRYANT
Loyola University of Chicago

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 figured prominently in the controversies that have constantly surrounded major events in race relations 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 based in part on current social science

AUTHORS' NOTE: *This article is based on a report submitted to the National Institute of Education Panel on the Effects of School Desegration. The work was*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 13 No. 3, February 1985 289-324
© 1985 Sage Publications, Inc.

evidence. More recently, the so-called Coleman Report or the Equality of Educational Opportunity Survey (Coleman et al., 1966) was used by the Johnson administration to accelerate the desegregation process (Grant, 1975). This major survey found that black student achievement increased in more integrated environments (i.e., when there was a greater proportion of white students).

The Coleman Report's findings led not only to a number of reanalyses by social scientists but also to an increasing number of systematic studies using before and after measurements (i.e., pretests and posttests) of achievement and 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 as well as those of other social scientists. By the mid-1970s a sufficient body of scientific studies accumulated, permitting several careful reviews.

Two of the most notable of these literature 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; St. John reviewed 64 studies, including 12 that were cross-sectional. Both found the findings inconclusive. The Bradleys concluded that evidence on the effectiveness of desegregation on black achievement was "inconsistent and inadequate." St. John quoted Light and Smith (1971), saying 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's seminal work on meta-analysis (1976, 1977).

Meta-analysis is a quantitative procedure for determining the average effect size of a hypothesis tested in many individual

supported by funds from the National Institute of Education (NIE-G-79-0128 and NIE-P-82-0070) awarded to the senior author. We gratefully acknowledge the assistance of Craig King. The views expressed in this article are our own and do not necessarily represent the views of the National Institute of Education.

studies. It offers numerous advantages over previous methods for aggregating the findings of independent studies (Light and Smith, 1971; Glass, 1977). Its major advantage is that it provides a single, precise, quantitative measurement of the average magnitude of program impact (typically in standard deviation units). Meta-analysis is applicable to most social science research and provides an important result that is easy to grasp. It also allows one to consider sample size and design quality. This technique, however, 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 been applied to the school desegregation literature by Crain and Mahard (1982) and Krol (1979). Their meta-analyses found small positive benefits for desegregation on black achievement (.16 and .08 standard deviations, respectively). Both studies are flawed, however. Krol's study illustrates the inappropriate application of Glass's method. For example, Glass (1977: 356) recommends using preexperimental designs lacking controls "*if* the treated group members' posttreatment status is a good estimate of their hypothetical posttreatment status in the absence of treatment." In the next section it will be shown that this suggestion may be unwarranted and ill-advised.

In a more recent meta-analysis Crain and Mahard (1982) took a traditional Glassian approach and included all available studies in their analysis. This approach is also inappropriate. Numerous desegregation 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 (as they lack both a predesegregation or pretest measure as well as a control group), but they were nevertheless 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 later. Such studies represent half of those included in Crain and Mahard's meta-analysis. Most important, however, Krol as well as Crain and Mahard paid insufficient attention to the threats to validity that could confound and bias the results of their meta-analyses.

The literature on school desegregation and achievement poses some special problems for the meta-analysis method. This literature is almost entirely quasi-experimental in composition and thus is 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. This latter term refers to 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. Although Glass was initially confident that his method could be used with quasi-experiments, his views gradually changed (see Glass and Smith, 1981). Examination of the quasi-experimental desegregation studies presented in the following sections indicates that selection is a persistent plausible rival hypothesis. That is, it is not canceled out. Therefore, a number of steps were 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.

This article focuses on the effect of school desegregation on black achievement. Although interest in these data is primarily methodological and stems from earlier work by the senior author on the secondary analysis of the Riverside School Study of desegregation (Linsenmeier and Wortman, 1978; Moskowitz and 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).

The next section describes the meta-analytic method used in this study. 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 last three sections present the results, an examination of the utility of meta-analysis for social policy, and a brief summary of the findings.

METHODOLOGY

To apply meta-analysis to quasi-experimental data one needs to obtain a measure of effect size (ES). The basic equation adapted from Cohen (1969) and Glass (1977) and extended to quasi-experiment is as follows:

$$ES = \frac{\left(\bar{X}_{E_2} - \bar{X}_{C_2}\right)}{S_{C_2}} - \frac{\left(\bar{X}_{E_1} - \bar{X}_{C_1}\right)}{S_{C_1}} \qquad [1]$$

where $\bar{X}_{Ei}$, $\bar{X}_{ci}$ = means for the treatment (i.e., desegregated) or experimental (E) groups and control (C) or untreated (i.e., segregate) groups, S_{Ci} = the standard deviation of the control group,[1] and i = 1,2 indicates time 1 (pretest) and time 2 (posttest). In a randomized or "true" experiment $X_{E_1} = X_{c_1}$, yielding the Glass ES equation. However, in a quasi-experimental situation it is likely that the groups will differ initially so that the Glass procedure would produce a biased estimate. That is, selection bias is a major threat to validity in this model. Thus, equation 1 provides a pretest adjustment to remove selection bias resulting from initial subject nonequivalence.

Meta-analysis involves summing the effect size estimates from all studies and dividing this total by the number of studies. The average effect size, Δ, is usually presented. This average can be computed several ways. For example, all ESs can be summed and averaged. Given that many ESs may be derived from a single study, this introduces bias from nonindependent measurements. It was largely for this reason that Landman and Dawes (1982) reanalyzed Smith and Glass's (1977) meta-analysis of the effectiveness of psychotherapy using the average ES from each study as the unit of analysis.

The desegregation literature is largely composed of quasi-experiments or even more poorly designed studies. As a consequence, it is susceptible to a variety of threats to internal validity (i.e., the ability to infer causality), such as selection bias noted above. It is risky to assume that these potential sources of bias can be treated as random errors that are self-canceling.

Matching was rarely used in the studies of desegregation, so the pretest adjustment procedure described in equation 1 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. To check the adequacy of this procedure, 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. Neither Crain and Mahard (1982) nor Krol (1979) attempted to correct or adjust for bias introduced by initial subject nonequivalence in their meta-analyses.

PRACTICAL LIMITATIONS

There are a number of problems in translating this small analytic model into an actual meta-analysis. First, the nonequivalent control group design (NECGD) used in most desegregation studies requires 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 nonsignificant results were obtained. The reliability of the tests is even less likely to be reported. In order to deal with this situation, a variety of indirect approaches have been proposed (see Glass, 1977) for converting reports of inferential statistics into ESs (also see Rosenthal, 1978).

Another common form of reporting results is the gain score. This is the change in each group from pretest to posttest. A simple algebraic manipulation reveals that the difference in the two gain scores is equivalent to the numerator in the basic equation, and can be used to estimate the effect size for quasi-experiments (equation 1). Thus if $S_1 = S_2$, gain scores can be used to derive ES for the NECGD quasi-experiment.

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 Stanley's (1966) 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 and one that is likely to lead to an overestimate of ES. Use of the standardized gain score creates a pseudo-effect equal to the control group gain. Moreover, if strict selection criteria are used, as they often are in compensatory education or competency testing remediation programs (so that only those with the lowest test scores are eligible), then regression effects will also be incorrectly included. Thus such case study data should be used only 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 in need, for example. As a consequence, researchers often attempt to solve the control group problem by using historical controls, or "cohort comparisons," according to Crain and Mahard (1982). In fact, this procedure has been recommended in some areas (see Gehan and Freireich, 1974). In education studies, historical control groups are often created using student data from the same grades during prior years (i.e., before the program innovation). This adds history to the list of possible threats to validity, given that these data are not obtained concurrently with the experimental (i.e., desegregation) data.

In general, historical controls have been found to overestimate treatment effects grossly and thus should be avoided if possible (Sacks et al., 1982). In education studies, for example, test scores were declining during the 1960s and 1970s so that earlier historical controls probably had higher scores. Such studies were not included in our analyses, but they comprised 17% of the studies in Crain and Mahard's (1982) meta-analysis. More recently, Crain (1983) included 8 such studies among his 20 best.

True Experiments

Although the preceding discussion focused on quasi-experiments, "true" or randomized studies are very useful. They provide unbiased estimates of the effect of desegregation on black achievement and indicate the bias resulting from quasi-experimental designs. Especially in studies of education, there has been a strong tendency for applied, field problems to be approached quasi-experimentally; laboratory and theoretical issues have been investigated using randomized studies. Few randomized studies have been conducted in the school desegregation area. Those that have been conducted, such as Project Concern (Iwanicki and Gable, 1978), often report results in a way that makes it impossible to derive effect-size estimates.

Crain (1983) identified 5 randomized studies among his top 20, 3 of which were based on data from Project Concern. Of these randomized studies (Rock et al., 1968; Samuels, 1971; Zdep, 1971) 3 were included among the 31 found acceptable in the present analysis. A more recent report from Project Concern (Iwanicki and Gable, 1978) was included in place of the two earlier reports used by Crain.[2]

DESIGN QUALITY

Although most school desegregation studies use the NECGD, the quality of the studies using this design varies. For example, some do not use standardized achievement test, some do not report pretest scores, and some do not separate student performance by grade level. All of these practices lower the quality of the data and limit their usefulness for meta-analysis. Moreover, as noted above, other quasi-experimental designs are often employed. A number of approaches to assessing quality have been developed. The best known is the validity approach developed by Campbell and Stanley (1966) and further refined by Cook and Campbell (1979). Essentially, the threats to validity indicate quality. Others (Boruch and Gomez, 1977; Sechrest and Yeaton, 1981) have stressed the implementation or integrity of the treatment. This is an important concept, though one that is difficult to measure.

The assessment of research quality is a new area that is critical in the synthesis of scientific studies. There has been much discussion of this issue (Mansfield and Busse, 1979; Eysenck, 1978; Glass, 1977, 1978) and the debate still continues (see Wortman, 1983). As the procedure section indicates, design quality is viewed as significant in selecting, coding, and analyzing the data in a research synthesis.

PROCEDURE

The meta-analysis approach first requires the retrieval of relevant scientific information (Glass et al., 1981). The importance of a thoroughly documented procedure at this point has been stressed by both Cooper (1982) and Jackson (1980). To that end, the cooperation of the authors of the two major studies systematically synthesizing the literature on the effects of school desegregation on black achievement (Crain and Mahard, 1978; Krol, 1979) was obtained. Both Robert Crain and Ronald Krol generously provided copies of the articles and the coding schemes used in their analyses. This data base was then extended and updated via literature searches including ERIC, dissertation abstracts, references in articles and books (especially St. John, 1975), and dozens of letters to authors and school district offices. A coding scheme was developed along with a list of studies to be included in the analyses. These are described below.

As the initial coding effort progressed, it became apparent that many studies would have to be rejected. It is imperative to describe these studies and the reasons for rejecting them from the analysis for two reasons: (a) Selection 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 another report (Bryant and Wortman, forthcoming). Both of these concerns are discussed in the next three sections.

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 factors for inclusion or exclusion. The observed threats to validity fall into one or more of four basic classifications developed by Campbell and his associates (Campbell and Stanley, 1966; Cook and Campbell, 1979). Thus the criteria used to reject studies (see Table 1) represent specific instances or threats to internal, external, construct, or statistical conclusion validity. In general, threats to construct or external validity were used to determine the relevance of a study for the meta-analysis. Threats to internal and statistical conclusion validity were used to determine the acceptability of a relevant study in the analysis (see Bryant and Wortman, forthcoming).

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, including those that at first appear to measure academic achievement of desegregated blacks but that, in fact, measure a different construct such as IQ (an ability measure); those that measure a different treatment, such as bus transportation; or those that use a different population, such as

TABLE 1
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) Heterogenous 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

[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.

*Criteria used to select NIE Core Studies.

whites or Chicanos (see criterion 3a, Table 1). Only studies using black controls were examined. This is the comparison recommended by St. John (1975) and should reduce or eliminate differential growth[3] of intellectual skills or "maturation" as a threat to validity. 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, however, both Crain and Mahard (1982) and Krol (1979) included studies that used white controls.

External validity refers to limitations in the generalizability of the study with regard to populations and settings, as well as treatment and measurement variables. One obvious reason for exclusion occurred if studies were conducted outside 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, Table 1). These may disguise or distort findings indicating how desegregation affects achievement. Moreover, when the dates of test administration are not described (criterion 5c, Table 1), problems arise in adjusting the effect size estimates to a proper time interval as well as in determining whether the pretest actually occurred prior to desegregation.

Internal validity is broadly concerned with whether the treatment (school desegregation) in fact affected the outcome (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 (1966). 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, Table 1) 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, Table 1) and thus made 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" problems stemming from differential test reliability and low intertest 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.

Statistical conclusion validity is concerned with the appropriateness of the statistical analyses used. 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 nonequivalent control group design (i.e., criterion 6h, Table 1) 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 that 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, thus making grade-equivalent scores actually noncomparable (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 1 resulted in the exclusion of 79 studies. Most suffered from more than one problem. A number of these criteria are sufficient in themselves to eliminate a study (i.e., are fatal flaws). All but three studies had such flaws. Overall, the majority of studies examined were excluded, including a number used in the previous meta-analyses (Crain and Mahard, 1978; Krol, 1979). A comparison of studies included and excluded is provided in Table 2. With the exception of Crain and Mahard (1978), only about half of the studies used in other major reviews were included. The 31 studies included in the analyses are listed in Appendix A. The studies were classified into effect size

TABLE 2
Comparison with Previous Research Syntheses

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

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.

A considerable amount of effort was spent 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. Although 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 (see Jackson, 1980).

One disadvantage of meta-analysis 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 to be acceptable contained only two published articles. Desegregation research is largely (and perhaps appropriately) a fugitive literature. The retrieval strategy described above probably located the target population of studies (Cooper, 1982). Moreover, use of the voting method (Light and Smith, 1971) to capture the general results of rejected studies (information to calculate effect sizes was lacking) revealed no statistically significant differences between accepted and rejected studies on math and verbal achievement (Bryant and Wortman, forthcoming).

THE NIE CORE STUDIES

After this screening process had been performed and the 31 resulting studies analyzed, the National Institute of Education (NIE) convened an expert panel to select the best studies in this area. The panel of six scholars (including the senior author) was supposedly balanced in both their attitudes and published work on desegregation—two pro, two con, and two neutral.[4] 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 1 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 Table 1.

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 because it suffered from an extremely high rate of attrition (criterion 3h, Table 1) that differed for segregated and desegregated students (i.e., 27% and 48%, 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, Table 1). The Walberg study added eight cases to the data base. Moreover, one of the "con" panelists wrote to the author of another study claiming negative findings (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; St. John (1975) reviewed only nine of them.

RESULTS

The Glass effect sizes (ESs) for 31 studies considered methodologically acceptable for performing a meta-analysis are presented in Table 3. The fourth row, labeled "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. Some reviewers have noted greater gains for mathematics than for verbal achievement (St. John, 1975; Krol, 1979); ESs for these two areas of achievement were also examined and are reported below.

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 (1979). However, the ESs for the better-designed quasi-experiments are considerably smaller (i.e., .32 and .18). The Hedges (1982) correction for bias was nearly identical to the NECGD studies, yielding an ES of .34. This is not surprising, given that it requires pooled pretest data from both segregated (control) and desegregated (treated) groups.

It is clear that the studies using the weaker OGPP design inflate 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 (see 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) noted, "The results of the analysis may be

TABLE 3
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: O X O		Nonequivalent Control Group: O X O - - - - - - - - - O O		Static Group Comparison: X O - - - - - - - - - O	
	No. of Obs.[1]	Mean ES & (σ^2)	No. of Obs.	Mean ES & (σ^2)	No. of Obs.	Mean ES & (σ^2)	No. of Obs.	Mean ES & (σ^2)
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[2] (0.68)	18	1.22[3] (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

NOTE: Plus sign indicates significantly different from nonstarred means within given column at beyond the .05 level by Scheffe test.

1. Number of observations refers to the number of discrete cases present. Each study could furnish more than one case, as data were coded by grade level and type of posttest. There were 31 "accepted" studies, yielding 106 observation ($\overline{X}$ = 3.42 observations per study).

2. Overall, unweighted, mean effect size. Weighting effect size by size of sample within each study yields a mean effect size of 0.42.

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

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).[5] The Glass ES for grades 1-6 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 results 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 4. These studies used the nonequivalent control group design and reported sufficient pretest information to calculate ESs.

The first column of the table indicates a sizeable and statistically significant difference between the overall unadjusted Glass ES estimate and the pretest-adjusted estimate (.42 and .16, respectively). The Glass estimate is similar to that reported 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 for the overall ES; 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).

The difference between the pretest-adjusted ES and the ES for studies without selection problems may result from differential regression. Given that the students involved in these studies generally score below the mean for their grade, their scores will

TABLE 4
Adjusted and Unadjusted Methods for the Meta-Analysis of 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.

regress to the higher mean at posttest solely due to 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. An assumed test reliability of 0.8 to 0.9 for these students accounts for the .04 difference.

The pretest adjustment method thus appears to remove the initial differences due to subject nonequivalence. It therefore provides a fairly accurate estimate of the overall actual benefit of desegregation on black achievement. According to Glass et al. (1981: 103), each .1 ES is equal to .1 grade equivalent or one month of educational gain. Thus desegregated students may be gaining about two months by attending an integrated environment.

The analysis indicates only a slight but statistically nonsignificant gain for the few cases where results greater than one school year were reported. Similarly, in only a few cases was the percentage of black students reported. When the difference between the percentage of black students 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%. That is,

TABLE 5
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).

students moving from almost completely segregated environments to predominantly white schools showed a sizable effect (1.06 ES using the Glass method). 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 5. As with the overall ES, both effects are positive, indicating a benefit for desegregated students. Contrary to previous research (Krol, 1979; 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. Nor did type of achievement measure interact with other variables to influence effect size. Thus a single overall estimate of achievement effects appears to be an appropriate measure of the impact of desegregation.

THE NIE CORE STUDIES

A similar analysis[6] was performed on the 19 studies selected by the NIE panel of experts. The results are presented in Table 6. 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 addition of these studies, the elimination of others, and the use of different control groups explains most of the differences among the results of the panelists (Cook, 1983).

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.[7] 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 difference is that negative ESs have been added by the panel to the core studies that largely, but not entirely, reflect preexisting 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 ($r = .76$) indicating that preexisting differences largely remain at the posttest. Thus subject equivalence is a persistent source of bias in these studies: for this reason, 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), though a number of the NIE panelists omitted this study from their analyses.

Cook (1983) has argued that the distribution of effect size estimates is seriously skewed. However, an analysis yielded a skewness value of .89, which is statistically significant only at the .05 level with a one-tailed test (Snedecor and Cochran, 1967). Even this finding becomes nonsignificant when the Hedges (1982) correction for bias is used (again, the ES is nearly identical).

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 classified by duration or length of desegregation, there

TABLE 6
Effect Sizes for NIE Core Studies

Name of Study (N=19)	# of Cases (N=62)	% Black		Grade Level		Achievement Effect Size		Pretest-Adjusted Fffect Size
		Seg.	Deseg.	Pretest	Posttest	Reading	Math	
Anderson (1966)	2	NA[1]	NA	2	4	.63	--	.95
		NA	NA	2	4	--	.59	.53
Beker (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	--
		NA	22	3	3	.02	--	--
		NA	22	3	3	--	.03	--
		NA	22	4	4	.02	--	--
		NA	22	4	4	--	.03	--

Evans (1973)	6	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	--
Rentsch (1967)	6	90	5	3	5	1.14	--	.15
		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
Slone (1968)	2	60	NA	4	5	.42	--	--
		60	NA	4	5	--	.49	--

(continued)

TABLE 6 Continued

Name of Study (N=19)	# of Cases (N=62)	% Black		Grade Level		Achievement Effect Size		Pretest-Adjusted Effect Size
		Seg.	Deseg.	Pretest	Posttest	Reading	Math	
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	NA NA NA NA	NA NA NA NA	NA NA NA NA	-.13 .11 .16 .29	-.29 -.28 .36 -.06	.11 -.24 .21 -.01
Zdep (1971)	2	NA NA	12 12	2 2	2 2	.34 --	-- -.15	.65 -.15
OVERALL MEAN[2]	(N= 62)	82.49	15.03	4.05	5.12	.28	.23	.14
MEAN FOR TREATMENTS LASTING ONE YEAR OR LESS[2]	(N= 20)	71.00	11.58	3.65	4.20	.30	.11	.13
MEAN FOR TREATMENTS LASTING MORE THAN ONE YEAR[2]	(N= 14)	95.31	17.50	4.00	5.81	.28	.39	.12

[1] Not ascertainable.

[2] Mean effect sizes, weighted by study.

is an interaction; mathematics shows larger effects for those studies longer than one year. Although relatively few cases are 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 constituted the majority of those reviewed by Krol (1979) and St. John (1975).

Cook (1983) tried to use these figures and those of the other panelists to conclude that desegregation "probably did not cause an increase in math skills" but "probably did cause an increase in reading skills." His analysis is based, in part, on a recomputation of the results from Table 6. However, he examined only the 11 studies with pretest adjustments. This includes only 17 cases, with 5 coming from the 2 negative studies added by a conservative panelist. Morevoer, it ignores the 10 cases where there were no selection differences. In our opinion there are too few studies and cases to justify such conclusions, especially given the contradictory results in other reviews that have included more studies.

DISCUSSION

What do these findings mean in the context of current social policy? One of the major problems confronting the use of meta-analysis in policy research is interpretation of the results. As noted above, Glass has provided a heuristic for translating effect sizes into a more meaningful metric—namely, grade equivalents.[8] The effect size found in both analyses reported here indicates about a two-month gain or benefit for desegregated students. Still, the meaning attached to this finding represents a value judgment. This is precisely where social science ends and social policy begins.

Sechrest and Yeaton (1981) discussed the problems involved in interpreting effect size estimates and suggested two ways of attaching social significance to such findings. The first involves a judgmental approach based on experts in a given field. This is similar to the consensus development process currently employed by the National Institutes of Health (NIH) to assess medical technologies (Perry and Kalberer, 1980). However, NIH has found that only neutral panelists (i.e., those with no prior published views) can reach consensus on the issues.

The NIE expert panel was distinctly nonneutral in its composition. Although all the members could agree that there was at least a positive effect for school desegregation, they could not reach agreement on a consensus statement. Many of the panelists excluded studies, substituted alternative control groups, and sought missing information directly from authors of the original studies in accordance with their prior beliefs about the effectiveness of desegregation. The panel thus disbanded with their initial views intact. If Sechrest and Yeaton's approach is to be tried, perhaps a second expert panel composed of truly neutral members should be convened to review the NIE panel's methods and findings.

Sechrest and Yeaton's second recommendation for interpreting these results involved the use of quantitative norms for comparison. Walberg (1983) employed this approach in evaluating the .17 ES he obtained for the NIE Core Studies. He found this a comparatively small effect with respect to other educational interventions, such as Bloom's (1976) mastery learning program, where ESs greater than 1 have been reported. Cook (1983) observed that this is not a valid comparison, given that such effects are unlikely to be maintained. However, a more fundamental issue is this: What is an acceptable comparison?

The answer depends on how one defines the problem. If the objective is to achieve educational benefits while maintaining racial equality, then alternative programs such as magnet schools become relevant comparison programs. If one focuses only on the educational benefits, as Walberg recommends, then all educational interventions become relevant comparisons. Typically, the former approach is taken in program evaluation and cost-benefit/cost-effectiveness analyses where alternatives for accomplishing the same programmatic or policy objectives are examined. This was done neither in the present study nor by the NIE panel, and thus may limit its utility for social policy formation.

Although various policy alternatives have been proposed and some have actually been implemented, no systematic, high-quality evaluative studies of these interventions have been conducted. The studies of busing examined here constitute the largest body of scientific information on the effectiveness of one social policy. Even the more scientifically sound of these studies are limited, in that variation in effectiveness among schools or school district programs cannot be fully explained due to a second major problem noted by St. John (1975) concerning equivalence of schools. The details of the educational programs involved in the desegregation studies are not reported. Thus it is not possible to determine effective from ineffective programs. However, researchers have developed procedures for improving educational practice in desegregated classrooms (Aronson and Bridgeman, 1979; Slavin and Madden, 1979). Such research based on sound social science theory is likely to lead to increased educational benefits for desegregated students.

Even if these results are seen as valid and the most socially meaningful policy alternative, policymakers must also ask if they are representative. In the present case studies that had numerous or severe threats to validity were excluded. As a result, the final sample of accepted studies is no longer representative of all the studies or all geographical locations. For example, studies of mandatory busing in the South Atlantic states from the early 1960s were more likely to be rejected; those of voluntary busing in New England or the Middle Atlantic states were more likely to be included because they involved stronger research designs. The tradeoff between methodological rigor (i.e., internal validity) and representativeness (i.e., external validity) is inevitable when the scientific research literature in question is quasi-experimental in nature (Wortman, 1983).

The decision to emphasize internal over external validity does not mean that the latter is unimportant in meta-analysis. On the contrary, perhaps the most powerful function of research synthesis as a tool in policy analysis is to identify specific settings and populations in which the intervention is most effective—an issue that may be largely obscured in individual studies (see Cronbach et al, 1980). The present strategy, emphasizing valid statements about cause and effect, represents a distinct departure from the traditional meta-analytic procedure in which all studies are combined regardless of methodological quality. Indeed, previous meta-analyses of the desegregation literature (e.g., Crain and Mahard, 1982; Krol, 1979) employed the traditional approach. Despite the differences in meta-analytic strategy, the results of all three studies are nearly identical, lending increased credence to both the validity and representativeness of their findings.

SUMMARY

The synthesis of scientific research using formal statistical procedures such as Glass's meta-analysis presents special problems when studies are methodologically flawed. The research literature on the effectiveness of school desegregation on black achievement is composed almost totally of quasi-experiments or weaker research designs. Although 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 were found (Glass and Smith, 1981; Gilbert et al., 1977; Sacks et al., 1982; 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 (Bryant and Wortman, forthcoming). 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, further reducing the set to 18 core 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.

Given that most of these studies suffered from initial subject nonequivalence, an adjusted effect size was calculated by subtracting 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 effect size is similar to that found for the larger set of 31 studies and also to Krol's (1979) finding. In examining the results of the two analyses reported above, the best overall estimate of the effect of school desegregation on black achievement appears to be about .2 standard deviations. 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 percentage of black students 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 1-7 and a decrease from 8-12. This result does not agree with other findings indicating larger benefits the earlier desegregation occurs. However, Crain and Mahard included very large kindergarten effect sizes; these were omitted in this study because no pretest scores were available and achievement tests are unreliable at that age. 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 Coleman Report's finding that effects are greatest in the most integrated environments.

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 (changes 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)

*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.

*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.

*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. 323-326.

*Thompson, E. W., and 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.

*Article included in NIE Core Studies.

NOTES

1. Cohen's estimate of effect size, d, is nearly identical. The denominator includes information from the treatment and control groups, "since they are assumed equal" (Cohen, 1977). Hedges (1982) maintains that the pooled within standard deviation should

be used because it 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 also impossible to calculate effect sizes from this study because standard deviations were not reported. Similar problems plague the earlier reports as well.

3. In fact, if differential growth is the only cause of change from time 1 to time 2, according to the fan spread model (Campbell and Erlebacher, 1970; Cook and Campbell, 1979), an increase in the mean difference over time will be 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 a selection × maturation interaction will not bias the estimate of effect size for quasi-experiments that are pretest-adjusted.

4. In fact, one of the neutral members had testified in numerous court cases against desegregation.

5. Both St. John's and Crain and Mahard's findings result entirely from the unusually large effect sizes found for grades K and 1. We included none of the former and only two cases of the latter because tests are notoriously unreliable for students of this age. Moreover, no pretest data were available for kindergarten students.

6. The NIE expert panel endorsed the pretest adjustment procedure described in equation 1.

7. This was identical to the mean effect size value obtained by all the panelists, with a range from .04 to .28.

8. Some panelists objected to this because the ESs often were not derived from grade equivalents. However, such a transformation is used perfectly legitimately to interpret the meaning of these unitless measures. The real issue is establishing the equivalence of ESs to some socially meaningful measurement.

REFERENCES

ARONSON, E. and D. BRIDGEMAN (1979) "Jigsaw groups and the desegregated classrooms: in pursuit of common goals." Personality and Social Psychology Bull. 54: 438-446.

BLOOM, B. S. (1976) Human Characteristics and School Learning. New York: McGraw-Hill.

BORUCH, R. F. and H. GOMEZ (1977) "Sensitivity, bias, and theory in impact evaluations." Professional Psychology 8: 411-434.

BRADLEY, L. A. and G. W. BRADLEY (1977) "The academic achievement of black students in desegregated schools: a critical review." Rev. of Educ. Research 47: 399-449.

BRYANT, F. B. and P. M. WORTMAN (forthcoming) "Methodological issues in the meta-analysis of quasi-experiments." New Directions for Program Evaluation.

———(1978) "Secondary analysis: the case for data archives." Amer. Psychologist 13: 381-387.

CAMPBELL, D. T. and A. E. ERLEBACHER (1970) "How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful," pp. 185-210 in J. Hellmuth (ed.) Compensatory Education: A National Debate. Volume 3, Disadvantaged Child. New York: Brunner/Mazel.

CAMPBELL, D. T. and J. C. STANLEY (1966) Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally.

COHEN, J. (1969) Statistical Power for the Behavioral Sciences. New York: Academic Press.

COLEMAN, J. S., E. Q. CAMPBELL, C. J. HOBSON, J. McPARTLAND, A. M. MOOD, F. D. WEINFELD, and R. L. YORK (1966) Equality of Educational Opportunity. Washington, DC: Government Printing Office.

COOK, T. D. (1983) Critical Examination of Meta-Analyses of School Desegregation and the Academic Achievement Gains of Black Children." (Report to National Institute of Education.) Evanston, IL: Northwestern University.

———and D. T. CAMPBELL (1979) Quasi-Experimentation: Design and Analysis Issues for Field Settings. Boston: Houghton Mifflin.

COOPER, H. M. (1982) "Scientific guidelines for conducting integrative research reviews." Rev. of Educ. Research 52: 291-302.

CRAIN, R. L. (1983) Is Nineteen Really Better Than Ninety-Three? (Technical Report). Washington, DC: National Institute of Education.

———and R. E. MAHARD (1982) Desegregation Plans That Raise Black Achievement: A Review of the Research (N-1844-NIE). Santa Monica, CA: Rand Corporation.

———(1978) "Desegregation and Black achievement: a review of the research." Law and Contemporary Problems 42: 17-56.

CRONBACH, L. J., S. R. AMBRON, S. M. DORNBUSCH, R. D. HESS, R. C. HORNIK, and Associates (1980) Toward Reform of Program Evaluation. San Francisco: Jossey-Bass.

DIRECTOR, S. M. (1979) "Underadjustment bias in the evaluation of manpower training." Evaluation Q. 3: 190-218.

EYSENCK, H. J. (1978) "An exercise in mega-silliness." Amer. Psychologist 33: 517.

GEHAN, E. A. and E. J. FREIREICH (1974) "Non-randomized controls in cancer clinical trials." New England J. of Medicine 290: 198-203.

GILBERT, J. P., B. McPEEK, and F. MOSTELLER (1977) "Progress in surgery and anesthesia: benefits and risks of innovative therapy," in J. P. Bunker et al. (eds.) Costs, Risks, and Benefits of Surgery. New York: Oxford Univ. Press.

GLASS, G. V (1978) "Reply to Mansfield and Busse." Educ. Research 7: 3.

———(1977) "Integrating findings: the meta-analysis of research," in L. S. Shulman (ed.) Review of Research in Education. Volume 5. Itasca, IL: Peacock.

———(1976) "Primary, secondary and meta-analyses of research." Educ. Researcher 5: 3-8.

———B. McGAW, and M. L. SMITH (1981) Meta-Analysis in Social Research. Beverly Hills, CA: Sage.

———and M. L. SMITH (1981) "Meta-analysis of research on class size and achievement." Educ. Evaluation and Policy Analysis 1: 2-16.

GRANT, G. (1975) "Shaping social policy: the politics of the Coleman Report." Teachers College Record 75: 17-54.

HEDGES, L. V. (1982) "Estimation of effect size from a series of independent experiments." Psych. Bull. 92: 490-499.

IWANICKI, E. F. and R. K. GABLE (1978) "A quasi-experimental evaluation of the effects of a voluntary urban/suburban busing program on student achievement."

Presented at the Annual Meeting of the American Educational Research Association, Toronto, March.

JACKSON, G. B. (1980) "Methods for integrative reviews." Rev. of Educ. Research 50: 438-460.

KLUGER, R. (1975) Simple Justice. New York: Random House.

KROL, R. A. (1979) "A meta analysis of comparative research on the effects of desegregation on academic achievement." Ph.D. dissertation, University of Michigan—Ann Arbor (University Microfilms 7907962).

LANDMAN, J. T. and R. M. DAWES (1982) "Psychotherapy outcome: Smith and Glass' conclusions stand up under scrutiny." Amer. Psychologist 37: 504-516.

LIGHT, R. J., and P. V. SMITH (1971) "Accumulating evidence: procedures for resolving contradictions among different research studies." Harvard Educ. Rev. 41: 429-471.

LINSENMEIER, J.A.W. and P. M. WORTMAN (1978) "The Riverside School Study of desegregation: a re-examination." Research Rev. of Equal Education 2: 1-40.

MANSFIELD, R. S. and T. V. BUSSE (1979) "Meta-analysis of research: a rejoinder to Glass." Educ. Research 6: 3.

MOSKOWITZ, J. M. and P. M. WORTMAN (1981) "A secondary analysis of the Riverside School Study of Desegregation," in R. F. Boruch et al. (eds.) Reanalyzing Program Evaluations. San Francisco: Jossey-Bass.

PERRY, S. and J. T. KALBERER (1980) "The NIH consensus development program and the assessment of health-care technologies: the first two years." New England J. of Medicine 303: 169-172.

RENTSCH, G. J. (1967) "Open-enrollment: an appraisal." Ph.D. dissertation, State University of New York, Buffalo.

ROCK, W. C. et al. (1968) A Report on a Cooperative Program Between a City School District and a Suburban School District. Rochester, NY.

ROSENTHAL, R. (1978) "Combining results of independent studies." Psych. Bull. 85: 185-193.

SACKS, H., T. C. CHALMERS, and H. SMITH (1982) "Randomized versus historical controls for clinical trials." Amer. J. of Medicine 72: 233-240.

SAMUELS, J. M. (1971) "A comparison of projects representative of compensatory busing; and non-compensatory programs for inner-city students." Ph.D. dissertation, University of Connecticut.

SECHREST, L. and W. YEATON (1981) "Empirical bases for estimating effect size," in R. F. Boruch et al. (eds.) Reanalyzing Program Evaluations. San Francisco: Jossey-Bass.

SHEEHAN, D. S. (1979) "Black achievement in a desegregated school district." J. of Social Psychology 107: 185-192.

SLAVIN, R. E. and N. A. MADDEN (1979) "School practices that improve race relations." Amer. Educ. Research J. 16: 169-180.

SMITH, M. L. and G. V. GLASS (1977) "Meta-analysis of psychotherapy outcome studies." Amer. Psychologist 32: 752-760.

———and T. I. MILLER (1980) The Benefits of Psychotherapy. Baltimore, MD: Johns Hopkins Univ. Press.

SNEDECOR, G. W. and W. G. COCHRAN (1967) Statistical Methods. Ames: Iowa State Univ. Press.

STAINES, G. L. (1974) "The strategic combination argument," in W. Leinfellner and E. Kohler (eds.) Developments in the Methodology of Social Science. Dordecht, Holland: Reidel.

STEPHAN, W. G. (1982) Blacks and *Brown*: The Effects of School Desegregation on Black Students. (Technical report). Washington, DC: National Institute of Education.

ST. JOHN, N. H. (1975) School Desegregation Outcomes for Children. New York: John Wiley.

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

WALBERG, H. J. (1971) "An evaluation of an urban-suburban school bussing program: Student achievement and perception of class learning environments." Presented at the annual meeting of the American Educational Research Association, New York.

———(1983) "Desegregation and educational productivity." (Technical Report). Washington, DC: National Institute of Education.

WEINBERG, M. (1977) Minority Students: A Research Appraisal. Washington, DC: U.S. DHEW, National Institute of Education.

WORTMAN, P. M. (1983) "Evaluation research: a methodological perspective." Annual Rev. of Psychology 34: 223-260.

———(1981) "Randomized clinical trials," in P. M. Wortman (ed.) Methods for Revaluating Health Services. Beverly Hills, CA: Sage.

———C. KING, and F. B. BRYANT (1982) "Meta-analysis of quasi-experiments: school desegregation and black achievement. Part I—Retrieval and coding." Ann Arbor, MI: Institute for Social Research.

ZDEP, S. M. (1971) "Educating disadvantaged urban children in suburban schools: an evaluation." J. of Applied Social Psychology 1. (ERIC number ED 053 186 TM 00716).

Paul M. Wortman is Director of the Methodology and Evaluation Research Program in the Center for Research on Utilization of Scientific Knowledge in the Institute for Social Research, and Professor in the Department of Medical Care Organization of the School of Public Health, both at the University of Michigan—Ann Arbor. He served as president of the Evaluation Research Society in 1983. He has been conducting evaluation research for over ten years. Among his current interests are research synthesis methods and medical technology assessment.

Fred B. Bryant is an Assistant Professor of Social Psychology at Loyola University of Chicago. His research interests include applying meta-analysis to quasi-experimental literatures and developing guidelines for selecting appropriate evidence.