DOCUMENT RESUME

ED 241 671 UD 023 589

AUTHOR Cook, Thomas; And Others
TITLE School Desegregation and Black Achievement.
INSTITUTION National Inst. of Education (ED), Washington, DC.
PUB DATE May 84
NOTE 227p.; Also printed separately as ED 238 999-239 005; Several pages have small, marginally legible print.
PUB TYPE Collected Works - General (020) -- Information Analyses (070)

EDRS PRICE MF01/PC10 Plus Postage.
DESCRIPTORS *Academic Achievement; *Black Achievement; *Black Students; *Desegregation Effects; Elementary Secondary Education; Literature Reviews; *School Desegregation

ABSTRACT

Seven papers commissioned by the National Institute of Education in order to clarify the state of recent knowledge about the effects of school desegregation on the academic achievement of black students are contained in this report. The papers, which analyze 19 "core" empirical studies on this topic, include: (1) "What Have Black Children Gained Academically from School Integration? Examination of the Meta-Analytic Evidence," by Thomas D. Cook; (2) "The Evidence on Desegregation and Black Achievement," by David J. Armor; (3) "Is Nineteen Really Better Than Ninety-Three?" by Robert L. Crain; (4) "School Desegregation as a Social Reform: A Meta-Analysis of Its Effects on Black Academic Achievement," by Norman Miller and Michael Carlson; (5) "Blacks and 'Brown': The Effects of School Desegregation on Black Students," by Walter G. Stephan; (6) "Desegregation and Education Productivity," by Herbert J. Walberg; and (7) "School Desegregation and Black Achievement: An Integrative View," by Paul M. Wortman. The 19 core studies examined in these papers were selected, based on their content and quality, from 157 works that looked at black students' academic achievement in desegregated schools. Authors of the selected works are Lewis V. Anderson, Jerome Baker, Orrin H. Bowman, Patricia M. Carrigan, El Nadel Clark, Charles L. Evans, E. F. Iwanicki and R. K. Gable, Robert Stanley Klein, M. A. Laird and G. Weeks, George J. Rentsch, L. W. Savage, Daniel S. Sheehan, Irene W. Slone, Lee Rand Smith, the Syracuse City School District, E. W. Thompson and U. Smidchens, D. W. Van Every, Herbert J. Walberg, and Stanley M. Zdep. (GC)

SCHOOL DESEGREGATION

AND

BLACK ACHIEVEMENT

The National
Institute of
Education
U.S. Department of
Education
Washington, D.C. 20208

SCHOOL DESEGREGATION AND BLACK ACHIEVEMENT

Thomas Cook
David Armor
Robert Crain
Norman Miller
Walter Stephan
Herbert Walberg
Paul Wortman

U. S. Department of Education

Office of Educational Research and Improvement

National Institute of Education

Washington, D.C. 20208

May 1984

TABLE OF CONTENTS

The seven (7) papers that have been collected to comprise this report originally appeared in separate and somewhat larger forms. The earlier versions are identified below via "see also" notes leading to their ERIC accession numbers.

A limited number of copies of this document are available from the National Institute of Education (NIE), 1200 19th Street, NW, Washington, DC 20208. When this limited stock is exhausted, requestors should order the document from the ERIC Document Reproduction Service (EDRS), P.O. Box 190, Arlington, Virginia 22304.

INTRODUCTION

The National Institute of Education has undertaken the most comprehensive and rigorous analysis to date of the effect of desegregation on Black student academic achievement. NIE commissioned papers from seven eminent scholars to clarify the state of research knowledge about the effects of school desegregation on the academic achievement of Black students, and the seven scholars are Thomas Cook of Northwestern University, David Armor of David Armor Associates, Robert Crain of the Rand Corporation, Norman Miller of the University of Southern California, Walter Stephan of New Mexico State University, Herbert Walberg of the University of Illinois-Chicago Circle, and Paul Wortman of the University of Michigan.

They were selected for their past extensive work on desegregation research, prominence in the field, knowledge of research methodology, and divergent viewpoints about the effects of desegregation on Black student academic achievement. NIE's intention was to find if under similar conditions, with the same set of data, and common ground rules, similarities and differences in analyses could be identified and clarified.

The seven scholars met first to discuss the state of research literature and to agree on a comprehensive list of criteria to be used in selecting the studies to be analyzed. A total of 157 empirical studies were identified that looked at Black students' academic achievement in desegregated schools. A comprehensive and rigorous list of criteria (listed below) were adopted and applied to the total set. This process resulted in a "core" of 19 highest quality studies (listed below) on this research topic, which the scholars then statistically analyzed to reach their individual conclusions. This analytical effort is a significant improvement over previous attempts at reconciling the controversial literature on this topic, and it is hoped that this effort by NIE will prove helpful to all parties concerned with the nationally important subject of school desegregation.

CRITERIA FOR REJECTION OF A STUDY

1) Type of Study

   a) non empirical
   b) summary report

2) Location

   a) outside USA
   b) geographically non specific

3) Comparisons

   a) not a study of achievement of desegregated Blacks (except in cases where we use a White comparison)
   b) multi-ethnic combined
   c) comparisons across ethnics only
   d) heterogeneous proportions minority in desegregated condition
   e) no control data
   f) no pre-desegregation data
   g) control measures not contemporaneous
   h) excessive attrition (review must provide specific justification for the inclusion of stndies with excessive attrition, but amount was not specified)
   i) majority Black in a segregated condition (unless the reviewer provides specific justification)
   j) varied exposure to desegregation (unless the reviewer provides a specific justification demonstrating that the variation in exposure time is not meaningful)
   k) groups are initially non-comparable (unless the reviewer provides a specific justification that the amount of divergence is not meaningful)

4) Study Desegregation

   a) cross-sectional survey
   b) sampling procedure unknown
   c) separate non-comparable samples at each observation

5) Measures

   a) unreliable and/or unstandardized instruments
   b) test content and/or instrument unknown
   c) dates of administration nnknown
   d) different tests used in pretests and posttests
   e) test of IQ or verbal ability

6) Data Analysis

   a) no pretest means
   b) no posttest means, nnless the anthor reported pretest scores and gains
   c) no data presented
   d) N's not discernible

19 CORE STUDIES

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

Baker, Jerome. A study of segregation in racially imbalanced urban public schools. Syracuse, New York: Syracuse University Youth Development Center, Final Report, May 1977.

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.

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

Clark, El Nadel. Analysis of the difference between pre- and post-test 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.

Evans, Charles L. Short term desegregation effects: The academic achievement of bused students 1971-1972. Fort Worth, Texas: Fort Worth Independent School District, 1973. (ERIC No. ED 086 759)

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.

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

Savage, L. W. Arithmetic achievement of Black students transferring from a segregated junior high school to an integrated junior 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, 165-182.

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 effects of integration -- Washington Irving and Host Pupils. Hearing held in Rochester, New York, September 16-17, 1966, U.S. Commission on Civil Rights.

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.W. Effects of desegregation on pupil 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)

Walberg, Herbert J. An evaluation of an urban-suburban school busing program: Student achievement and perception of class learning environments. Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February 1971. (ERIC No. ED 047 076 UD 011 284)

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)

What Have Black Children Gained Academically From School Integration?: Examination of the Meta-Analytic Evidence

Thomas D. Cook
Northwestern University

INTRODUCTION

My assignment is to comment on the following essays by Armor, Crain, Miller, Stephan, Walberg and Wortman in order to help readers decide what should be concluded from their evaluations of how school desegregation has affected the academic achievement of black children. All but two of the essays contain a meta-analysis by the author. Crain's paper is one of the exceptions. Instead of conducting a meta-analysis, he critically discusses some of the assumptions behind the others' efforts and concludes that he will stand by the results of his own prior meta-analytic work (Crain & Mahard, 1983). I shall refer to his prior meta-analysis based on 93 studies more than to his essay in this volume. Walberg is the other exception. He devotes most of his essay to a review of factors other than desegregation that raise academic achievement. He does this to make the point that, if the purpose of desegregation is to raise the achievement of black children, then more effective means exist to do this than desegregation. Walberg does, however, reanalyze three prior meta-analyses--by Krol (1975), Crain & Mahard (1982), and Wortman, King, and Bryant (1982)--in order to make the further point that, in his estimation, the average effect sizes they present do not reliably differ from zero. I intend to deal with his statistical analysis to a small extent, but will not deal directly with his larger point about relative efficacy.

The first part of the present paper deals with the meta-analytic work of Armor, Miller, Stephan and Wortman, and is largely restricted to the 19 studies selected by the panel. The purpose is to arrive at an estimate for this sample of how desegregation has affected the achievement of black children. I try to restrict my commentary to the most important points and assumptions made by the authors, and make no attempt at a comprehensive analysis of any single person's work in order to be comprehensive about its strengths and weaknesses. This is to keep the focus on the desegregation issue. In the second part of the paper, I take my own results, which are both similar to and different from those of the panel, and discuss several ways they can be interpreted. In particular, I ask how generalizable are results from the panel's 19 studies when they are compared to the results from larger data bases; I probe the extent to which my findings speak to the information needs of groups with different stakes in school desegregation; and I speculate about whose interests the panel's results might advance or prejudice.

RESULTS

1. The Studies Examined. Individual panel members considered different subsets of the 19 studies that most of them deemed methodologically adequate. Armor dropped the study by Rentsch on grounds, first, that the desegregated group and the segregated controls differed by so much initially; second, that the pretests and posttests involved different measures; and third, that the desegregated control group contained some white children. He also dropped the study by Thompson & Smidchens on grounds that the segregated controls were in classes made up of only 42% minority students. However, he included the study by Carrigan, even though its segregated control group members were in classes that were hardly more "segregated"--50% minority. Indeed, Miller and Stephan dropped the Carrigan study because of its questionable control group. In a few other cases, Armor selected control groups within a study that differed from the choice of all other panelists. The net result of Armor's preferences was lower effect sizes since (1) Rentsch obtained some of the largest effect sizes; (2) Carrigan resulted in both positive and negative effect sizes; and (3) both Rentsch and Carrigan involved multiple comparisons, so their results were disproportionately weighted whenever comparisons were the unit of analysis rather than individual studies.

Miller dropped both Carrigan and Thompson and Smidchens from his analyses because the segregated controls were not segregated. He also differed from the other analysts in preferring to compute an effect size per study instead of per comparison. Much has been written in the meta-analysis literature on this topic, and our preference is to compute or report effect sizes each way. However, if only one choice is available, we favor a sample of studies because this does not weight the results in favor of school districts where desegregation was tested using several grades.

Stephan also omitted the studies by Carrigan and by Thompson & Smidchens. However, he also objected to the studies by Iwanicki & Gable and Slone on grounds that they dealt with the second year of desegregation while other studies dealt with the first year. He further objected to Slone because the segregated controls were attending a school that was 40% white. This left Stephan with only 15 studies to analyze. Since the studies he omitted all tended, with the exception of Slone, to have zero or negative effect size estimates, it is clear that Stephan's sampling decision disposed his analysis towards a larger average effect size than other panelists.

Wortman differed from the other panelists in two important ways. First, he preferred his own selection of 31 "superior" studies to the panel's 19. However, his analyses of the 31 showed that designs without control groups produced higher effects size estimates than designs with control groups. Hence, I treat his analyses based on studies with controls differently from the analyses without controls for, among other possible artifacts, maturation and testing effects can inflate estimates of the desegregation effect. Second, in his analyses of the panel's 19 studies, Wortman was more strict than the others about what he would accept as valid information about variances. Since such information is crucial for computing effect sizes he was able to produce estimates that also controlled for pretest differences between the desegregated and segregated control groups for only 11 of the 19 studies favored by the panel. One of these was the study by Carrigan. Omitted were Clark, Evans, Iwanicki & Gable, Klein, Laird & Weeks, Slone, Syracuse, and Thompson & Smidchens. Since Wortman preferred somewhat different standards of methodological adequacy than the panel, I sometimes include estimates computed from his analyses of the 11 panel studies, and at other times estimates based on the larger subset of his preferred studies that involved designs with control groups. These studies should overlap heavily with the panel's selection criteria.

The panelists provided estimates for reading and math combined, for reading alone, and for math alone. It is interesting to note that there is no obvious relationship between gains in mathematics and reading when the desegregated are compared to the segregated. To compute a correlation of reading and math gains would not be useful because of the small number of studies and comparisons for which there were measures of both reading and mathematics gains. However, of Armor's 18 relevant comparisons, math and reading gains had the same sign in seven instances, different signs in eight, and three instances were indeterminate because of zeros. Of Miller's 13 comparisons, seven had the same sign and six the opposite; while of Stephan's comparisons there were 13 with the same sign, 11 with the opposite, and one was indeterminate. Math and reading gains were not clearly related, and little is gained by adding them together. Consequently, I prefer to present results separately from each knowledge domain. However, for purposes of continuity with the panelists some of my reanalyses will involve reading and math scores combined. When that happens, my analyses--like those of the panelists--weight reading slightly more than math because more reports included reading than math measures.

2. Panelists' Results. Using his own preferred set of studies based on a sample of comparisons. Armor obtained an effect size of .06 for reading and .01 for math; Miller obtained an effect size of .16 for reading and .08 for math; Stephan's values were .15 and .00; while in my analysis of Wortman's resutls for the eleven studies with pretest adjustments, the mean effects were .26 and .08. (Wortman's own results from the panel's 19 studies were .28 and .23, but this includes studies where no pretest adjustments were made. His estimates from his total sample of 31 studies were .57 and.33, but these are based on some studies without control gorups. Thus, I consider both of these last sets of estimates to be problematic).

If we turn now to estimates of reading and math combined, Armor's overall estimate was .04, Stephan's was .14 (but .07 when computed as gain per 8-month school year), Miller's was .12, while Wortman's was .17 derived from the studies of his own choosing that had control groups.

If one took the panel's estimates at face value they would appear to support the following conclusions:

a. Desegregation did not cause a decrease in the achievement of black children.

b. It probably did not cause an increase in math skills, for the mean gains vary from 0 to .08 standard deviation units.

c. It may have caused an increase in reading skills, for the mean gains vary from .06 to .26.

The range estimate for reading deserves comment, since the upper bound comes from our analysis of Wortman's eleven studies where pretest adjustments could be made. This is a considerably smaller sample than the other authors analyzed, and so should be treated as particularly tentative. Omitting it gives a revised range that petmits a fourth conclusion, which I believe to be better justified than the third conclusion immediately above.

d. The gain in reading was somewhere between .06 and .16 standard deviation units. This is between two and six weeks of gain if we follow the rule of thumb of Glass et al (1981) and associate a gain of one-tenth of a standard deviation with one month's gain in knowledge.

The small discrepancies between the panelists in mean estimates principally reflect differences in (1) the studies included for review; (2) the way effect sizes were computed; and (3) a preference for some types of control groups over others within a few studies. I shall resist the temptation to discuss each of these issues in order to make judgments for each of them about the methodological option to be preferred, after which point estimates of gains could be computed. While such an exercise would result in easily remembered single number estimates of reading and math gains, the resulting precision would be misplaced. In meta-analysis, varying the assumptions underlying an analysis is desirable because it makes heterogeneous those facets of research where no "right" answer is available and fallible human judgment is required. To attempt to legislate a single "right" way either to compute effect sizes or to sample studies would be counterproductive so long as none of the analysts is clearly wrong. Indeed, the idea of selecting a panel of methodologically sophisticated experts with different views on school desegregation is predicated on the particular utility that would result if the panel's estimates of desegregation's effects converged despite the differences in values and methodological predilections of individual panelists. It is more reasonable to expect "convergence" as a range than a point. To search for the elusive "true" point estimate of effect could involve laborious debates about fine points of methodology and substance that might occur within a range of estimates that many would think has few practical implications.

Speaking personally, I am impressed by the degree of correspondence between the panelists when only the 19 core studies are considered. None achieves negative estimates; all achieve larger estimates for reading than math: and the largest single difference--between Armor and Miller for reading gains--is of a magnitude many would consider small--viz., a difference of about one month of gain.

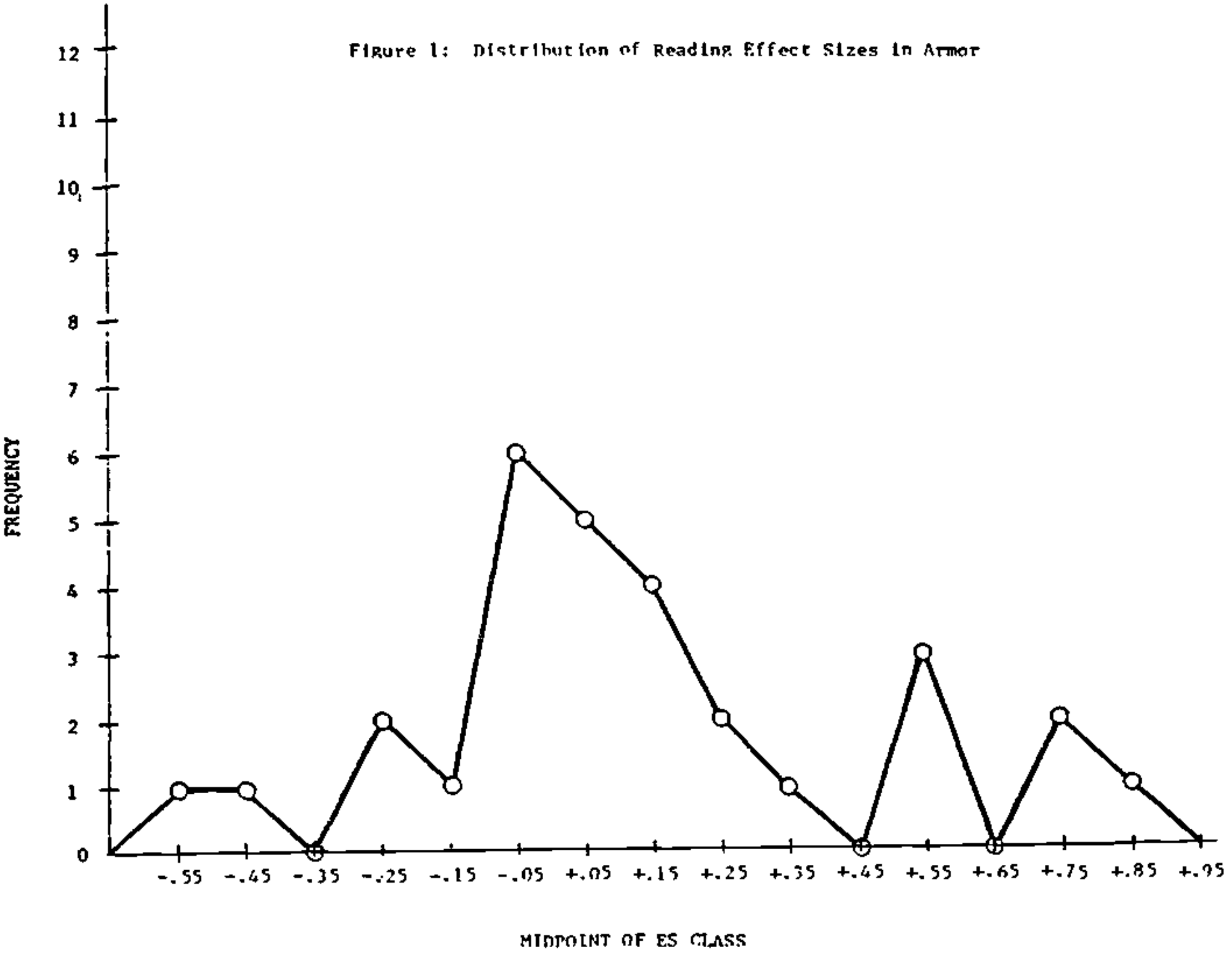
The convergence is all the more dramatic since, across all dependent variables, Krol obtained an estimate of .10 from his own meta-analysis of "better" desegregation studies, while a similar estimate resulted from Crain & Mahard (1983) when one aggregates across all their dependent variables for the randomized experiments and studies with both pretest-posttest measurement and control groups of segregated black children. Combining math and reading and analyzing only the studies preferred by the present panelists, Armor's estimate was .04, Miller's was .12, and Wortman's was .17 for all the studies he found with pretests and black control groups, while Stephan's estimate was .14 without his correction for the length of time desegregation had been taking place--a correction that none of the other panelists made. The average of the panelists' values is .11, only slightly higher than the estimate obtained by Krol and Crain & Mahard. (However, as we later see, Crain rejects this estimate, preferring to base his judgment on studies where desegregation occurs at kindergarten or first grade.)

3. The Distribution Problem. As a measure of central tendency the mean depends on a normal distribution of scores. In Figures 1 through 4, we present frequency distributions of reading effect sizes for Armor, Miller, Stephan, and Wortman based on the studies they chose to analyze. (For Wortman we add the math data since he presents reading effect sizes for only eleven studies where pretest adjustments were made, and this results in a particularly poor estimate of the distribution). In all cases except Miller, the sample sizes are based on comparisons rather than studies. But irrespective of the unit of analysis, the distributions are visibly skewed, with a disproportionate number of effect sizes falling in the upper range.

Table 1 presents the medians and modes corresponding to the reading mean. The median is computed for a sample of both comparisons and studies and is defined as the value of the (N+1)/2th case. To compute a mode with so few cases, we constructed a scale composed of categories with intervals of .10 standard deviation units whose midpoints are presented in Figures 1-4. Each effect size was assigned to its respective category, with scores of zero being assigned in equal proportions to the category 0 to .10 and 0 to -.10. For Miller, no value is reported for the median of comparisons since he only provided data on studies. Sometimes, no mode is presented for Wortman because his smaller sample of studies from the panel's set that had pretest adjustments often makes it difficult to determine any modal category with more than three cases falling into it.

Table 1 shows that mean effect sizes for reading are larger than median effect sizes irrespective of whether the latter are computed as a median

Figure 1: Distribution of Reading Effect Sizes in Armor



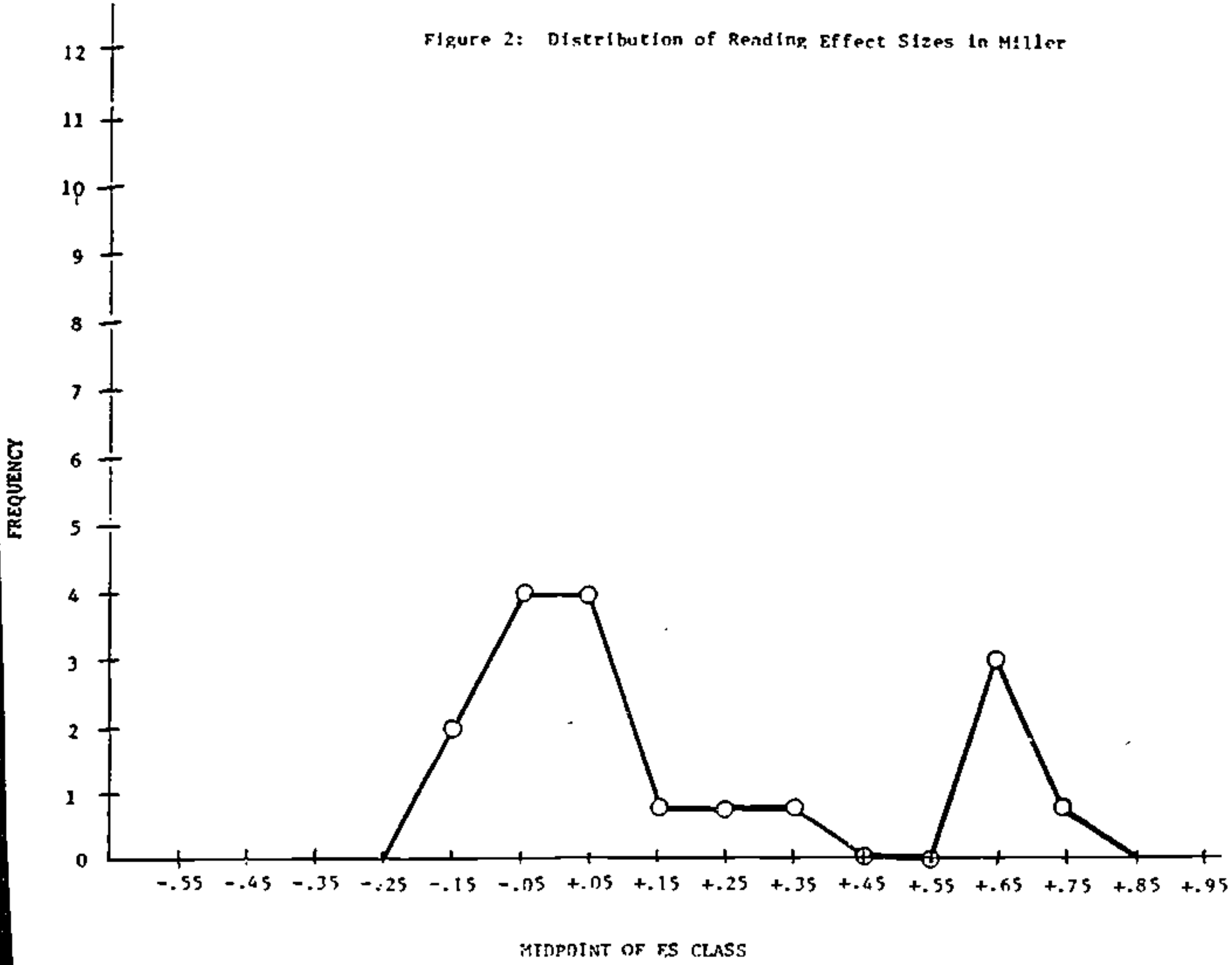


Figure 2: Distribution of Reading Effect Sizes in Miller

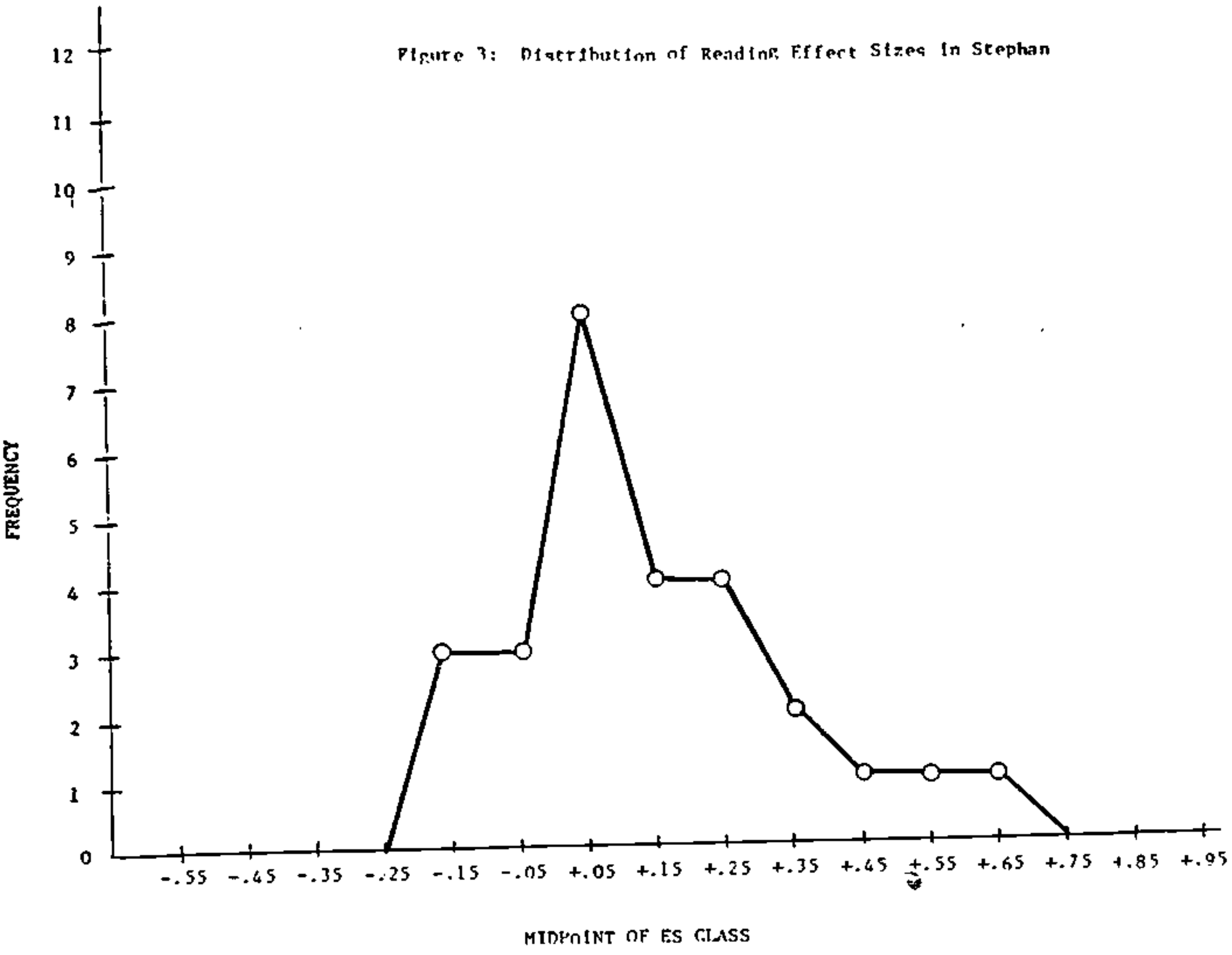


Figure 3: Distribution of Reading Effect Sizes in Stephan

Figure 4: Distribution of Reading and Math Effect Sizes Combined for the Pretest-Adjusted Studies of Wortman

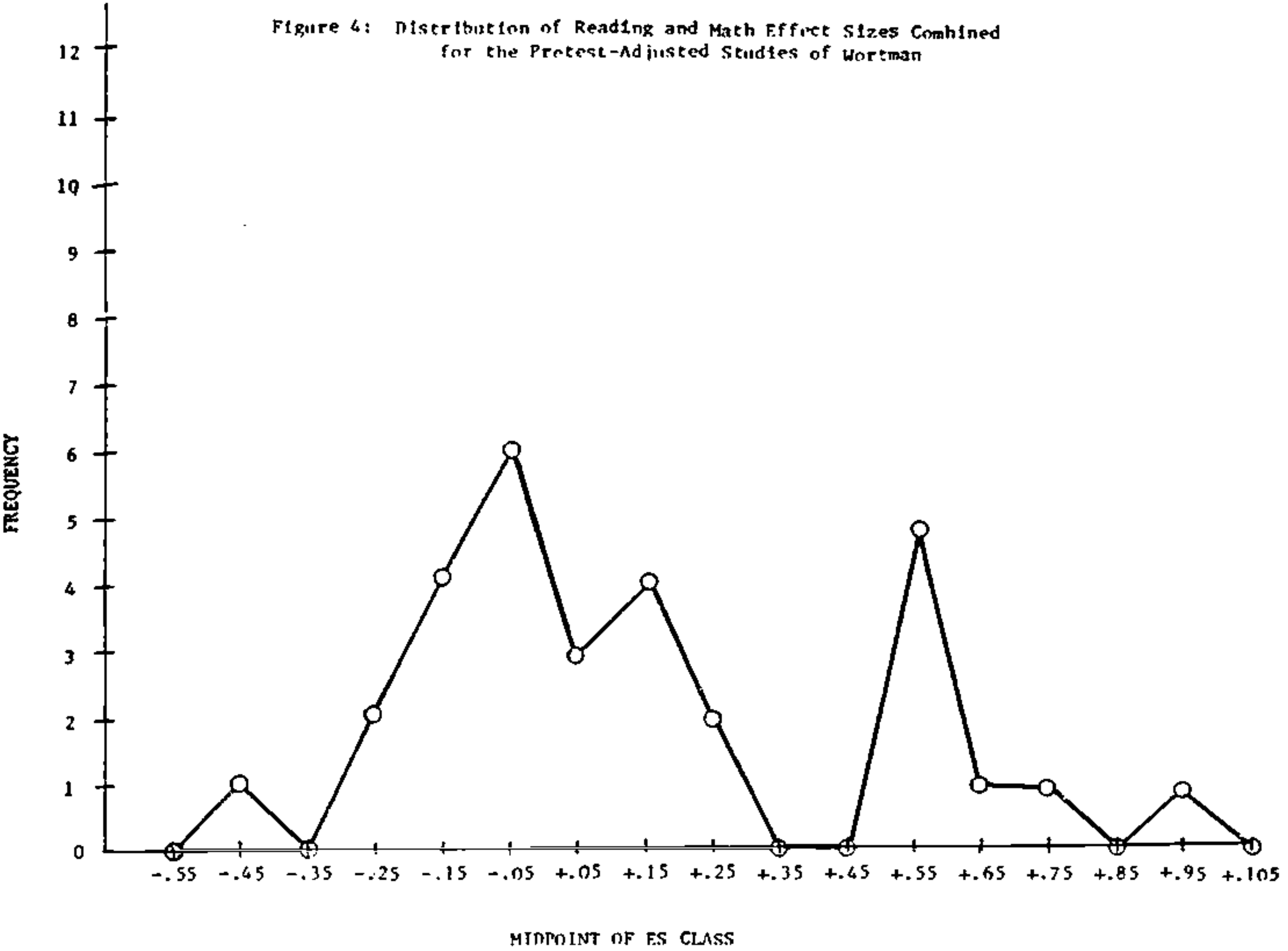


Table 1

Central Tendencies for Reading - Author's own Preferred Studies

	Mean	Median of Comparisons	Median of Studies	Midpoint of Modal Category of Comparisons
Armor	.06	.00	.00	-.05 & +.05
Miller	.16	--	.06	-.05 & +.05
Stephan	.14	.08	.08	+.05
Wortman[a]	.26	.15	.04	--

[a] In Wortman's case "preferred" studies refers to those of his selection from the panel's core 19 for which pretest adjustments could be made. It does not refer to his analysis of 31 studies.

of comparisons or of studies. It also shows that the mode is smaller than the other measures of central tendency and hovers around zero. Indeed, the mean of the mean effect sizes across all four panelists is .15, the mean median of comparisons is .08, the mean median of studies is .05, while the modal categories are of effects between +.05 and -.05.

Table 1 was recomputed based on the 17 core studies most panelists agreed upon. That is, Thompson & Smidchens was omitted since three of the four panelists who did meta-analyses questioned it; and Carrigan was omitted since at least two of the panelists objected to the questionable nature of their "segregated" controls. In computing the data for Armor, the missing values for Rentch were taken from Wortman. Stephan provided his own estimates for the studies by Iwanicki & Gamble and Slone that he preferred to leave out of most of his own analyses. As Table 2 shows, having a common set of studies reduced the dispersion of mean effect size for reading. The range for the panelists--Wortman excepted because his analysis is not based on the 17 studies, and I did not want to take his six missing estimates from other panelists since that would involve estimating about 30% of the scores--the range shifted from .06--.16 to .13--.16. However, even with the same 17 studies per analyst, the table still shows that medians are lower than means, and that modes are lower than medians.

A corresponding table for math from the author's own preferred set of studies is in Table 3. Modes could not reasonably be computed due to the smaller number of math than reading comparisons. However, the means are consistently higher than the medians.

Combining math and reading allows modes to be computed again and results in the same basic relationship between measures of central tendency. This is true whether one uses the author's own set of preferred studies (Table 4) or the common set of 17 (Table 5). The individually preferred studies produced a range of mean estimates from .06 to .16, or median estimates from .00 to .08, and of mode estimates from -.15 to +.05.

These differences in central tendency result because the distribution of effect sizes is skewed. The skewness means that, if one were willing to assume that the present results are applicable to the nation at large today--a dangerous assumption--then (1) for any school district that desegregates the most reasonable expectation is that there will be no effects on black achievement, for the mode suggests that this outcome is obtained more often than any other; (2) 50% of the school districts will probably raise achievement by about three one-hundredths of a standard deviation (the average median of studies across the panelists), while 50% of them will probably raise it by less than this; but (3) the national impact will be to raise the achievement of black children in reading by between two and six weeks and to raise achievement in math, if at all, by something less than three weeks--the upper range of mean estimates. However, (4) a minority of school districts could expect to make larger positive gains. Using Miller's reading estimates for the moment, larger gains appear to have been obtained by Anderson (.733), Beker (.400), Syracuse (.691), and Zdep (.671). In mathematics, the outliers were less common but still visible (Anderson .669, Klein .333, and Van Every .543).

Table 2

Central Tendencies for Reading - 17 Common Core Studies

	Mean	Median of Comparison	Median of Studies[e]	Midpoint of Modal Category of Comparisons
Armor[a]	.13	.03	0	-.05 & +.05
Miller[b]	.16	--	.06	-.05 & +.05
Stephan[c]	.13	.07	.08	+.05
Wortman[d]	.26	.15	.04	--

[a] Based on N of comparisons; Carrigan and Thompson & Smidchens omitted; Rentsch added and given Wortman values.

[b] Based on N of studies; Carrigan and Thompson & Smidchens omitted.

[c] Based on N of comparisons; Carrigan and Thompson & Smidchens omitted. Thus, Iwanicki & Gable and Slone added.

[d] Based on N of comparisons. The sample size is considerably smaller than with other analysts, since Wortman omitted all instances where the control group standard deviation was not specifically given. This resulted in the omission of Clark, Evans, Iwanicki & Gable, Klein, Lard & Weeks, Slone, Syracuse, and Walberg, as well as Carrigan and Thompson & Smidchens. No mode was ascertainable.

[e] The medians are from Miller's Table 2 for each author based on N of studies rather than comparisons.

Table 3

Central Tendencies for ES Values in Math - Author's own Preferred Studies

	Mean	Median of Comparison	Median of Studies	Midpoint of Modal Category of Comparisons
Armor	.01	-.05	-.06	--
Miller	.08	--	.07	--
Stephan	.04	.02	.02	--
Wortman	.03	-.02	-.05	--

[a] In Wortman's case "preferred" studies refers to those of his selection from the panel's core 19 for which pretest adjustments could be made. It does not refer to his analysis of 31 studies.

Table 4

Central Tendencies for Reading and Math Combined - Authors' own Preferred Studies

	Mean	Median of Comparisons	Median of Studies	Midpoint of Modal Category of Comparisons
Armor	.06	.00	.00	-.05
Miller	.12	--	.06	-.15 & +.05
Stephan[b]	.07	.05	.05	-.05
Wortman[a]	.16	.08	.01	-.05

[a] In Wortman's case "preferred" studies refers to those of his selection from the panel's core 19 for which pretest adjustments could be made. It does not refer to his analysis of 31 studies.

[b] These are estimates per school year

Table 5

Central Tendencies for Reading and Math - 17 Common Core Studies

	Mean	Median of Comparisons	Median of Studies[e]	Midpoint of Modal Category of Comparisons
Armor[a]	.08	0	0	-.05
Miller[b]	.12	--	.06	-.15 & +.05
Stephan[c]	.07	.03	.06	+.05
Wortman[d]	.16	.08	.01	-.05

[a] Based on N of comparisons; Carrigan and Thompson & Smidchens omitted; Rentsch added and given Wortman values.

[b] Based on N of studies; Carrigan and Thompson & Smidchens omitted.

[c] Based on N of comparisons; Carrigan and Thompson & Smidchens omitted. Thus, Iwanicki & Gable and Slone added. Estimates of effect per school year.

[d] Based on N of comparisons. The sample size is considerably smaller than with other analysts, since Wortman omitted all instances where the control group standard deviation was not specifically given. This resulted in the omission Clark, Evans, Iwanicki & Gable, Klein, Laird & Weeks, Slone, Syracuse, and Walberg, as well as Carrigan and Thompson & Smidchens.

[e] The medians are from Miller's Table 2 for each author based on N of studies rather than comparisons.

But Stephan's estimates make the studies with outlying results seem less extreme, and some different outliers emerge. He computes effect sizes in a way that controls for the length of time children have been under study in a desegregated school. When reading effect sizes are computed per eight-month school year, the outliers are pulled in because they tended to come from studies lasting two or three years. The new values are: Anderson (.42), Baker (.13), and Zdep (.66). (Stephan leaves Syracuse out of his sample). For mathematics, the positive outliers now become: Anderson (.24), Klein (.33), and Van Every (.14). Stephan's computation of effect sizes leads to less variable and less skewed estimates than the other panelists, which is why medians and modes make less of a difference to his computations of central tendency than to others. But the choice of a measure of central tendency still makes a difference in Stephan's estimates, for both reading and reading and math combined.

However, Stephan's work does present a puzzle. He is the sole panelist to compute a median, and about midway in his report he mentions that the median gain in verbal achievement (reading) is .13. (His corresponding means were .17 for the sample of comparisons and .15 for the sample of studies.) I have examined Stephan's effect sizes from his Table 1 and have been unable to arrive at the same value. My own estimate based on a sample of comparisons and omitting the studies he leaves out is .08. Readers should scrutinize Stephan's Table 1 and estimate for themselves the effect size for reading scores above which 50% of the effect sizes fall and below which 50% fall.

4. The Confidence Problem. Our reanalysis of the panelists' studies using multiple measures of central tendency should not be interpreted to mean that, in our opinion, desegregation has had no effect on most schools. There are two reasons for a low level of confidence in the results presented in Tables 1 through 5. First, we do not know the underlying distribution of mean effect sizes (however computed) for the population of school districts that have already desegregated. It is not clear how representative the panel's core set of studies are. Second, with so few comparisons and studies, we cannot have much confidence in the sample distributions presented in Figures 1-4. A dozen new cases could radically alter each of the estimates of central tendency. With such a poorly estimated and unstable distribution, it is not clear that the mean would remain unchanged even if more cases were added from the very same population that the present sample is supposed to represent.

Statistical significance tests are typically used to make inferences about the level of confidence one should ascribe to findings. (Because of lay misunderstandings of the word "significance," we prefer to talk of tests of statistical reliability rather than statistical significance.) Walberg has maintained that for measures of math and reading combined, none of the estimates obtained by Krol, Crain & Mahard and Wortman, King & Bryant reliably differ from zero. In the current case, our calculations of reliability indicate that: (1) for Armor, the mean estimates for math alone and for reading and math combined do not differ from zero, but the estimate for reading does so marginally (p is less than .10); (2) for Miller, the estimate for math does not reliably differ from zero, but the estimates for reading alone and for reading and math combined do so; (3) for Stephan, the effect for math is not reliable, while for reading and for math and reading combined, conventional levels of statistical reliability are reached irrespective of whether the mean is computed with or without correction for the length of desegregation; and (4) for Wortman, the effects for reading and for reading and math combined both differ from zero even when we consider only the small sample of studies with pretest adjustments.

These statistical tests are themselves partly problematic. In all cases except Miller, the analyses are based on a sample of comparisons. But since some studies produce more than one estimate of effect size, the assumption of independent errors may not be met. This particular problem does not occur in Miller's analysis. There, the small sample of studies increases the dependence on the assumption of a normal distribution of effect sizes. But as the difference between the various measures of central tendency indicates, the distribution of effect sizes may not be normal. Hence, all the statistical test results reported above (and in Walberg) should be treated with some caution. As they stand, they suggest that neither the mean reading effect nor the mean effect for reading and math combined is due to chance.

However, to complicate matters, it is not likely that the medians and modes differ from zero. The standard error of a median is normally set at 125% of the value of the standard error of the means from the same distribution, reflecting the greater instability of medians. By this criterion, no medians reliably differ from zero for reading or for reading and math combined. No estimate of the reliability of modes is necessary since they hover so closely around zero. However, the medians and modes are based on so few cases that estimates could shift radically once a dozen new values are added to the distribution.

If the population of effect sizes is indeed skewed, it is not clear which measure of central tendency is to be preferred. The mean represents national impact at some abstract, aggregate level, and is of use to those persons and groups most interested in gaining a national perspective on education and society. The mode represents what should happen to the typical school, and so may be of most interest to any school district or judge considering desegregation, especially if the district in question deffers from those where desegregation has produced large impacts in the past--characteristics we shall explore below. For any commentator willing to assume that the distribution of effect sizes in the population approximates the (unclear) sample distributions we have obtained, it is important to decide at a high level of consciousness on the different utilities implicit in different measures of central tendency.

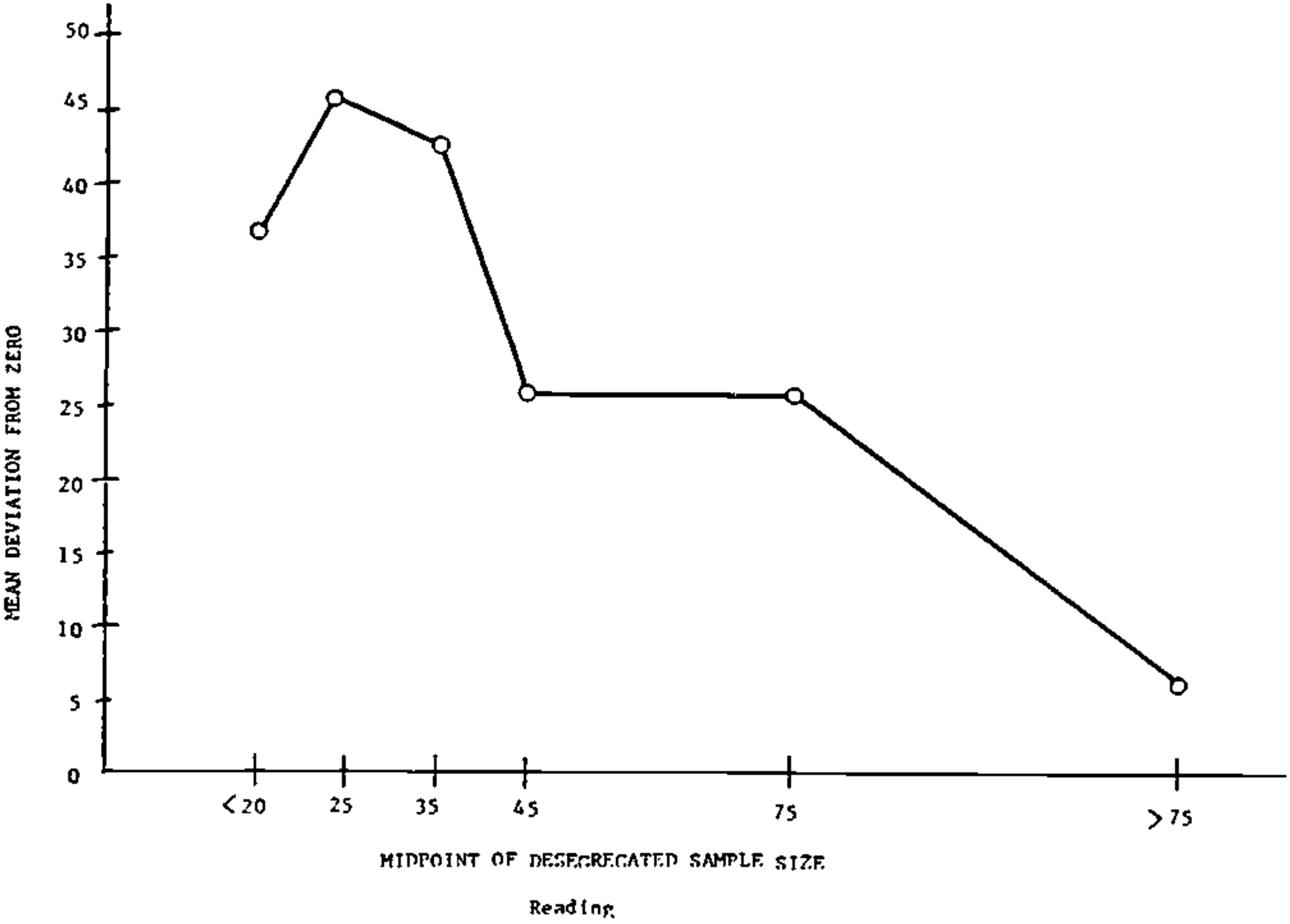
5. Why Do Some School Districts Show Larger Gains in Reading? The skewness in the distributions indicates not only that the mean may be a misleading measure of central tendency, but also that it might be productive to probe the reasons why some school districts are outliers. Discovering what they did to achieve larger gains could, for instance, be used to develop specific guidelines for desegregation plans, which school districts could then select if they believed they were suitable

for their schools. But since desegregation is an amorphous set of activities that differs from site to site, and since we have so few studies, no one should expect a definitive answer to the question of what characterizes school districts with large reading gains. At most, one should expect grounded hypotheses to emerge. Our discussion is in two parts: which were the districts with large gains; and what differentiates them from other districts?

a. Which Were the School Districts with Larger Reading Gains? Before probing substantive reasons for high reading gains, it is important to raise three methodological issues that reduce confidence in judgments about the identification of valid outliers. The sample sizes in the studies under review vary considerably, from 12 desegregated children in Zdep to over 1,000 in Sheehan and Marcus. Several panelists analyzed the relationship between sample size and effect size, concluding that smaller samples tended to produce larger estimates but that the relationship was not reliably different from zero. Considering classical sampling theory in isolation, we would not expect sample sizes to be linearly related to effect sizes without transformation of the original metrics. In a normal distribution with mean equal to zero, we would expect smaller samples to produce larger estimates, but in equal proportions each side of zero. This is equivalent to a negatively accelerated decay function when plotting effect size against sample size, irrespective of the sign of the effect. Figure 5 presents the mean reading effect size, free of sign, for studies with desegregated samples of 20 or less, between 21 and 30, between 31 and 40, 41 and 50, between 51 and 100, and over 100. An overall relationship is apparent that might well be of the expected quadratic form, though with such a small sample of studies it is hard to be sure. More important, though, is that with such a sample of studies, it is possible for more of the studies with smaller samples to fall on one side of the mean than the other. If we take the studies identified from Miller's estimates as outliers we note the following individual sample sizes in the desegregated groups for analyses of reading: Anderson (34), Baker (36), Syracuse (24), and Zdep (12). This is a total of 106 desegregated children. Since a total of 2812 were studied for reading, the outliers responsible for the higher mean estimates constitute about 4% of the total sample of desegregated children, but are about 25% of the studies Miller analyzed (4 of 17). If we add Rentsch to the list of outliers because analysts other than Miller and Stephan place him there, then the outliers represent 30% of the schools studied (5 of 17) but only 7% of the children.

A second methodological reason for caution in substantively pursuing why some school districts have large gains is also related to sampling instability. If we were to define positive outliers in terms of their gains in both reading and math, few of the outliers would be the same as when reading was considered alone. Thus, the unweighted gain in Anderson,

Figure 5: Relationship between Sample Size and Magnitude of Effect Size Irrespective to their Sign



using Miller's estimates, was .70, in Beker was .19, and was .26 in Zdep. (it was .035 for Rentsch in Miller's analysis.) When a joint criterion is used to define outliers, only Anderson clearly emerged. Indeed, the three other studies had negative estimates for math. Pursuing the instability theme further leads us to note that the second largest negative outlier for reading (Van Every, -.17) is based on a desegre- gated sample of only 20, and the math estimate is +.54. We are not arguing that desegregation should have affected both reading and math. We are only suggesting that we would be more confident of having identified valid outliers if reading and math gains were correlated among the potential outliers.

The third methodological issue concerns how effect sizes were computed. All the panelists are commendably sensitive to the need to control for differential growth rates between the nonequivalent desegregated and segregated control groups, and all go about the task in similar--but not quite identical--ways. The adequacy of statistical adjustment for selection-maturation depends on many factors, including the (unknown) true selection difference, the reliability of measures, the comparability of within-group regression lines, etc. In meta-analysis, the hope is that , across all the studies examined, the inevitable imperfections in the analysis of any one study will even out so that the average bias due to selection-maturation will be zero. However, there is no presumption that the bias will be zero in any single study. Yet in analyzing outlier effect sizes, one has to assume that the average selection and selection-maturation bias among the outliers is zero. However, one might easily have capitalized on chance and have isolated the subset where adjustment has been the least adequate. Indeed, in four of the five outlier cases the desegregated children outperformed the segregated initially, and in the other cases the means were essentially identical.

Thus, the possibility cannot be ruled out that the outliers reflect: (1) sampling instability due to small sample sizes; (2) sampling instability that makes high reading gains not synonymous with general achievement gains; and (3) an underadjustment for initial group differences in reading achievement. It is within the limitations afforded by these three points that I now examine substantive characteristics of the outliers for reading.

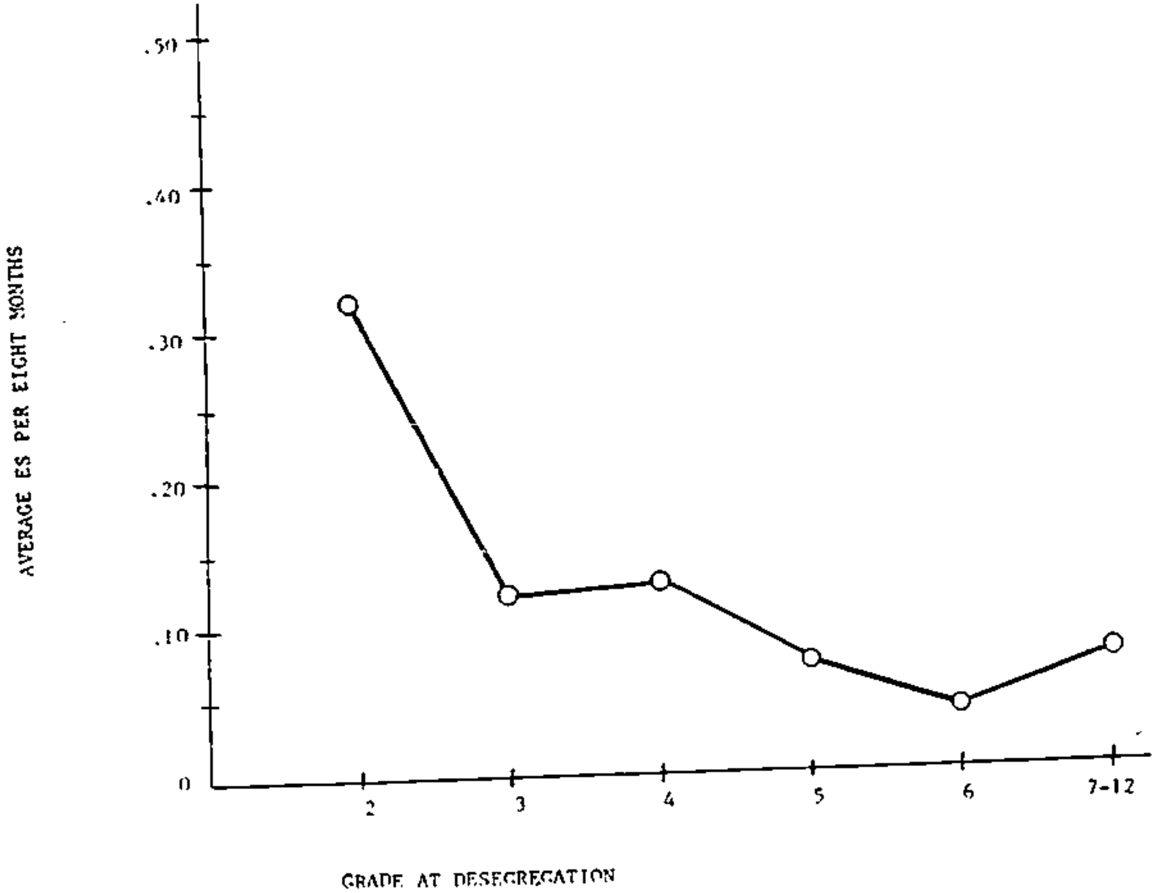
b. The Characteristics of Outlier School Districts. As previously discussed, one characteristic of the outlier school districts on Miller's list is that they evaluated longer periods of desegregation--up to three years in some cases. The relationship between effect sizes and length of desegregation is not clear due to sampling instability, with all the panelists who tackled the issue concluding that effect sizes seem larger in the five studies with two years of desegregation than in the nine studies with one year of desegregation. However, estimates seem to be lowest of all in the three studies with three years of desegregation. Since two-year studies predominate among the studies with larger effects in Miller's Table 2, it suggests that effect sizes may be related to the amount of desegregation that has taken place.

The predominance of two-year studies among the districts with larger effects also leads me to prefer Stephan's estimates for defining outlier school districts. But to use his data, I averaged his estimates across grades to give a single reading mean per study. The outliers fall into two groups: Anderson (.49), Syracuse (.58) and Zdep (.66) are in the one, and Klein (.23) and Rentsch (.22), in the other. Even listing these outliers raises once again the specter of instability, since Klein would not be an outlier for Miller, while Beker would be for Miller but not for Stephan!

Two substantive factors are associated with Stephan's larger effect sizes. One factor concerns when desegregation takes place. Figure 6 shows effect sizes per eight months of desegregation plotted against when desegregation began. The latter values are taken from Wortman rather than Stephan, since the information about grades in Stephan's Table 1 appears to be based on the grade at which desegregation began in some cases and on the grade when it ended in others. Figure 6 shows a clear negatively accelerated decay curve, with larger effects the earlier the desegregation. None of the panelists obtained effects of grade on achievement that were as clear cut as this, probably because they computed linear relationships, truncated at inappropriate grade levels, did not adjust effect sizes for the length of desegregation, or they assessed the grade of children when the study ended. Figure 6 suggests that at second grade, a gain is obtained of about .30 standard deviation units per eight-month year--though this estimate is based on only four studies --that at the third grade the gain is .12 (five studies), while it is .14 at the fourth grade (based on the nine studies).

In trying to explain why a small set of school districts produced large reading gains that skewed the distribution of effect sizes, it is important to probe whether the desegregation was voluntary or mandatory. According to Crain's report in this volume, all of the school districts I have identified as positive outliers had voluntary programs. This is perhaps not surprising, since the programs were voluntary in 15 of our 19 studies. For reading, only three school districts showed overall negative effects in Stephan's analysis--Sheehan & Marcus (-.07), Smith (-.01) and Van Every (-.12). The first and last of these were mandatory programs. Of the two other mandatory programs in the panel's sample, the study by Carrigan was omitted from some analyses but, when aggregated across grades, it produced a small negative effect.

Figure 6: Relationship between Grade Level at Desegregation and Mean Effect Size per Eight Months of Desegregation



The other mandatory study produced a trivial gain of .02 across grades (Evans). It is clear, then, that mandatory programs were not associated with reading gains but that voluntary programs were.

However, the relationship between effect size and the voluntary/mandatory nature of desegregation could only be considered causal for these four cases of mandatory desegregation if all other interpretations of the relationship could be ruled out. However, two of the studies--Evans and Sheehan & Marcus--were done in Texas, were the only ones to use the Iowa Test of Basic Skills, and were two of the only three studies of desegregation activities that began in the 1970's. (The other study with apparent negative outcomes--Van Every--took place in Flint, Michigan, began in 1969, used the SRA test, and had very small samples.)

Just as it would be wrong to conclude with confidence that mandatory programs produce no gains in reading, so it would be wrong to conclude from the panel's core studies that desegregation beginning in the earlier grades results in larger positive gains. There are signs of each relationship, but with only four mandatory programs and four second grade samples it is inevitable that we have not made heterogeneous all the sources of irrelevancy that might have produced spurious results. The reality is that if the sample sizes of studies is too small to permit a meaningful analysis of central tendency across 19 studies, it is even less appropriate for conducting responsible internal analyses to try to explain why some school districts seem to have achieved larger effect sizes than others.

This is true, not only of the potential explanatory factors analyzed above, but also of other factors about which individual panalists have speculated. Stephan points out that studies conducted at an earlier date tend to show larger effects, while Miller suggests that school districts with larger effects may have introduced enrichment programs at the time desegregation occurred and may have had smaller percentages of blacks in the desegregated classrooms. With the small samples on hand, it is inevitable, first, that no strong probes of the impact of such moderator variables is possible; and, second, that many interpretations remain to explain why some districts achieved particularly large positive or negative gains.

The points we want to stress are that: (1) the form of the distribution of effect sizes is not clear either for the population of school districts that have desegregated or even for the small sample of districts we have analyzed; (2) there may be districts that benefitted more from desegregation than other districts--but if so, it is not clear whether they are outliers for irrelevant methodological reasons (small sample sizes, unstable measures, or initial group achievement differences not completely adjusted away) or for relevant substantive reasons; and (3) of the relevant substantive reasons, several are contenders as explanatory constructs, but their unique contribution cannot be unconfounded from the contribution of the factors. The factors at issue include: the child's grade at desegregation, the number of years of desegregation, whether the desegregaton is voluntary or mandatory, the percentage of whites in the class, the copresence of desegregation and new enrichment programs, and the year in which desegregation took place.

6. Summary of the Reanalyses. A casual reading of the panelists' papers leads to the four conclusions mentioned earlier that are based upon the panel's 19 studies and seem quite consonant with the findings of prior meta-analyses by Krol and by Crain & Mahard that involved larger samples. These conclusions are: (1) desegregation does not decrease the achievement of black children; (2) it probably does not increase math achievement; (3) it probably raises reading scores; and (4) the increase in reading scores is somewhere between .06 and .16 standard deviation units or about two and six weeks. These last estimates were computed from 17 studies, about half of which dealt with a single year of schooling, and then usually the first one after formal desegregation began.

Our own analyses corroborate the first two of these findings. We continue to find no evidence that desegregation decreases achievement or that it increases achievement in math. Our differences involve the conclusion about reading. The present analysis suggests that whether there is an effect or not depends on the measure of central tendency used, with statistically reliable results emerging from mean gains but not from median or modal gains. The implication of the lower median or modes is that the mean differences are found, not so much because the "average" effect of desegregation on reading is positive but because--in the panel's sample at least--some school districts made atypically large reading gains that skewed the distribution of effect sizes.

It is therefore difficult to make an estimate of the size of the reading effect. There is one range estimate for the mean (between .13 to .16 when the same 17 studies from the panel's 19 are used with each analyst's own effect size computations--see Table 2), another range estimate for the median (.00 to .08 irrespective of the samples used--see Table 1 or 2), and yet another for the modal effect (between -.05 and +.05--see Tables 1 and 2). Combining the reading and math effect sizes makes no difference to the conclusion that central tendency values differ. The estimated means vary between .07 and .16 for the 17 common studies; the study medians vary between .00 and .06; and the mode falls between +.05 and -.05.

Why do some schools achieve unexpectedly large reading gains? With so few studies, this question cannot be answered in any definitive way. There are at most indirect suggestions that such schools may have desegregated in the 1960's, had voluntary plans, included the earlier grades in their evaluation design, been studied for longer time periods, have had a higher percentage of white children in desegregated classrooms, and may have introduced enrichment programs at the same time as desegregation. Such variables could have had independent or joint impacts, and it is inevitable that other variables could be thought of that should be added to any list of possible explanations of why some districts gained so much more than others in reading. Among the possibilities is chance, for it is noteworthy that the outlier studies had smaller sample sizes and that, with the exception of Anderson, the districts with the largest gains in reading were not the districts with the largest gains in math. While it is not necessary for desegregation to impact on both--and Stephan gives an ex post facto rationale for why desegregation should affect reading but not math--we would be more confident of having identified valid outliers had there been more of a consistency in gains between reading and math.

If the present analysis had not taken place, there would have been what I interpret to be an impressive consistency of results for reading and math combined. When they defined better studies their own way and combined all measures and grades, both Krol and Crain & Mahard reached comparable mean estimates of .10. (For Crain & Mahard, the value is derived from the combined results of their randomized experiements and their two longitudinal designs with black segregated controls.) Using their own preferred set of studies and considering math and reading only, the present panelists arrived at estimates varying around this. Armor obtained .04, Miller .12 and Stephan .14, and Wortman .17 when his two strongest designs were weighted and averaged based on part of his sample of 31 studies. These estimates are generally higher than the values of Krol and Crain & Mahard, but not by much. Indeed, I suspect that few commentarors would find much of a difference between a gain of one month and of one and one-half months (.10 versus .15).

The present analyses have muddled these waters by suggesting that the means above are noticeably higher than their corresponding medians or modes and by further suggesting that the choice of a measure of central tendency depends in part on knowledge of the distribution of effect sizes in the population. But with such a small sample, the true distribution cannot be confidently ascertained. For those who accept my analyses, I have substituted a low degree of certainty about the effects of desegregation for the higher degree that used to pertain but that depended on distributional assumptions which may be wrong. Social science analyses often increase uncertainty, and this is to be preferred to a premature certainty about something wrong or misleading. However, it is even more preferable to reduce quickly new sources of identified uncertainty. In the present case, this means examining the distributions obtained by Crain & Mahard (1983) for their better studies to see if they are skewed.

7. A Comparison of the Present Results with Crain & Mahard. Crain & Mahard (1983) insist that the effects of desegregation are best assessed from randomized experiments and from studies where desegregated schooling begins at kindergarten or grade one so that the child has never known segregated schooling. When the randomized experiments and the studies with kindergarten and first grade samples were studied separately, Crain & Mahard obtained estimates of .30 in each case. They therefore interpreted this as the best estimate of the effects of

desegregation on the achievement of black children. Such an effect is moderately large by many of the (arbitrary) standards used for assessing the effects of educational interventions, as Walberg's essay in this volume attests. It is certainly a more optimistic value than obtained in the meta-analyses reviewed here. Hence, we will consider the estimates of Crain & Mahard in some detail.

It is clear that their estimates decrease to some extent when we consider medians and modes rather than means. Crain kindly supplied me with the distribution of effect sizes for the seven comparisons involving randomized experiments, with Zdep omitted. The mean was .27, the median .24, and the mode could not be computed. For the kindergarten and first grade samples evaluated using before-after designs and black segregated control groups, the mean based on 17 comparisons was .31, and the median and mode were each .26. I do not know what the mean, median and mode were for all the studies and all the grades with before-after measures and black controls. Nonetheless, the data above suggest that the medians and modes do not reduce to zero in the studies that Crain and Mahard prefer for estimating the effects of desegregation.

Unfortunately, the results of Crain & Mahard are not easy to interpret as estimates of generalized causal impact. First, nearly all the randomized experiments were part of Project Concern and so offer little comfort as to the generalizability of effects. Also, with so few degrees of freedom in the analysis of randomized experiments, it is not likely that the mean effect reliably differs from zero. Second, only one of the kindergarten and first grade samples of Crain & Mahard was included in the present panel's sample--Carrigan--despite the specification of both Crain & Mahard and the present panel that before-after designs and black controls characterized better studies. This discrepancy in the number of comparisons presumably occurs because of differences in strategies used to estimate standards deviations and--principally--because Crain & Mahard were willing to accept pretest measures that the present panel would not accept because it required that pretest and posttest measures tap into the same conceptual domain. For understandable reasons, the pretest measures of very young children tend to reflect "academic readiness" rather than the academic achievement that is assessed at the posttest. If the usual selection bias operated and the children attending desegregated schools were more able or more motivated than their segregated counterparts, then the reduced pretest-posttest correlation caused by differences between the readiness and achievement measures would probably result in overestimating the effects of desegregation in each study (Campbell & Boruch, 1975). Consequently, it is unlikely that valid estimates of the effects of desegregation were obtained with the kindergarten and first grade samples of Crain & Mahard, though the authors have indeed identified a significant issue. After the first generation of desegregation in a district, no students enter desegregated schools from segregated ones--nearly all begin and end their schooling in desegregated classes. Consequently, it is of special importance to learn how desegregation is related to the achievement of very young children.

The estimate of Crain & Mahard that most closely approximates the work of the present panel is based on all grade levels, all outcome measures, before-after designs, and black control groups. As mentioned earlier, the estimate they obtained was .10, and this is much closer to the panel's estimate than the probably inflated value of .30 provided by studies of kindergarten and first grade children for which initial differences were not well-controlled. However, nothing in the present panel's work specifically refutes an implicit claim--in Crain & Mahard--that desegregation may have larger impacts at younger grades. To say that .30 may be inflated is not to say the true value for the youngest children is .10. The issue of grade differences in effect sizes has not been solved by either the present panel or Crain & Mahard, and must remain an issue for further research.

INTERPRETATION

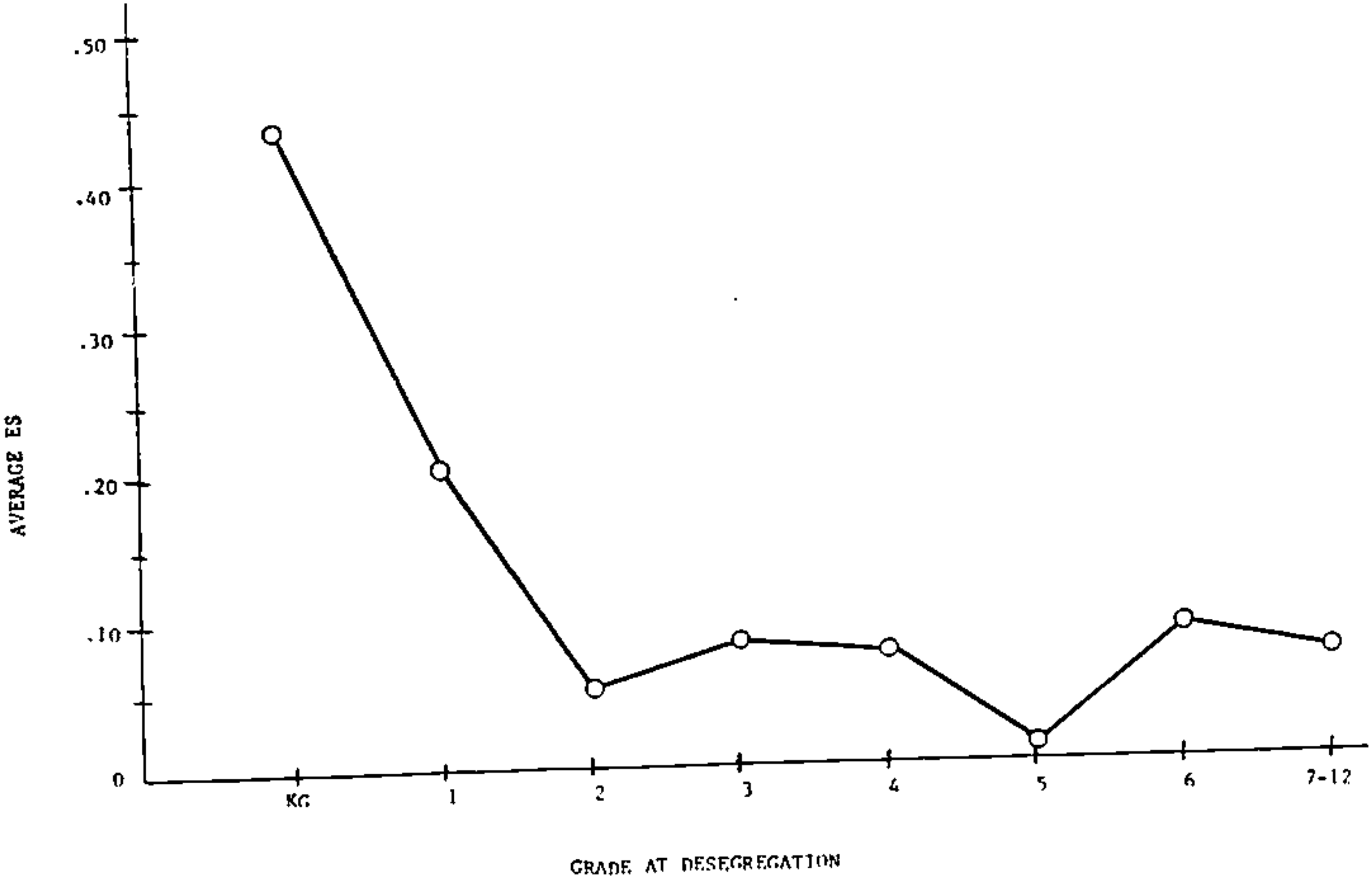
I want now to interpret the meaning of both the absence of gains in mathematics and the presence of reading gains of between two and six weeks. To do this, I broach two issues. First, I ask what implications the findings have for various stakeholder groups, and in so doing I also explore how generalizable the findings are beyond the 19 studies examined. Second, I ask what implications this meta-analysis project has for theories of research synthesis.

1. Stakeholder Analysis

a. Protagonists of School Desegregation. The analyses I have presented might give some comfort to protagonists of school desegregation, particulary those who support it for reasons of equal access, the improvement of race relations, or the enhancement of self-esteem rather than for reasons of academic achievement. For such protagonists the crucial finding from all the analyses of all the scholars is that school desegregation does not decrease the achievement of black children. If it did, this would represent an undesirable side effect of desegregation with which protagonists would probably have to deal ethically, ideologically, and politically. My guess is that it is more difficult to argue that a decrease in achievement is of no consequence than it is to argue that the absence of an increase is of no consequence. Unintentionally decreasing achievement would be a worrisome side effect of desegregation that no protagonist could ignore.

Protagonists of school desegregation can also take some succor from an as yet imperfectly corroborated trend in the data. This is that achievement gains may be larger in younger children who have not had to go through as long a prior experience in segregated classes. Indeed, one of the major points in Crain & Mahard--that we could not independently test--is that achievement gains are greatest of all if black children have never been segregated. This is a very important point, for many of the advocates of desegregation view it as a means of providing desegregated--or preferably, fully

Figure 7: Relationship between Grade Level at Desegregation and Average Effect Size in Crain and Mahard (1983)



integrated--education to all children for all of their school career. From this perspective, the group of children who start out in segregated schools are not the group of greatest interest. Of more concern are those who have never been segregated and will never experience the historically circumscribed difficulties associated with being among the very first children to transfer within a desegregated school district. Such pioneers move into environments that are novel, not only for them but also for teachers, administrators, parents and local leaders. Because of the novelty, more mistakes are likely to occur than is the case at a later date when new cohorts of children come through the system, and teachers, administrators and parents should have benefitted from earlier mistakes. Later cohorts might be expected to benefit more from desegregation, both because they have never known segregated schooling and because the school personnel are more experienced with education in mixed racial settings.

Protagonists of desegregation might also note that over half of the studies examined by the present panel involved only one year of desegregation. Moreover, the typical fall-spring testing sessions involve less than a complete school year. Thus, most of the studies involved only a small fraction of the total time that children experience desegregation, especially if they enter desegregated schools in the early grades. Protagonists of school desegregation might wonder if its full impact has yet been evaluated and they may point to the larger effects in two-year studies to suggest that the cumulative impact of desegregation may be much larger than its first year effect. The major problem with this argument is that the studies testing three years of desegregation produced no effects. Consequently, protagonists of desegregation would have to discredit the three-year studies in order to make the case that desegregation has not yet been tested at its presumptively most efficacious. However, it is not difficult to discredit these studies since they are only three in number and they undoubtedly differ from the majority of studies in many ways that are correlated with lower achievement gains.

b. The Perspectives of Antagonists of School Desegregation. The present analyses should bring most succor to antagonists of school desegregation. Where before they would have had to acknowledge the gains in reading caused by desegregation and would have had to argue that their practical implications are trivial--as Armor has done in his present essay--antagonists can now point to analyses which suggest that there have been no real gains in reading because of desegregation in most school districts. This involves a shift in the argument--from how meaningful the obtained reading gains are considered to be, to whether there are any gains at all with value worth debating. But although the medians and modes in Tables 1 through 5 could be used by antagonists of school desegregation, I have tried to stress how unstable these estimates are and how much they might be changed by adding just a dozen more cases to the distribution of effect sizes.

Antagonists of school desegregation can also point to the opaque trend in the data for mandatory programs to result in zero effect sizes and for larger effects to be found with voluntary programs. Few antagonists of desegregation oppose plans in which local authorities agree to desegregate and receiving schools voluntarily accept pupils who volunteer to go to the receiving schools (or whose parents "volunteer" for them). The objection is to mandatory desegregation which, in both my analysis and Stephan's, produced no reading or math gains. (This comparability was achieved despite the fact that Stephan classified only two of the panel's studies as mandatory, whereas using the essays in this volume by Crain and Armor, I classified four as mandatory, although one was by Carrigan.) However, little confidence can be placed in the idea that mandatory desegregation plans cause no reading gains. Given the small number of studies overall, and of mandatory studies in particular, the mandatory/voluntary distinction was correlated with the year desegregation took place, the test used to measure achievement, the region of the country (two studies were in the Dallas/Ft. Worth area), and was probably also correlated with many other factors that would emerge as soon as one examined in detail the specifics of the mandatory desegregation studies by Sheehan & Marcus, Evans and Van Every.

Antagonists of school desegregation can also point to the paucity of clearcut evidence about desegregation plans that will raise school achievement. Protagonists of school desegregation, and persons whose job it is to plan the desegregation effort in a particular community, want to know what types of desegregation will be effective. They prefer this specific question to the more global: "How effective is desegregation in general in raising achievement?" All the parties concerned with desegregation research realize that there is no standard desegregation treatment, but many of the protagonists of desegregation hope to discover a set of activities that, when implemented in newly desegregated schools, will raise achievement, among other things. The present analysis has pointed with little confidence to some possible elements of effective desegregation plans. But nothing in the list of elements is new, and after the panel's reviews, nothing is better "proven" as a causally efficacious element of desegregation plans than was the case before. Antagonists can point,therefore, to the saliency the present review gives to the continuing uncertainty about the elements of desegregation that enhance achievement. This is not to say that the present meta-analysis proved all-or even most--of the prospective causal elements, or even that it probed the better corroborated among them. All we maintain is that it probed some of them, but failed to make us any more confident that we know how to put together desegregation plans that will raise achievement in reading and math.

c. Persons Planning Desegregation Activities. Irrespective of their personal beliefs about the desirability of desegregation, mandated or otherwise, there are some groups of persons who have to plan desegregation activities. One such group consists of judges, civil servants, consultants, and school district officials who develop desegregation plans for school districts or metropolitan areas. Such persons want to know about the types of desegregation plan, or the major elements within an overall plan, that will produce the kinds of outcomes they most value from desegregation. The present panel's work provides nothing of substance to help such planners. It might, however, make a minor contribution to undermining their morale, for the difference in outcomes between the means, medians and modes suggests that the effects of their labors on achievement are likely to be minimal, at least in the short term and to the extent the backward-looking analyses on which this review is based are pertinent to the immediate future.

This last point is crucial. For many theorists of evaluation, its function is less to summarize what has happened in the past and more to discover what might be effective in the future. In this context, it is worth noting that the major difficulties with meta-analysis concern the possihility that the bias in one direction may be greater than in the other across all the studies under review. The panelists dealt exhaustively with biases that might lead to false conclusions about whether the relationship between desegregation and learning gains is causal, but few of them considered biases that limit the generalizability of findings and hence their presumed utility for planners. In fact, 16 of the 19 studies were begun in the 1960's, and only one is later than 1975. The dearth of later studies is striking, and Armor's essay contains an important paragraph expressing indignation that so few evaluations of school desegregation were undertaken in the 1970's, a decade characterized by so many large-scale evaluations in other areas within education. Most of the 19 studies under examination were dissertations or local efforts by the staff of a school district. This may explain why the sample sizes are so small, the documentation of desegregation activities so meager, and the measurement plan so sparse.

Another constant bias is obvious. The panel was constrained to examine how desegregation impacted on the achievement of black children. Yet for most planners, achievement does not exist in a vacuum. The utility of the achievement gains caused by desegregation can vary in meaning depending on whether the desegregation activities in question also reduce or widen achievement gaps between blacks and whites, are or are not accompanied by an increase or reduction in interracial prejudice, are or are not accompanied by white flight, are or are not associated with self-esteem gains, are or are not associated with community support, are or are not related to changes in real estate values. are or are not associated with the founding of magnet or lab schools, etc. By examining just school desegregation and black achievement, much of the interpretative context vital to planners is lost.

A second group of planners is composed of teachers, both those contemplating desegregation and those already teaching in desegregated classrooms. In theory, research could be of help to those in identifying practices they can implement that will improve the functioning and results in classrooms. However, the present meta-analytic efforts do not speak to such learning needs. The teacher's needs are more micro than macro, more concerned with process than outcome, and with explanation than descriptive causation. The question on which the panel worked is a question that meets the interests of central government officials with responsibility for oversight more than it meets the interests of those who must plan for desegregation in specific school contexts.

d. Persons Honestly Seeking To Learn What Desegregation Has Accomplished. The panel's papers help those who would honestly understand what desegregation has accomplished by questioning the utility of so global a label as "desegregation." Miller's analysis shows that, after the mean effect size is accounted for, more variance remains than is due to chance. This suggests that systematic forces have to be taken into account over and above whether desegregation took place if there is to be any reasonable prediction of effect sizes. Elementary consideration of the decentralized structure of educational decision-making suggests that desegregation plans will differ from location to location and that, even where they appear similar on paper, there will be local adaptations to suit local conditions. From the perspective of someone seeking to learn what desegregation has achieved, elementary questions need to be asked: "What does desegregation mean?"; "What are the criteria that should be used to create clusters of desegregation activities?"; and "How well do the different clusters or types of desegregation predict differences in achievement outcomes across districts?" At present, persons interested in learning about school desegregation are more likely to have learned to identify the more pertinent questions than they are to have learned answers to these questions.

But there are some persons interested in the effects of desegregation, very globally conceived, most of whom are government officials with oversight responsibility, journalists, or scholars. The present essay may help sensitize them to the possibility of considerable differences in effects from district to district and to the possibility that, across all districts, effects may be highly variable and even skewed. The possibility of skewness might present them with a problem. Although the mean represents the global impact of desegregation painted on a broad national canvas, it is of no comfort to judges and school districts contemplating desegregation or to teachers worrying about how to handle a racially mixed class. For some of these people, the mode is more immediately meaningful than the mean. It may be less meaningful in the future, of course, if (1) there really are outliers, (2) the causes of large gains can be explained, and (3) school districts can adopt the causal elements present in the schools with large effects. But we do not yet know what these elements are. In the absence of such knowledge, the differences between the means, medians, and modes highlight anew the conflicting information needs of the many groups in the national educational system who have a stake in desegregation. The differences are most apparent (1) with respect to what should be evaluated--desegregation in general, a specific type of desegregation plan, the particular plan in a particular district, or elements within plans?; and (2) with respect to what should be assessed--achievement, school discipline, race relations, self-esteem, enrollment figures, local tax support for education, local political support for desegregation, home values, etc.? But the differences in information needs are also apparent with respect to (3) which measures of central tendency is most appropriate. Different measures speak more to the interests of some stakeholders than others.

2. Theories of Research Synthesis. The present panel represents a unique attempt to probe to what extent experts with three different presumed commitments would converge on a common answer about how desegregation has affected the achievement of black children. Crain and Wortman had already concluded in review articles or papers that desegregation increased achievement; the opposite conclusion has been drawn by Armor and Miller; while Stephan and Walberg had published on the issue but had taken more neutral stances, although Walberg has given court testimony largely opposed to desegregation. The hope was to achieve a common estimate of effect size despite the different commitments, based on a theory that the results would be more credible, and perhaps even more valid, if they could be replicated across the heterogeneity associated with the analysts' prior professional commitments.

In general, the effect sizes for math and reading combined did reflect the prior commitment. Highest were those of Wortman (.17), and Crain, who stressed the results from his kindergarten and first grade samples and from the randomized experiments he studies (.30 for all outcome measures combined). The next highest estimate was from Stephan (.14 without corrections for length of desegregation), and lowest of all was Armor (.04). The person least fitting expectations was Miller, whose .12 value was intermediate.

Actually, the theoretical rationale for pluralism of analysts was only partially realized, given the decision made before the panel met to restrict the meta-analyses to "good" studies and to use Wortman's prior work to generate that list. One of the major points in meta-analysis where ideology and other commitments enter in is when relevant studies are selected for analysis. Panel members were free to suggest studies for the core list, and Armor succeeded in having two studies added that had negative effect sizes (Sheehan & Marcus, and Walberg). He also made a strong and persistent case for excluding Rentsch and including Carrigan. But few considered calls were heard to add other studies, even though Crain had a list of 93 that he and Mahard considered relevant, more than half of which may have been randomized experiments or longitudinal designs with segregated black control groups. In retrospect, the decision to restrict the selection criteria to a common set rather than let the panelists select their own, and the failure to assess each of Crain's 93 studies according to the panel's criteria of adequate methodology, may have unnecessarily restricted both the sample of studies and the heterogeneity in assumptions on which the theory behind the use of multiple panelists depends.

It is not difficult to see why the decision was made to restrict the meta-analyses to "better" studies. After all, Krol has found smaller estimates with his "better" studies, as also had Wortman, King and Bryant. But Crain obtained larger estimates with his "better" studies. Obviously, chance differences in the studies available, or differences of opinion about what makes better studies, may have contributed to the apparent puzzle about whether superior methods were associated with larger or smaller effect sizes. Another point is also worth keeping in mind. Although one of the rationales for pluralistic panel members was the credibility and validity afforded by convergence, a second rationale is that divergence in their results might serve to force out the differences in assumptions between advocates and opponents of desegregation, thereby sharpening the focus for future research. Yet the likelihood of such differences being forced out is presumably greater the more freedom panelists have to select studies for review.

Another decision that was made before the panel convened was to use meta-analysis. This technique depends most heavily on the assumption that the average bias is zero with respect to threats to internal, external, construct, statistical conclusion, or any other type of validity (Cook & Leviton, 1980). This assumption is usually dealt with in either or both of two ways. First, a subsample of studies is isolated for which the assumption is made that the bias is zero, and the estimate from this sample is then compared to the estimate for the remaining subsample where bias might be a problem. If there are no differences in the estimates, the conclusion is drawn that the biasing force in question has not operated. The second strategy is to assume the source of bias away by postulating that the total sample studied is heterogeneous with respect to the threat in question. This last assumption is more credible the more the sample differs on irrelevancies correlated with the major outcomes.

Desegregation research is problematic for the meta-analyst since Wortman has shown that studies without control groups might be biased, and few analysts are willing to use norms or white children as "control groups." The need for control groups entails that few studies will meet minimal methodological characteristics. The sample of studies will also tend to be highly variable, given the wide range of desegregation activities in the decentralized education sector and the wide range of children, grades and times studied. Consequently, small samples of possibly abnormally variable estimates will be meta-analyzed. It is difficult to imagine arriving at confident estimates of distribution and central tendencies in this situation; and it is also foolhardy to expect to break the data down in multiple ways so as to examine the correspondence in estimates across different types of desegregation activities, different years when desegregation began, different regions of the country, etc. Consequently, to rule out threats one has to rely on there being "enough" variability in region, year of study, type of activities implemented, etc. But given the small samples, it is not easy to be confident of "enough" heterogeneity in conceptual irrelevancies, hence the low level of confidence I have placed in most of my own conclusions and those of the panelists.

These meta-analytic endeavors point to another problem with the method that overlaps with the problems in using small samples to estimate populations that may be complex and highly variable. Once one has postulated that a skewed distribution may be present, the guiding question becomes the explanatory one: "Why are there outliers?" Explanation is not a strong point of meta-analysis. To explain, presumes that we have measures of the potential explanatory constructs for a large sample of studies. Rarely is this the case with meta-analyses, for their availability depends (1) on the extensive measurement of what is implemented as part of a treatment--in the desegregation studies examined, little was available from reports to help with this; and (2) on the extensive measurement of causal micro-mediating processes. For desegregation and reading, such measurement might include, but not be limited to, the assessment of dominant language patterns inside and outside of classrooms. But the sample size of studies with such measures might be expected to be low since the relevant hypothesis about language patterns had not been developed when the earlier evaluators did their work. Indeed, the theory developed because of their work and the anomalies in the data which the work revealed. Since the number of studies with adequate measures of potential explanatory variables will often be low in meta-analysis for reasons of cost and because of the dynamic, evolving nature of theoretical explanatory constructs, meta-analysis will rarely result in confident explanation. This was certainly the case in trying to explain the outliers in Figures 1 through 4. Many potential explanatory forces were isolated, but none of them could be unconfounded from each other with the sample sizes and measures on hand.

Conclusions

My own reading of the panelists' papers and my own analyses lead me to the following conclusions about how school desegregation has influenced the academic achievement of black students. The conslusions are based on only about 17 studies, and their generalizability is unknown.

1. Desegregation did not cause any decrease in black achievement.

2. On the average, desegregation did not cause an increase in achievement in mathematics.

3. Desegregation increased mean reading levels. The gain reliably differed from zero and was estimated to be between two and six weeks across the studies examined. Only one panelist (Stephan) computed the reading effect per 8 month school year. His estimate is between five and six weeks of gain per year. But since none of the studies involved more than three years of post-desegregation research, it is not possible to compute the mean gain over a child's total school career in desegregated classrooms.

4. The <u>median</u> gains were almost always greater than zero but were lower than the means and did not reliably differ from zero. The <u>modal</u> gains were even less than the median gains and varied around zero.

5. The differences between the means, medians, and modes result because the distribution of reading effects appears to be skewed, with a disproportionate number of school districts seeming to obtain atypically high gains.

6. Studies with the largest reading gains can be tentatively characterized along a number of methodological and sustantive dimensions, including: small sample sizes, the study of two or more years of desegregation, desegregated children who outperformed their segregated counterparts even before desegregation began, and desegregation that occurred earlier in time, involved younger students, was voluntary, had larger percentages of whites per school, and was associated with enrichment programs.

7. None of the above factors can be isolated, singly or in combination, as causes of any of the atypically large achievement gains in reading that were obtained in some school districts.

8. The panel examined only 19 studies of desegregation, with most panelists rejecting at least two of them on methodological grounds. When the results for each study (or each comparison) are plotted for reading or mathematics, the distributions are based on so few observations that I could not accept the assumption that the obtained distributions closely approximate what the underlying population distributions are. Because of the small samples and apparently non-normal distributions, little confidence should be placed in any of the mean results presented earlier. I have little confidence that we know much about how desegregation affects reading "on the average" and, across the few studies examined, I find the variability in effect sizes more striking and less well understood than any measure of central tendency.

References

Campbell, D. T. and 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 results." In C. A. Bennet & A. A. Lumsdaine (Eds.), Evaluation and Experience: Some Critical Issues in Assessing Social Programs. New York: Academic Press, 1975.

Cook. T. D. and Leviton, L. "Reviewing the literature: A comparison of traditional methods with meta-analysis." Journal of Personality. 1980, 48, 449-472.

Crain, R. L. and Mahard, R. E. Desegregation Plans that Raise Black Achievement: A Review of the Research. Santa Monica, CA: Rand Corporation, June 1982.

Crain, R. L. and Mahard, R. E. "The effect of research methodology on desegregation achievement studies: A meta-analysis." American Journal of Sociology, 88, 1983.

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

Krol, R. A. A Meta-analysis of Comparative Research on the Effects of Desegregation of Academic Achievement. Unpublished dissertation, 1978. Ann Arbor, Michigan: University Microfilms (# 6907962), 1979.

Wortman, P. M., King, C., and Bryant, E. B. Meta-analysis of Quasi-experiments: School Desegregation and Black Achievement. Ann Arbor, Michigan: Institute for Social Research, 1982.

The Evidence on Desegregation and Black Achievement

David J. Armor
David Armor and Associates

The debate over the costs and benefits of school desegregation, particularly in its mandatory forms, continues unabated today, nearly 30 years after the fateful Brown decision by the U.S. Supreme Court. No issue has been more central to this debate than the question we address here: the impact of desegregation on Black student achievement.

Indeed, it is remarkable that this question remains in controversy today, considering the extent of school desegregation over the past twenty years and especially given the mandatory methods imposed by the courts over the past fifteen years. One wonders how many courts have ordered busing, how many agencies have allocated time and money, and how many Black parents have willingly sent their children to distant schools out of their neighborhoods, on the assumption that desegregation would yield academic benefits for Black children.

Obviously, more is at stake in desegregation policy than the academic progress of students. Desegregation is a highly desirable social policy regardless of its educational benefits, and many educators and parents will and should seek it despite research findings. On the other hand, it is one matter to agree that school desegregation is a desirable policy and quite another to make it compulsory regardless of other considerations. The moral imperatives permitting coercion in social policy make it unlikely, in my opinion, that our courts would have abandoned the traditional neighborhood school policy in favor of mandatory busing without the belief that they were actually benefiting the education of Black students. Why else would so many courts hear evidence, and so many legal journals publish treaties on this issue?

Aside from the legal importance of the achievement question, it does have immediate relevance to educational policy-makers, especially in this day of tight budgets. It is beyond dispute that we need programs to enhance minority achievement. The key question is, what kinds of programs? In recent years significant amounts of time and money have been devoted to improving racial balance in schools, justified in part by its supposed educational payoffs. Is this resource investment in fact yielding a fair return, in terms of improving minority achievement, or would other programs have greater impact? In other words, are racial balance activities cost-effective when compared to other available alternatives? If not, we should re-order our priorities and invest in programs that promise to work.

Finally, the issue of desegregation and Black achievement should have more than a passing interest to parents of Black children, who for years have borne the heaviest personal cost of desegregation by enduring long bus rides, separation from familiar surroundings, and curtailment of extracurricular activities. It is quite likely that, over the long run, Black parents' support of busing for the purpose of desegregation would lessen if desegregation was found to have minimal impact on their children's rate of learning.

For all these reasons, the National Institute of Education must be commended for bringing together, for the first time, a representative panel of experts to review the evidence and pass judgment on this difficult but vital issue. At the same time, more than one observer will be surprised at the small number of studies (19 in all) meeting the minimal scientific standards established by the panel, and perhaps shocked that only three of these studies have been conducted within the past ten years, when school desegregation has been at its peak.* It is almost as though educational researchers and their funding agencies--including NIE--believe that the issue is settled, or no longer important. It is clearly an important question, and even a cursory review of the available literature shows that it is clearly unsettled. Hopefully, this panel will offer a consensus judgment that will finally settle the controversy.

Before turning to the studies selected for review by the NIE panel, I will comment briefly on several other comprehensive review efforts. To a large extent the approach taken by the panel culminates an evolutionary sequence that can be observed in the previous attempts to grasp the essential truths in this varied and complex literature.

PREVIOUS REVIEWS

Much of the early disagreement over the desegregation and achievement issue stemmed from reliance on a single study, or on a small number of studies where variation in results and conclusions might be expected (e.g., Armor, 1972 and 1973; Pettigrew, 1973). Yet disagreement persists even among the comprehensive reviews, all of which investigate many of the same studies.

The first review to encompass a large number of studies was carried out by Weinberg (1970). Like his most recent review, Weinberg covers a lot of studies but makes little or no attempt to select studies according to their methodological adequacy for causal inference (Weinberg, 1977). As we shall see, his conclusion that desegregation significantly benefits minority achievement was undoubtedly affected by his failure to consider a study's scientific rigor.

The second comprehensive review by St. John (1975) made considerable progress over Weinberg. Not only was her study coverage broad, but she additionally classified studies according to the research design employed, allowing her to observe the relationship between methodology and the impact of desegregation. When St. John took design rigor into account, she reported that the evidence was mixed, preventing a firm conclusion about the benefit of desegregation for Black achievement. A later review by Bradley and Bradley (1978) did not expand on the state of the art over St. John. They did conclude that methodological flaws impaired the entire group of studies, and that nothing could be decided. A distinct advance was made in Krol's (1978) review, where he applied formal "meta-analysis" to 55 studies, as that phrase has been used by Glass (1978) and others. The technique Krol used involved two critical

*Different panelists, including myself, will take methodological exception to some of these studies.

steps that are lacking in previous reviews. First, studies were screened for minimal methodological adequacy (e.g., appropriate treatment condition and quantitative results) and coded as to a variety of conditions related to the type of research design and other study attributes. Second, achievement test results were converted to quantified standardized estimates by taking the ratio of test score means to their standard deviations. This allows estimates of the magnitude of segregation effects, as well as the impact of specific study characteristics on those effects.

Using this approach Krol concluded that the average effect of desegregation on Black achievement is .16 standard deviations, which (depending on the type of achievement test) amounts to anywhere between 1½ to 3 months of progress during an academic year. However, this effect was not statistically significant, and the effect for that subset of studies with a valid control group was only .10, which again was not significant. The major limitation for the Krol study is that the number of studies was small, and no adjustment was made for control group selection bias; that is, for treatment-control differences prior to treatment. Moreover, the way he estimated effects for studies without control groups assumed that a control group would experience no gain. This is not a tenable assumption for achievement test data, where some academic growth is the norm for most students at least through the 10th grade.

The most recent large-scale review was carried out by Crain and Mahard in several stages (1982). The latest version of this review also uses the meta-analysis approach, with quantified effect estimates and study characteristics coded for some 93 studies. Although the number of studies is larger than in Krol's review, Crain and Mahard intentionally included studies with weaker design characteristics in order to test the impact of design flaws on desegregation effects. Their overall effect size mean is .065 standard deviations, which is both negligible and non-significant.

Crain and Mahard do find differential major effects for grade level, with an average effect size nearing .3 for students desegregated at the kindergarten or 1st grade level, but dropping off markedly to near 0 in the 2nd and higher grades. On the basis of this finding, they argue that desegregation can have a significant effect on Black achievement, providing it starts in or before the 1st grade; it will have little or no effect on students starting desegregation in later grades. It is not clear from the study whether this effect occurs only at these early grade levels, or whether it is cumulative. In any event, there are some further methodological problems with this conclusion. It appears, for example, that none of the studies which have tested kindergarten and 1st graders have been adjusted for possible selection bias, which continues to be a major problem in this field. We will take this issue up once again in our concluding section, after reviewing the NIE studies.

NIE STUDY PROCEDURES

It is clear from the foregoing review that there is still disagreement among the experts about the effect of desegregation on Black achievement. The purpose of the NIE panel is to establish methodological guidelines for selection of studies, to review the studies so selected, and to decide what these studies say about the effect of desegregation on Black achievement. I will comment briefly on these guidelines, leaving their major exposition in the capable hands of Dr. Wortman.

Study Selection Guidelines

The major reason for variations in conclusion of major reviewers is that they are looking at different sets of studies, which vary greatly as to their adequacy for making a causal inference. By establishing "minimum" standards for selecting studies, the NIE panel does not mean that the resulting set is "pure." Indeed, there may be no such studies in existence. The very nature of the process being studied prevents the ideal experiment, where one can eliminate all confounding factors but the factor being tested. It is believed, however, that studies selected according to these guidelines have the best chance for arriving at a decision about whether desegregation itself--and not other factors--was responsible for changes, if any, in Black achievement.

For example, the guidelines exclude cross-sectional studies, because they do not allow determination of whether desegregated students have actually gained on the achievement test in question compared to segregated students, or whether differences simply reflect prior differences between segregated and desegregated students that persist over time. Likewise, longitudinal (over-time) studies without a control group of some kind are also excluded since some academic growth can be expected of nearly all students during their school career, regardless of desegregation experiences. A segregated control group is necessary if one wishes to conclude that desegregated Black students have gained or lost in comparison to Black students who remained in segregated schools.

Thus, in addition to the usual requirements of quantifiability, relevance, and so forth, all selected studies fulfill a basic quasi-experimental design, with pre- and post-tests as well as a segregated control group (where segregation is defined as 50 percent or more Black). We do not imply, however, that there are no further methodological problems. Only one of the studies selected is a randomized experiment and therefore the control group is not generally equivalent to the treatment group prior to the start of desegregation. Wortman's preliminary analysis shows that the correlation of pre-test and post-test effect sizes is .74. This condition raises a serious threat to causal inference, because--just as in a cross-sectional study--any observed differences between desegregated and segregated students after desegregation could simply reflect pre-existing differences between the treatment and control groups.

Fortunately, the selection criteria also require pre-test means to ensure that adjustments can be made to remove the pre-treatment effects. As we shall see, adjusting the control groups for initial differences has a significant impact on one's conclusions from these 19 studies.

I disagree somewhat with two of the guideline provisions. First, the adjustment method to be described in the next section is not infallible and is itself based on a number of assumptions. While it probably works well for modest pre-test differences, there is no guarantee that it corrects properly for gross differences between treatment and control groups, say those approaching or exceeding one standard deviation. Since researchers are reluctant to compare the growth patterns of white and Black students precisely because their differences approach this magnitude, I question whether it makes sense to compare two groups of Black students who exhibit similar differences.

Second, the guidelines do not require equivalent pre- and post-tests, but only that the content is similar and that the same test is used for both treatment and control groups. For example, SRA reading might be used as the pre-test and Iowa reading as the post-test. Although one can convert each test score to a standardized score, using that test's standard deviation, this converted mean still reflects test content, thereby preventing us from establishing that the treated group actually changed on the criterion in question. Moreover, if this issue is combined with substantial pre-test differences, it is quite possible that spurious effects can arise (e.g., high-achieving Black students can show greater relative gain from the CTBS at time 1 to the Stanford at time 2 than low-achieving Black students, and more than high-achievers would show from CTBS at time 1 to CTBS at time 2).

Fortunately, only one study (Rentsch, 1967) embodies both features and, accordingly, I have excluded it from the review in the next section. I have also excluded the Thompson and Smidchens (1979) study on two grounds: its segregated control group averages only 42 percent Black, which means it is not segregated by the 50 percent criterion, and no pre- or post-standard deviations are available for the purpose of computing a standardized effect estimate. A sensitivity analysis is shown in the discussion section to test the impact of these exclusions on my results.

Analysis Procedures

The fact that pre-test differences have a high positive correlation with post-test differences in the studies being reviewed makes it imperative to adjust post-test scores for pre-test differences. If this is not done, then desegregation effect estimates will be biased by pre-existing differences between segregated and desegregated students.

In general, I have followed the procedures outlined by Wortman (1982), with several refinements which are described here. Ideally, what one would like to have is a population standard deviation for each grade and test, so that truly standardized means could be calculated independent of sample variations. Unfortunately, this information is not readily available, and it is not available at all if one wishes to use estimates for Black populations alone. Therefore, sample estimates of standard deviations must be used for calculating adjusted effect estimates.

My procedure differs from Wortman's only in the fact I pooled standard deviations wherever possible to improve the reliability of the standard deviation estimate. If the data shows an apparent fan-spread effect, indicated by higher post-test standard deviations than pre-test standard deviations, then standardized effects were computed separately for time 1 and time 2 means using pooled standard deviations for each time. If no fan spread was apparent, then all standard deviations were pooled for the estimate.

Moreover, I made estimates even where some or all sample standard deviations were missing. If only pre- or post-test standard deviations were available, then they were pooled for the population estimate. In a couple of instances I used standard deviation estimates from other studies in our NIE set, providing they were based on the same test. The advantage of this approach is that a greater number of adjusted effect est'mates are available than in Wortman's approach. This analysis feature is fairly critical, since many otherwise excellent studies in our set have all of the design requirements and the pre- and post-test means, but lack only standard deviation estimates (sometimes from only one time period). It seems improper to exclude such studies from effect size means when other standard deviation information can be sued to provide reasonable approximations.

Other less important analytic issues will be raised in the study-by-study discussion, to which we now turn.

REVIEW OF THE STUDIES

A summary of desegregation effects on Black achievement from each of 17 studies reviewed is tabulated in Table 1. More detailed information, including pre-test means, gain scores, and pooled standard deviations are shown in an appendix table, along with Wortman's effect estimates (which are very close to mine in most instances where he computes them). Table 1 also shows the results of significance testing carried out by each study's author, denoted by an asterisk next to the effect estimate if it exceeds the .05 level.

Anderson

The first study in the group, a voluntary transfer plan in Nashville, shows the largest effect sizes of the studies reviewed, for both math and reading. It is not only statistically significant (by the author's test), but educationally large as well, with reading gains nearing 1 standard deviation. Note that the study has converted test scores into T-scores relative to each grade level, so that decreases in the means are not inconsistent with increases in raw score means. Also, given this type of standardization, fan spread cannot be detected and so all sample standard deviations were pooled for the estimate. Since the two groups were equal on pre-test means, fan spread should not be a problem in any event.

Beker

This study evaluates a voluntary transfer plan in the North. Our analysis differs somewhat from Wortman (other than using pooled standard

TABLE 1

SUMMARY OF THE EFFECTS OF DESEGREGATION ON BLACK ACHIEVEMENT

Study Author	Grade Levels Tested Pre - Post[a]	Desegregation Effect Size Reading	 Math
Anderson	2S - 4S	+.89*	+.54*
Baker	2F - 2S	+.34	-.28
	3F - 3S	+.17	-.04
Bowman	3F - 5S[c]	+.03*	-.05
	3F - 5S[c]	-.55	-.37
Carrigan	KS - 1S	-.55	--
	1S - 2S	+.13	--
	2S - 3S	-.19	--
	3S - 4S	+.21	--
	4S - 5S	+.10	--
	5S - 6S	-.11	--
Clark	6F - 6S	-.01	-.12
Evans	4F - 4S	-.03	-.12
	5F - 5S	+.06*	+.26*
Iwanicki	2S - 3S	.00	--
	4S - 5S	.00	--
	6S - 7S	.00	--
Klein	10F - 10S	.00	-.08
Laird & Weeks	1S - 4F	+.54*	.00
	3F - 5F	+.24*	-.18
	4F - 6F	+.19	.00
Savage	9 - 11	+.15	-.08
Sheehan	4F - 5S	-.16*	-.21*
Slone	4S - 5S	+.27	+.47*
Smith	6S - 9S	-.06	+.13
Syracuse	4F - 4S	+.75*	--
	3F - 4S	.00	--
Van Every	4F - 6S	-.46	+.51
Walberg	3,4F - 3,4S	-.02	--
	5,6F - 5,6S	-.21	--
	7,9F - 7,9S	+.08	--
	10,12F-10,12S	-.25	--
Zdep	2F - 2S	+.53	-.17

* Significant at .05 level or better by author's test

a S denotes spring, F denotes fall

b In standard deviation units

c First entry uses regular segregated control group; second entry uses segregated control group with an enriched program.

deviations). Wortman used a control group of Black students who were accepted for the voluntary transfer plan but who ultimately turned it down. There was another potential control group of students who were accepted, but could not be accommodated in the transfer program due to lack of space. Since this group did not differ to any significant degree from the "refuser" group, I pooled the two groups to improve N's and standard deviation reliabilities. Compared to Wortman, this procedure yielded higher effects for reading but lower effects for math. The author did not compute a formal test so far as I can discern, but his discussion implied significant positive effects for 3rd grade reading, significant negative effects for 2nd grade math, and no other significant effects.

Bowman

The Bowman study is the only one I have included which uses different pre-and post-tests (N.Y. State and Iowa, respectively). One reason I included it was the fact that the pre-test showed only modest differences between the desegregated and the control groups (about ½ standard deviation), and also because it has a second and novel control group: Black students remaining in a segregated school and classroom but with an enriched educational program. Interestingly, while there are no large effects of desegregation compared to the regular controls (although the author reports a significant t-test for reading), there is a very large effect (non-significant according to the author) showing that segregated enriched students gained more than desegregated students. (In the Appendix all means are divided by their respective standard deviations, and therefore appear in standardized form.) Sensitivity analysis shown later evaluates the effect of including or excluding the segregated-enriched control group.

Carrigan

The Carrigan study evaluates a mandatory "one-way" busing program, arising from the closure of a predominately Black school. One might object to the control group here, because it was just at 50 percent Black. Nonetheless, it was in an area undergoing transition and does just barely meet the definition being used here.

Pre-test means are not shown in the Appendix, since Carrigan did not tabulate them for subjects in the study for both the pre- and post-test (there were some dropouts and missing data). Given the small N's such inconsistencies might bias the standard deviation estimates, so I simply pooled all standard deviations for a single estimate, which can then be divided into the gain score for the effect size. Wortman apparently used the existing pre- and post-standard deviations (with inconsistent N's), thereby accounting for the variations with my estimates. However, the estimates averaged across all grades are very close.

Clark

Clark evaluated a voluntary transfer program in North Carolina. This is the first study in the NIE set where all design criteria are met except pre- and post-standard deviations. Presumably because of missing standard deviations, Wortman analysed the SCAT verbal test; although even here only a single standard deviation is available. I have chosen the STEP reading test, although the results are similar to those for the SCAT. For a pooled standard deviation I have used the estimate from Savage (see below) whose standard deviation averaged 14 at the 9th grade level. According to STEP norm tables, the 6th grade standard deviation should be about 1 point lower than the 9th, but I have used 14 from Savage as a conservative estimate. Given the small change, a standard deviation in the 13 to 15 range will not alter the effect estimate. I also used 14.0 as the standard deviation for the SCAT quantitative test, although this is probably conservatively high (thereby producing a smaller negative effect). Fan spread should not be a problem here, since pre-test means are virtually identical for the two groups.

Evans

This study evaluates a comprehensive, two-way mandatory program in Ft. Worth, one of only two such programs in the NIE set. Again, all design requirements were met except for pre- and post-standard deviations, so we used those from Sheehan, who assessed Black outcomes at the same grades in the sister city of Dallas (using the same test). I interpolated for an estimate of 4th grade Spring and 5th grade Fall. It should be noted that all standard deviation values here are lower than those shown for national norms.

Iwanicki and Gable

This study is the only one of several evaluating Project Concern, a voluntary program in New Haven, Connecticut that qualified under the panel's guidelines. Unfortunately, this study focuses on the second year of desegregation, so this factor should be taken into account when interpreting the results. Considering the similarity of the pre-treatment means at each grade level, however, (which reflect the end of the first year of desegregation), and the fact that the control group was drawn randomly from a group meeting Project Concern's requirements, including agreeing to participate when an opening occurs, it appears there were no first-year effects either.

The study does not include standard deviations, but assuming that Black students gain anywhere from ½ to 1 standard deviation in one year (more in earlier years), which is the pattern in our data, then the standard deviations are probably in the 10 - 15 range. This assumption is consistent with white student means reported by Iwanicki which are anywhere from 11 to 18 points higher than the Black means. In any event, since the similarity of pre-test means diminishes the concern for fan spread, and since the gains are identical for grades 2 and 4, the effect size for those grades will be 0 regardless of the standard deviation estimate. For grade 6 we used a conservative effect estimate of 0, even though the effect would be negative if we had a specific standard deviation estimate.

Klein

This study of voluntary transfers in the South is one of only two studies in our set at the high school level. Two control groups were available, one randomly selected from all-Black high schools and one matched on I.Q.. The latter group was selected, due to clear selection effects when transferees were compared to the randomly selected controls. We still have a pre-test difference of 7 points, but it would be 11 points if the random group was used. Only a single standard deviation is available from an analysis of variance table, so the possibility of fan spread cannot be taken into account. However, since the control group has a lower pre-test mean and since each group gained the same amount, any fan spread effect should change our 0 effect into a negative effect, thereby making 0 a conservative estimate.

Laird and Weeks

This Philadelphia study evaluates a voluntary program brought on by overcrowding in a Black school. Students were bused to one of two white schools, Day and McCloskey. The Black students bused to Day were highly biased compared to control students, with both IQ and pre-test means averaging at or near 1 standard deviation above the controls in grades 4 and 5 (in fact, their IQ's equalled white means in the receiving schools). Therefore the McCloskey students were selected for analysis. Since post-test standard deviations differed considerably from pre-test standard deviations, time-specific effect estimates were derived.

The effects in this study are quite large and significantly positive for reading at grades 4 and 5, but negligible and non-significant for math at all grade levels. The authors used matched samples for their significance tests.

Rentsch

The results from this two-year evaluation of the volunteer busing program in Rochester (grades 3,4, and 5) are excluded from Table 1 on methodological grounds. First, the pre-test and post-test were different tests, and the author did not make it clear which tests were used and when they were administered. Second, pre-test differences between the desegregated and segregated control groups neared or exceeded 1 standard deviation. Most devastating of all, information received after the panel had selected this study revealed that white students were included in the study, and the selection method used for the bused students makes it highly likely that the desegregated group had two to three times as many white students as the control group. This possibility could explain why the desegregated group had such higher pre-test means.

The average reading effect for the three grades in the Rentsch study is +.50, while the average math effect is -.11. Sensitivity analysis will show the effect of including or excluding this study on my overall conclusions.

Savage

This evaluation of a Richmond, Virginia voluntary evaluation plan is the only study in our set to investigate the high school level. Three of the four standard deviations for reading were about equal and similar to published norms, but a fourth was 2½ times larger (post-test for controls) and reflected a possible computational or typing error. These three standard deviations were pooled for reading; pooling was done separately for pre- and post-standard deviations for math due to fan-spread indications.

Sheehan

This study of the Dallas plan may be especially significant because of its large N (nearly 2,000 students), a time span of two years, and being the only other evaluation of comprehensive two-way mandatory busing in this set. While the negative effect of desegregation is not large here, the size of the N renders it statistically significant--the only such negative effect in the set.

Slone

An example of pairing is illustrated in this New York City evaluation, although it was implemented in only a few schools. The desegregation started in Fall, 1964, but the pre-test was given in Spring, 1965, so this study also represents a test of second year effects. On the other hand, Slone presents reading tests from Spring, Grade 3 (1964) showing that the desegregated and segregated groups started out with the same relative difference in reading achievement (25.5 months vs. 21.5 months) prior to desegregation. These pre-test differences of about ½ standard deviation would make pre- and post-standard deviations desirable, but they are not available. Only a single pooled standard deviation is used for the effect estimate.

Smith

This Tulsa, Oklahoma study is the only one in the NIE set to study school desegregation due to residential patterns; it is also one of the longest-term studies. The desegregated schools have a higher proportion Black than the other studies, averaging about 42 percent.

Syracuse

This study evaluated an "open enrollment" busing program in Syracuse, New York. Matched and unmatched controls were available; only the matched groups were used here. The control group for the 4th grade group was drawn from a different school than attended by the bused students originally. An overall standard deviation estimate was computed from a t-statistic; since the groups were virtually equal at pre-test, no fan spread correction is required.

A third grade group bused for two years to another receiving school is also reported in Table 1, but not analysed by other members of the panel. This group is of considerable interest because it is longer-term and, especially, its control group is drawn from the same school as the bused group. Only gain scores are reported, but the author reports that the matching was successful and that there were no significant differences between bused and matched control students. The standard deviation estimate is borrowed from Beker's 3rd grade, Spring and Smith's 6th grade Spring estimates, but its size is immaterial given the equality of the gain scores.

Thompson and Smidchens

This study was excluded because the "segregated" control group averaged only 42 percent Black. Sensitivity analysis will assess the impact of this exclusion on our final effect estimates.

Van Every

This is a unique study of school desegregation brought on by a new housing project located in a predominantly white school attendance zone; the control group is drawn from a Black segregated school with socio-economic characteristics comparable to the desegregated group. No difference between pre- and post standard deviations was found, so one pooled estimate was used. Although Van Every reports a non-significant post-mean difference, there appears to be a calculation error. Both the reading and math differences appear to be statistically different.

Walberg

This study evaluates the Boston METCO program, a voluntary city-to-suburb busing plan like Project Concern. Grades 3 and 4 are combined, as are 4 and 5, and so on, due to small N's in the control subjects. No differences between pre- and post-standard deviations were observed, so over-all pooled estimates are used at each grade level. Math results are unreported here because of unreadable figures on xeroxed copy.

Zdep

The final study evaluates another voluntary metropolitan plan. The pre- and post-tests are from the same publisher, but the two different forms are not directly comparable and hence the raw score "gains" presented in Table 1 are presented only so the reader can derive post-treatment means. When converted to standardized "scale" scores from published norms, the bused group gained 4 more points on reading and lost 2 on math when compared to the control group (the national standard deviation of the scale scores is 10). Zdep found one of the largest effects on reading in the set, but the small N renders it statistically non-significant.

The Wortman Effects

The Wortman formula always computes effect estimates separately for time 1 and time 2, and uses only the control group standard deviations. One can see from the Appendix that whenever identical groups and tests are being assessed, in most cases my estimate agrees closely with Wortman's. The main discrepancies arise in the Carrigan and Walberg studies, where absence of pre- and post-means on the same group of persons led me to use only the gain scores and a pooled standard deviation. Even for these studies the effect estimate averaged across all grade levels is very similar. The discrepancy in the Beker study arises because I combined two groups of segregated students for the control group: those who "refused" to join the busing program, the group used by Wortman, and those who accepted but could not be accommodated.

The important difference between the Wortman formula and the approach used here is the number of effect estimates obtained. By pooling standard deviations and by estimating standard deviations from other information, effect estimates are obtained for every study. Even though a precise standard deviation is not available, in many cases the treatment-control initial scores and gain scores are so similar that the effect will be near zero no matter what standard deviation is used. These near-zero effects can have a significant impact on overall effect estimate averages.

DISCUSSION

Although the number of studies in the set reviewed here is not large, the advantage of the panel's approach is that most studies exhibit above-average methodology, and most appear to be carefully conducted. Most important, each study meets reasonable standards for possible causal inference: a pre-post design with a control group. What is lost in numbers, then, is gained in design quality, which is essential in arriving at a sound judgment about the impact of desegregation on Black achievement.

The studies also exhibit a variety of desegregation settings and types, although they are weighted more towards voluntary programs than mandatory, a definite limitation for generalization. On the other hand, for this reason this set may provide a good test of the hypothesis, since it is probably the case that voluntary programs offer better opportunities for positive effects more support from the community, self-selection of families most desirous of the experience, and so forth.

The other major restriction on generalization is that the longest-term study here is only three years in duration, thereby complicating inference for desegregation experience spanning the whole school cycle. Given this panel's search, apparently there are no longer-term studies of adequate quality.

Taken as a whole, what do these studies tell us about desegregation and Black achievement? There are several ways to approach an answer to this question.

First, we can consider the significant tests carried out by the author of each study. Of the 47 different grades and tests in these studies that were subjected to statistical analysis, only 11 were found significant at an acceptable level, and two of these were negative effects. We would add three more significant results out of 53 possible if the Rentsch study were to be added to the set. Thus the overwhelming majority of these studies, taken individually, found no significant effects of desegregation on Black achievement.

The meta-analysis technique employed by the panel provides a second and more reliable method that goes beyond this simple counting exercise. We can arrive at an overall assessment of desegregation's impact by averaging the size of effects across all studies and grade levels. I adopted two alternative strategies in computing these overall averages. First, I computed the average of the effect estimates shown in Table 1, which reflects a group of studies that differs somewhat from the total group adopted by the panel. Second, for sensitivity purposes, I averaged effects for the original set of studies as selected by the panel. This second set of averages therefore includes results from the Rentsch study and the Thompson and Smidchens study and excludes the extra grades 1 analysed from the Bowman and Syracuse studies.

The average effect sizes are shown in Table 2. For the set of studies I selected, the average effect is .06 of a standard deviation for reading and .01 for math. Neither of these two average effect sizes are significantly different from 0 by statistical test. When we consider those studies as originally adopted by the panel, the effect for reading rises to .11 and the math effect falls to 0. The reading effect is still not significantly different from 0. The average reading effect size of .11 for the panel's original studies is somewhat smaller than Wortham's average effect, primarily because of his decision not to calculate effect estimates for a number of studies with effects near 0 (due to incomplete standard deviation information).

For the sake of discussion, let us assume that the more liberal effect estimate of .11 for reading held up across a larger number of studies, so that it would be statistically significant. We must still decide whether a reading effect of this size would be educationally significant.

First, we must keep in mind that the unit of measurement here is variation in Black scores, which is known to be smaller than that for Black and white students combined, or for national norm data, perhaps on the order of two-thirds or three-fourths. Therefore, even if one found an effect of .11 in a larger group of studies, the effect in terms of national norms is still less than .10 or less than one month of a school year. Since the achievement differential between Black and white students averages between 1 and 1.5 standard deviations, an average effect of .11 for Black reading achievement means that desegregation alone could close the gap by less than 10 percent.

Second, such an effect might be ẽducationally significant if it was cumulative over time; that is, if a Black child gained .11 or one month of a school year for each year the child was in a desegregated school. Is there any evidence for such a possibility in this group of studies? This possibility can be tested to some extent by dividing up studies according to duration and computing average effects for one-year studies, two-year studies, and three-year studies. I have carried out this analysis for reading scores using the panel's original 35 grade levels.

TABLE 2

THE AVERAGE EFFECT OF DESEGREGATION ON BLACK ACHIEVEMENT

	Average Effect Size[a]	
Study Grouping	Reading	Math
Table 1 Studies	.06	.01
(N)[b]	(33)	(18)
Original Panel Studies	.11	.00
(N)	(35)	(22)

[a] In fractions of standard deviation. One-tenth of the black student standard deviation (.10) is equivalent to about one month of educational growth as measured by most standardized tests.

[b] Number of grade levels for which the average is computed.

If desegregation effects are cumulative, one should see increasing effects sizes as the duration of desegregation increases.

The results for reading are summarized in Table 3. The average effect is +.04 for one-year studies, +.37 for two-year studies, and -.16 for three-year studies. While the two-year studies do have larger effects on the average than one-year studies, the three-year studies show an average negative effect (due largely to the Van Every study). Therefore, there is no evidence from these studies--the best available--that there is any cumulative effect of desegregation. This conclusion must be qualified, of course, by the fact of the relatively small number of cases for any given duration period.

What about the grade at which children are desegregated? When we compute average effects by grade level, the studies here reveal average effects of -.55 for desegregation begun at grade one (one study), .35 for grade 2, and inconsistent effects near zero for other grades. This set of high-quality studies does not support Crain and Mahard's finding of large effects for grade 1 (and kindergarten) but no effects for grade 2 and higher grades.

Finally, it is noted that there are several studies with very sizable reading effects: Anderson, Syracuse, Zdep, one grade from Laird and Weeks, and two grades from Rentsch. Without these six grades (out of 35 in the set), the reading effect would be near 0. Therefore, even the overall average reading effect of .11 is not a consistent effect of desegregation. It would be more accurate to summarize our studies by saying there are six grades with substantial reading effects ranging from .5 to .8 and 29 grades with much smaller reading effects that average out to about 0.

No matter how one summarizes these desegregation effects, the conclusion is inescapable: the very best studies available demonstrate no significant and consistent effects of desegregation on Black achievement. There is virtually no effect whatsoever for math achievement, and for reading achievement the very best that can be said is that only a handful of grade levels from the 19 best available studies show substantial positive effects, while the large majority of grade levels show small and inconsistent effects that average out to about 0.

The fact that only a small fraction of these studies show substantial effects, even though all grade levels were desegregated, suggests strongly that factors other than desegregation are the real causes of the large achievement gains documented in these studies. We have no way to investigate what these factors might be, but one hypothesis is that they are due to unique educational programs available in those few schools. Indeed, given the much larger effects demonstrated in many purely academic interventions (see Walberg's paper in this volume for a discussion of some of these interventions), this hypothesis may be the only reasonable explanation for the considerable variation observed in the panel's selected studies.

TABLE 3

THE EFFECT OF DESEGREGATION ON BLACK READING ACHIEVEMENT, BY YEARS OF SEGREGATION[a]

Length	Average Reading Effect Size	
One year	+.04	(N=23)
Two years[b]	+.37	(N=9)
Three years[c]	-.16	(N=3)

[a] Using only the original panel studies, including Rentsch and Thompson & Smidchens.

[b] Anderson, Laird & Weeks, Rentsch, Savage and Sheehan.

[c] Bowman, Smith and Van Every.

IMPLICATIONS FOR POLICY

Although the findings of each paper in this volume differ to some extent, the range of difference is small in comparison with previous debates on this issue. With the exception of Crain, all panelists find no effects for math achievement, and find that reading effects are positive but quite small and not educationally significant in all but a few studies. Perhaps a majority of the panel also agrees that the average reading effects are considerably smaller than what might be expected from special educational interventions.

What, then, should the policy directions be from this consensus of experts? It seems to me there are four audiences whose future actions might be influenced by these results.

The community of educational researchers might justifiably decide that enough research has been done on the issue of desegregation and achievement, and that their energies and resources should be devoted to more fertile pastures. There will be some, of course, who will find sufficient flaws in all 19 of these "best" studies to recommend one more large-scale, well-funded study to provide a definitive answer. I would not quarrel with such a study, but at this point the probability of a negative or indeterminate answer (given current knowledge) is high, thereby making its cost hard to justify.

For educational policy makers, I think these results offer an excellent opportunity to reconsider priorities for programs designed to enhance minority student achievement. Desegregation is simply not a cost-effective technique to accomplish this goal. However desirable racial balance may be for other purposes. it is not going to reduce the achievement differential between white and Black students. It is time to solve educational problems with educational solutions, and many promising directions are documented in the Walberg paper.

The courts and civil rights activists should also take note of these findings. The studies reviewed here tell us nothing about whether segregation caused the Black-white achievement gap, but they do tell us that desegregation by itself will not close it to any important degree. There is controversy about the role played by achievement issues in the original Brown decision, but there is no question that many lower courts have been influenced by achievement results when fashioning desegregation remedies. One hopes that the results here will relieve judges of the misconception that they are benefiting the academic progress of minority students by ordering desegregation plans.

Finally, these findings may offer relief to many Black parents who have willingly endured the hardships of cross-town school transfers because of the mistaken belief that their children will benefit academically. Many will continue to endorse such transfer for other reasons, but many others may well be happy to discover that their child can get just as good an education in a neighborhood school close to home.

This does not mean we should abandon desegregation: it remains a goal all panel members share. I think it does raise serious questions about compulsory desegregation methods such as mandatory busing. There is little justification for forcing parents and children into expensive, time-consuming cross-town bus rides when there is no educational advantage. For those of us who want to pursue the goal of integrated educatiou, we should support comprehensive voluntary transfer programs, on a metropolitan basis where necessary. It should be made clear to all participants, however, that simply changing to schools that are more racially balanced than one's neighborhood school is no guarantee of a superior education. Indeed, they may be giving up possible advantages of special programs in their own school--programs designed specifically to enhance education and proven to work.

THE EFFECT OF DESEGREGATION ON BLACK READING AND MATH ACHIEVEMENT

Study and	Test and	Desegregated		Segregated		$Gain_D$-	Pooled sd		Wortman	Author
Grade/Year	(N_D/N_S)	Pre $\bar{X}$	$Gain_D$	Pre $\bar{X}$	$Gain_S$	$Gain_S$	(T_1/T_2)	Effect	Effect	Test
Anderson	Metro (T-scores)									
2/60 - 4S/63	(34/34)	44.3	2.3	46.4	-4.8	+7.1	8.0	+.89	+.95	+
MATH:	(34/34)	44.6	3.6	43.8	-1.3	+4.9	9.0	+.54	+.53	+
Beker	Stanford (GE months for paragraph meaning)									
2F/64 - 2S/65	(25/32)	15.9	6.7	16.3	5.2	+1.5	2.3/6.7	+.34	+.23.	?
3F/64 - 3S/65	(11/28)	24.2	8.5	20.0	5.5	+3.0	6.6/8.9	+.17	-.04	?
MATH:	(25/32)	15.6	4.7	16.7	7.1	-2.4	4.3/6.6	-.28	-.02	?
(Concepts)	(11/28)	20.6	7.6	20.3	7.9	-0.3	6.9/9.3	-.04	+.59	?
BOwman	Iowa (Pre-test is NY State: scores here are standardized by test sd's)									
3F/67 - 5S/70	(12/36)	2.80	-.06	2.33	-.09	+.03	4.7/12.0	+.03	+.02	+
	(" /21)	(Sed. Enriched)		2.24	+.61	-.55		-.55	--	0
MATH:	(12/38)	2.16	+.14	2.05	+.19	-.05	2.7/7.0	-.05	-.06	0
	(" /21)			1.95	+.51	-.37		-.37	--	0
Carrigan	California (Age-equivalent)									
KS/65 - 1S/66	(17/23)		7.1		10.0	-2.9	5.3	-.55	-.41	0
1S/65 - 2S/66	(16/21)		7.5		6.7	+0.8	6.3	+.13	-.02	0
2S/65 - 3S/66	(25/23)		6.6		8.3	-1.7	9.0	-.19	+.30	0
3S/65 - 4S/66	(11/23)		11.6		9.2	+2.4	11.2	+.21	-.13	0
4S/65 - 5S/66	(13/24)		9.1		7.3	+1.8	17.4	+.10	+.33	0
5S/65 - 6S/66	(13/21)		3.3		5.1	-1.8	16.9	-.11	-.31	0
Clark	STEP (Coverted scores)									
6F/69 - 6S/70	(108/88)	250	4.9	248	5.1	-0.2	14.0 *	-.01	--	0
MATH: (SCAT)	(108/88)	254	5.5	254	7.2	-1.7	14.0 *	-.12	--	0
Evans	Iowa (GE months)									
4F/71 - 4S/72	(393/180)	32.0	3.0	29.0	3.0	0.0	10.0/11.6**	-.03	--	0
5F/71 - 5S/72	(381/181)	39.0	2.0	37.0	1.0	+1.0	11.6/13.2**	+.06	--	+
MATH:	(392/179)	33.0	4.0	32.0	5.0	-1.0	8.3/9.8**	-.12	--	0
	(386/181)	40.0	5.0	39.0	2.0	+3.0	9.8/11.3**	+.26	--	+

Study and Grade/Year	Test and (N_D/N_S)	Desegregated Pre $\overline{X}$	 $Gain_D$	Segregated Pre $\overline{X}$	 $Gain_S$	$Gain_D$- $Gain_S$	Pooled sd (T_1/T_2)	Effect	Wortman Effect	Author Test
Iwanicki	Woodcock									
2S/76 - 3S/77	(64/50)	102	13	100	13	0.0	?	.00	--	0
4S/76 - 5S/77	(66/48)	125	5	124	5	0.0	?	.00	--	0
6S/76 - 7S/77	(70/65)	136	2	134	5	-3.0	7	.00	--	0
Klein	Cooperative									
10F/65-10S/66	(38/38)	104	13	97	13	0.0	31.6	.00	--	0
MATH (Z-scores)	(38/38)	.23	.03	-.16	.11	-0.08	1.0	-.08	--	0
Laird & Weeks	Philadelphia Achievement									
1S/63 - 4F/65	(20/140)	3.7	5.1	4.2	4.0	+1.1	1.7/2.3	+.54	--	+
3F/63 - 5F/65	(13/140)	7.2	4.2	6.7	2.2	+2.0	1.4/2.5	+.24	--	+
4F/63 - 6F/65	(10/147)	8.4	4.1	9.1	3.7	+0.4	2.2/2.6	+.19	--	0
MATH:	(19/138)	4.9	2.3	5.6	2.9	-0.6	1.6/3.0	.00	--	0
	(16/139)	6.6	2.6	6.8	3.4	-0.8	2.0/2.9	-.18	--	0
	(14/167)	7.7	4.3	8.5	4.3	0.0	2.9/2.8	.00	--	0
Savage	STEP (Converted scores)									
9/68 - 11/70	(42/42)	269	10.6	271	8.5	+2.1	14.2	+.15	+.14	0
MATH:	(42/42)	256	3.6	253	3.8	-0.2	11.5/16.0	-.08	-.05	0
Sheehan	Iowa (GE months)									
4F/76 - 5S/78	(810/1115)	27.6	9.2	29.0	11.8	-2.6	10.0/13.2	-.16	-.16	-
MATH:	(810/1115)	28.3	8.2	29.2	10.3	-2.1	8.3/11.3	-.21	-.16	-
Slone	Metro (GE months)									
4S/65 - 5S/66	(86)	40.2	11.0	34.9	8.8	+2.2	8.1	+.27	--	0
MATH:	(98)	38.1	5.1	36.7	2.1	+3.0	6.4	+.47	--	+
Syracuse	Stanford (GE months)									
4F/65 - 4S/66	(24/24)	34.5	9.2	34.3	4.0	+5.2	7.2	+.75	--	+
3F/64 - 4S/66	(12/12)		11.4		11.4	0.0	8 to 9	.00	--	0

Study and Grade/Year	Test and (N_D/N_S)	Desegregated Pre $\overline{X}$	 $Gain_D$	Segregated Pre $\overline{X}$	 $Gain_S$	$Gain_D$- $Gain_S$	Pooled sd (T_1/T_2)	 Effect	Wortman Effect	Author Test
Smith	Stanford (Raw score for paragraph meaning)									
6S/65 - 9S/68	(124/150)	16.8	18.5	18.1	19.7	-1.2	8.8/12.0	-.06	-.05	0
(Comput. MATH: raw)	(124/150;	10.5	12.3	9.3	10.5	+1.8	4.1/7.2	+.13	+.10	0
Van Every	SRA (GE months)									
4F/66 - 6S/69	(20/21)	31.6	11.5	29.4	16.2	-4.7	10.3	-.46	-.44	0
MATH·	(20/21)	29.6	19.0	30.8	15.2	+3.8	7.4	+.51	+.53	0
Walberg	Metro (Raw)									
34F/68-34/69	(90/17)		1.8		2.0	-0.2	7.9	-.02	+.11	0
56F/68-56/69	(61/29)		3.6		5.0	-1.4	6.8	-.21	-.24	0
79F/62-79/69	(124/25)		2.1		1.5	+0.6	7.8	+.08	+.21	0
HSF/68-HS/69	(72/14)		1.7		3.2	-1.5	6.0	-.25	-.01	0
MATH:										
Zdep	Coop. Primary (Raw scores--pre is 12A, post is 23A)									
2F/68 - 6S/69	(12/15)	14.5	8.4	16.0	4.5	+3.9	6.9/7.8	+.53	+.65	0
MATH:	(12/15)	26.3	-1.9	26.3	-1.0	-0.9	6.8/5.4	-.17	-.15	0

*Estimated from Savage **Estimated from Sheehan

BIBLIOGRAPHY

Anderson, Louis V. The effect of desegregation on the achievement and personality pattern of Negro Children. Ph.D. dissertation, George Peabody College for Teachers (University Microfilms No. 66-11237).

Armor, David J. "The Evidence on Busing." Public Interest, 28, 90-126, 1972.

Armor, David J. "The Double Double Standard: A Reply." Public Interest, 30, Winter, 1973.

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 (Microfilm No. 73-19176).

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.

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

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

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.

Evans, Charles L. Integration evaluation: Desegregation study 11 -- academic effects on bused black and receiving white students, 1972-73. Forth Worth, Texas: Forth Worth Independent School District, 1973 (ERIC No. ED 094 087).

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

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 S. 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.

Krol, R.A. A meta-analysis of comparative research on the effects of desegregation on academic achievement. Unpublished doctoral dissertation, Western Michigan University, 1978. (University microfilms No. 79-07962), 1979.

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.

Pettigrew, T.F. "Busing: A review of the Evidence." Public Interest, 30, Winter 1973.

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

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 R. 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.

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

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., & 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).

Walberg, Herbert J. An evaluation of an urban-suburban school busing program: Student achievement and perception of class learning environments. Paper presented at the Annual Meeting of the American Educational Research Association, New York, 1971.

Walberg, Herbert J. Desegregation and Educational Productivity. National Institute of Education, 1983.

Weinberg, M. Desegregation Research: An Appraisal. Bloomington, Ind., Phi Delta Kappa, 1970.

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

Wortman, Paul M. School Desegregation and Black Achievement: An Integrative Review. University of Michigan, 1983.

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).

Is Nineteen Really Better Than Ninety-Three?

Robert L. Crain

The Rand Corporation and
The Center for Social Organization of Schools
Johns Hopkins University

In this volume, a group of scholars have come together to assess the state of our knowledge about the effects of school desegregation on black achievement test scores. The scholars were selected to represent a range of personal ideologies. Thus this project should provide a near-perfect opportunity to array a group of social scientists along a continuum from left to right and demonstrate that the scientific conclusions they draw are consonant with their personal politics. Doing so would present strong evidence that our worst fear is true--that social science is not really science, and government, in employing social sc'nce, has merely been financing propaganda. Perhaps one can draw this conclusion from the panel's work, but I don't think so.

First, it is not so easy to attach political positions to working social scientists. It makes good sense to classify me as a "liberal;" I have testified in a number of court cases, and while this has sometimes been as a court-appointed expert or on behalf of a school board resisting desegregation, it has usually been as an expert called by the plaintiffs in a suit trying to bring about desegregation. Other members of this panel have testified for school boards resisting desegregation or have been called to present the anti-busing position in congressional hearings. But in at least two cases putting labels on members of the panel is not so easy to do. Paul Wortman was selected as a liberal mainly because he had completed a literature review showing positive effects of desegregation on black achievement; and Walter Stephan was selected as a "neutral" because he is the author of an earlier review concluding that there were few positive effects of desegregation. But every scientist whose data support a black position is not necessarily a liberal, just as every scientist who agreed with Copernicus was not anti-Christian.

It is also not so easy to show a correlation between personal ideology and scientific position. It is true that I, the obvious liberal on the panel, am the co-author of a literature review (Crain and Mahard, 1982) arguing that desegregation seems to raise Black achievement by .3 standard deviations, a larger estimate than any other member of the panel has made; and the panel's most obvious conservative, David Armor, has produced the smallest estimated achievement effect of any member of the panel. But if political position were dominant here, its effect would have to appear in the way the panel selected the 19 studies it considered best. Paul Wortman read the studies gathered by Mahard and me (1982) and by Krol (1978) and recommended to the panel a group of 31 studies as being of superior quality; the 18 that the panel chose to accept from that offering are in fact only slightly less positive in their assessment of desegregation than the ones they declined to use.

There is little evidence of bias in their choice. It is true that when the panel veered from its normal course of using only the data provided by Wortman, it did so to add one study which had found a negative effect of desegregation and to add additional data strengthening a second study in the group of 18 which had found a negative effect. But this is not very strong evidence for an ideological interpretation of the actions of the authors. Finally, one might simply note that when the liberals, Crain and Mahard, reviewed the literature on desegregation, they gathered together 93 studies whose mean effect of desegregation on black achievement was +.08 standard deviations, pooling reading and math effects together; the conservative David Armor reviewed 19 studies and found an effect on reading scores of +.11 and on math scored of .00--an average of .055. It is hard to believe that approximately 180° of political ideology are accurately translated into the selection of two samples whose mean treatment effects differ by only .025 standard deviations.

Ideology does appear in some of the essays in this volume, including this one; but it tends to show up mostly in the conclusions and interpretations--in the words rather than the numbers. One reason it does not show in the numbers is that it is very difficult for contemporary social scientist to disagree about methodology. The technique used here for assessing effect size was proposed by Wortman as neither a liberal nor a conservative solution; it was accepted by all the members of the panel regardless of personal ideology.

But this is not to say that there are no differences worth noting among the panelists, or that these differences have not consequences. There is an important division among the members of the panel, but on a methodological, not ideological, issue--the question of whether one, in reviewing literature, should select only the better studies and concentrate on them, or review all the studies one can find. There is in this panel a rather neat correlation between the number of studies one chooses to look at and the size of the effect of desegregation one finds. Crain and Mahard, using 93 studies, conclude that desegregation raises black achievement something on the order of 1/4 to 1/3 of a standard deviation. Wortman, reviewing 31 studies, concludes that the gain is perhaps 1/5 of a standard deviation. The others, using 19 or fewer studies, conclude that desegregation raises black achievement by perhaps 1/8 of a standard deviation or perhaps less. I would like to argue that in this particular case, it is not an accident that the number of studies reviewed is related to the conclusions drawn.

The question of whether one should selectively review literature or review all of it has been a subject of considerable debate among scientists using what is now called meta-analysis--the computer-assisted review of studies of a particular question. At first thought, the argument that one should choose the best studies and leave the chaff aside seems unquestionably the right answer. Certainly the counterargument that one should include all the studies because error is a random variable--that with a large enough sample of studies errors will cancel themselves out and reveal the truth--seems quite inadequate.

Selection of the good studies seems like the obvious answer only as long as we sleepily think that our task is only to find the competent evaluations of a particular program and compute an overall average program effectiveness score. Most of the meta-analyses done to date and most of the literature reviews discussed by Herbert Walberg in this volume are in fact of this type, but there is no reason they must be this simple. First, one often wants to know more about a new intervention than simply whether it works; we often need to know how and why as well. And even if we only want to know whether there is an overall treatment effect, there are better ways than throwing away most of the research. Suppose there are 100 studies of an innovation. Rather than choosing the ten supposedly best studies and computing an average effect size, one might include all 100 studies in the review, choosing by empirical statistical analysis the 10 best. Alternately, one might evaluate all 100 studies and assign different weights, such as is done in survey research, to those studies which are particularly weak or strong; rather than counting each study equally, one might count the particularly weak studies as being only a fraction of the better studies. Alternately, one might do as Mahard and I did and construct an additive model, assuming that any study which had a particular weakness would overpredict or underpredict the treatment effect by a fixed amount "x," and then estimate x through some statistical procedure. All three of these alternatives are ways of emphasizing the best studies after an empirical analysis of all of them. All else equal, of course we would prefer to select the best studies from a group through an empirical analysis rather than from an a priori judgment.

Viewed this way, the only argument in favor of prior selection is that of efficiency. In many cases this can be a convincing argument. With limited resources one cannot afford to spend vast amounts of time wading through dozens of weak studies in order to gain a modest amount of information. Given the short duration of this project, it might have been impossible for the panel to review all 100-odd studies of desegregation and Black achievement. Perhaps selecting a small group was the only workable plan. But this does not mean that it was a good plan.

In this paper we will argue, first, that selection of a small group of preferred studies from a pool using criteria chosen in advance of examining the studies is in principle a mistake. We will then go on to show that in this case, a mistake in principle was also a mistake in practice: the panel, in selecting 19 studies from the pool of 100, led themselves into a serious error.

The Theoretical Problems with Prior Selection

The analogy to weighting in survey research is useful. In surveys, it is often the case that particular classes of respondents are especially valuable for analysis, and these respondents are oversampled. However, the total sample is then no longer representative of the general population. The solution is to assign a weight, a multiplier, to each of the oversampled cases so that if three times as many cases in one particular class are selected, each is treated as only 1/3 of a case in the final analysis. The selection of some studies to include in a meta-analysis while others are rejected is essentially a decision to assign a weight of 1 to some studies and a weight of 0 to all others. The simplest way to justify doing so is to divide the studies into a small number of discrete categories, arguing that every study in certain categories is worth examining while none of the studies in the other categories is. Unfortunately, anyone that has read literature such as the desegregation-achievement material knows how difficult it would be to justify doing this.

If one does not accept the idea that the studies can be neatly divided into two discrete categories, one good and one bad, then a more systematic approach is to rank the studies by quality, putting the best studies at the top of this list and then moving down the list until we find an appropriate cut-off point so we can discard studies below a certain level of quality. There are several problems with this approach. The first is that study quality is a multi-dimensional concept; a study which is good in one respect may not be in another. Even if studies that are good in one respect tend to be better than average in others, how does one choose to rank one study which is very good in category A and only moderately good in category B above or below another study which is very good in B and only above average in A? While I have not attempted a formal proof, I believe that the Arrow paradox (1951) can be used to show that such a ranking is impossible unless one is willing to assign definite numeric values to, for example, the relative merits of increasing the sample size versus using a pretest measure of higher reliability. If it is not possible for one person to rank the studies unequivocally from best to worst, it is certainly impossible for a group of scholars to do so--meaning that one cannot expect the readers of a meta-analysis to agree with the author that the right decision has been made about study selection.

At this point the reader may argue that I am being a bit pedantic; that all science is imperfect, and more importantly is dependent on scarce resources. With only a certain amount of money and time available, one should not spend it rooting through hundreds of useless studies, carefully recording all their faults. If one used the weighting procedure suggested earlier, one would have to read each study, enter its data into the computer, and perhaps compute weights designed, for example, to minimize the variance in the overall estimate by assigning low weights to classes of studies which have relatively large variability in their estimates of treatment effect. Alternately, if one uses the algebraic model that Crain and Mahard used, one must run regression equations trying to estimate the proper amount to add or subtract from the treatment effects generated by studies of a particular kind. All of this takes time and money away from the main objective, which presumably is to find the best studies and see what they say.

It seems to me that the best way to settle this argument is empirically. We have here an example of each kind of research. Can we compare them and conclude whether the selection of a small number of supposedly better studies is a wiser strategy than a brute force analysis of the entire literature?

The Real-World Problems with Prior Selection of Desegregation Studies

The problem with selecting the best studies of desegregation and black achievement is not merely that the multiple criteria which can be used for selection are imperfectly correlated; the criteria are in fact negatively correlated. The data which Mahard and I assembled on the 93 studies demonstrate this. Methodologically superior studies presumably have larger sample sizes, longitudinal research designs, and evaluate situations which more accurately represent the policy being investigated. In this case, more recent desegregation plans are more interesting to study than earlier desegregation plans because they presumably represent contemporary policy more accurately; and the students being studied should be students who have experienced desegregation from kindergarten or first grade, since that is the way desegregation is done in perhaps 95% or more of all desegregation plans in the United States. Table 1 shows the intercorrelations among these four criteria.

Table 1: Correlations among Study Methodological Attributes and Study Outcomes

	Samp. Size	Longit. Design	Late Date Deseg.	Early Grade Deseg.	Effect Size
"Quality"					
Sample Size (Large)	--	-.23*	.33*	-.10	-.04
Longitudinal Design Yes)	-.23*	--	.03	-.05	.13*
"Representativeness"					
Date of Deseg. (Later)	.33*	.03	--	-.19	-.08
Grade Deseg. began (at early grade)	-.10	-.05	-.19*	--	.24*
Outcome: Effect Size (+)	-.04	.13*	-.08	.24*	--

The correlations are, on the whole, negative. Studies which have large sample sizes tend not to be longitudinal. The more recent the desegregation plan being studied, the less likely it is that the study will be of students who were desegregated at kindergarten or first grade. (The latter negative correlation is almost a necessity since a brand new desegregation plan has not had time for its youngest students to reach an age where they can be easily tested.) If one wants to choose the best studies from among this field, there are hard trade-offs to be made.

The last line of Table 1 shows the correlations between the various methodological dimensions and the overall effect size. We know that most studies of desegregation show a positive effect on black achievement, although our readers cannot be expected to agree on whether that effect is large or small. But given that the effect is positive, and given our assumption that longitudinal designs are preferable to others, it makes sense that there should be a significant positive correlation between using a longitudinal design and the magnitude of the treatment effect. Wortman notes this, pointing out that the average treatment effect of the thirty-one studies he selected is considerably higher than the average treatment effect of the pool of 93 which Crain and Mahard used. But by the same criteria, if nearly all desegregation plans in the United States begin desegregation at Kindergarten or first grade, and there is a strong positive correlation between the grade where desegregation is begun and the treatment effect (see the lower right of Table 1), it follows that the grade at which desegregation began is also an important selection criterion. It would be extremely difficult to have anticipated this in advance of seeing this correlation. But the problem is serious. Imagine that a desegregation plan is adopted in some city, and a local researcher decides to evaluate it. The chances are good that he or she will choose to study the plan during its first year or two. The researcher will not want to wait until the plan has been in place for a decade and is no longer of policy interest or newsworthy. The chances are also good the researcher will the evaluation by studying the test performance of students in the middle elementary grades. These are the youngest grades where students can be easily and accurately tested. In a typical design, the students will have attended segregated schools until the end of second grade, be pretested, transfer to desegregated schools, and be posttested a year later. This is a very clean design, resembling a laboratory experiment. But it is not a study of the right problem. The experience of the students being studied--segregation for three years followed by one year of desegregation--is quite atypical, a transitory stage in the school district's desegregation process. Their younger siblings and all future students in this school system will have four years of desegregation at the end of grade three. And according to Table 1, their achievement gains as a result of desegregation will be considerably more positive than that of the students being studied by this (or most) researcher(s). The 93 studies Mahard and I located included 295 samples of students; of these, four-fifths received a mixed schooling, partly segregated and partly desegregated.

This illuminates the main problem with the prior selection approach--that it assumes the methodological criteria which define a good study are known in advance. This is an assumption we normally take for granted. We know what sort of design is superior and what sort inferior and therefore can make an a priori decision about the quality of any particular study. However, it is unlikely that in practice we can ever actually do this. First of all, one usually cannot know until the data has been examined which of several competing methodological criteria are most important. If there are various threats to validity, the importance of any particular threat depends a good bit upon the particular type of research being done. For example: if achievement test scores are the dependent variable, then reliability of pretest and posttest measures is likely to be less of a problem than if the study deals with measurement of psychological attitudes. Second example: studies of student absenteeism. At the same time, a study of juvenile delinquency might choose to include the studies using self-reported delinquency and exclude studies using delinquency reported by official sources on the grounds that official reports of delinquency are notoriously inaccurate. The same criteria are applied in directly opposite ways in two studies depending upon the subject being studied.

In the case of the effects of desegregation on minority achievement we have found a methodological error--studying students whose education was a mixture of segregation and desegregation -- which is so specific to desegregation research that it was not even recognized as an error and source of bias until our review was done. Table 1 suggests that studies of the effects of desegregation on minority achievement, which use as subjects students who have not experienced a complete desegregation treatment beginning in kindergarten or grade 1, will underestimate the effects of desegregation. One might assume that such an error would be quite rare, since virtually every desegregation plan in the United States begins in kindergarten or grade 1 at the latest. However, a large majority of researchers who have studied the effects of desegregation committed this error, of studying students whose desegregation began not in the normal fashion at the beginning of their entry into school but only after they had received some education in segregated schools, and the reason they have done so is obvious: they wanted to publish quickly on this timely topic, and they wanted to study students who were old enough to be reliably tested.

The panel, in selecting the nineteen studies which they considered to be methodologically superior, did not require that the students being studied have a desegregation experience beginning in kindergarten or first grade. They used instead various other criteria, including that the study be longitudinal; and herein lies the problem. Table 2 shows the relationship between design type and grade at which students are desegregated.

Table 2: Use of Longitudinal Design and Inclusion of Sample in Panel Substudy, by Grade of First Desegregation

Grade	Percent of studies with longitudinal design	Percent of studies included in substudy	n
KG	18%	0%	11
1	41%	4%	44
2	53%	14%	36
3	63%	13%	54
4	47%	21%	38
5	42%	10%	40
6	40%	8%	25
7-12	59%	6%	49

Only two studies (18%) of students desegregated at kindergarten are longitudinal. The reason is obvious--it is difficult to pretest students who have not yet learned to read. And neither of these two studies were selected by the panel. The second column shows the percentage of studies at each grade selected by the panel. Mahard and I found a total of twenty studies of desegregated black students with desegregation beginning in kindergarten or first grade and which contained a segregated black control group. The panel used the data from only one of these studies. The remaining nineteen studies were discarded, usually because these very young children did not provide accurate pretests for longitudinal analysis. Eight of the twenty studies we identified used cohort comparison--comparing the scores of kindergarten and first grade students after desegregation to the scores of the students who had been in kindergarten and first grade the preceding year. The panel, making a rather conventional scientific decision, had judged these studies to be of inferior quality and excluded them. While it is true that in principle a cohort comparison is inferior to a longitudinal experimental or quasi-experimental design, this is precisely an example of the situation where there are competing methodological criteria, and the choice cannot be wisely made in advance of looking at the data. In this case a cohort study is superior because it enables us to study students who had begun desegregation in first grade.

Estimating the Effect of Desegregation

The nineteen studies selected by the panel of scientists show an overall effect of desegregation on achievement which is slightly more positive than the Crain-Mahard larger sample. Whereas we find an average desegregation effect in all 93 studies of .08 standard deviations, our estimate for the 18 of our studies selected by the panel is significantly higher, .16. This is likely the result of discarding non-longitudinal studies. If desegregation has a positive effect, then it follows, as Wortman notes, that accurately done desegregation studies will show a positive effect and the panel's exclusion of technically inferior studies should produce a higher estimate of the effect of desegregation than our strategy of including every study regardless of quality. We arrive at this same conclusion in a different way. By coding the different types of research design as a variable for each study, we show that technically better research designs are correlated with more positive effects of desegregation. As Table 3 indicates, studies in which the performance of blacks in desegregated schools are compared to performance of whites, or the performance of the testmaker's norming sample, often conclude that desegregation has failed to improve black achievement. On the other hand, studies which compare desegregated blacks to segregated blacks--either in a "cohort" design (the segregated blacks are the students in the same grade in the years before desegregation), a "cross-sectional" design (with no pretest) or a longitudinal design--are twice as likely to show positive as negative results; and randomized experiments show positive results eight or nine times as often as negative results.

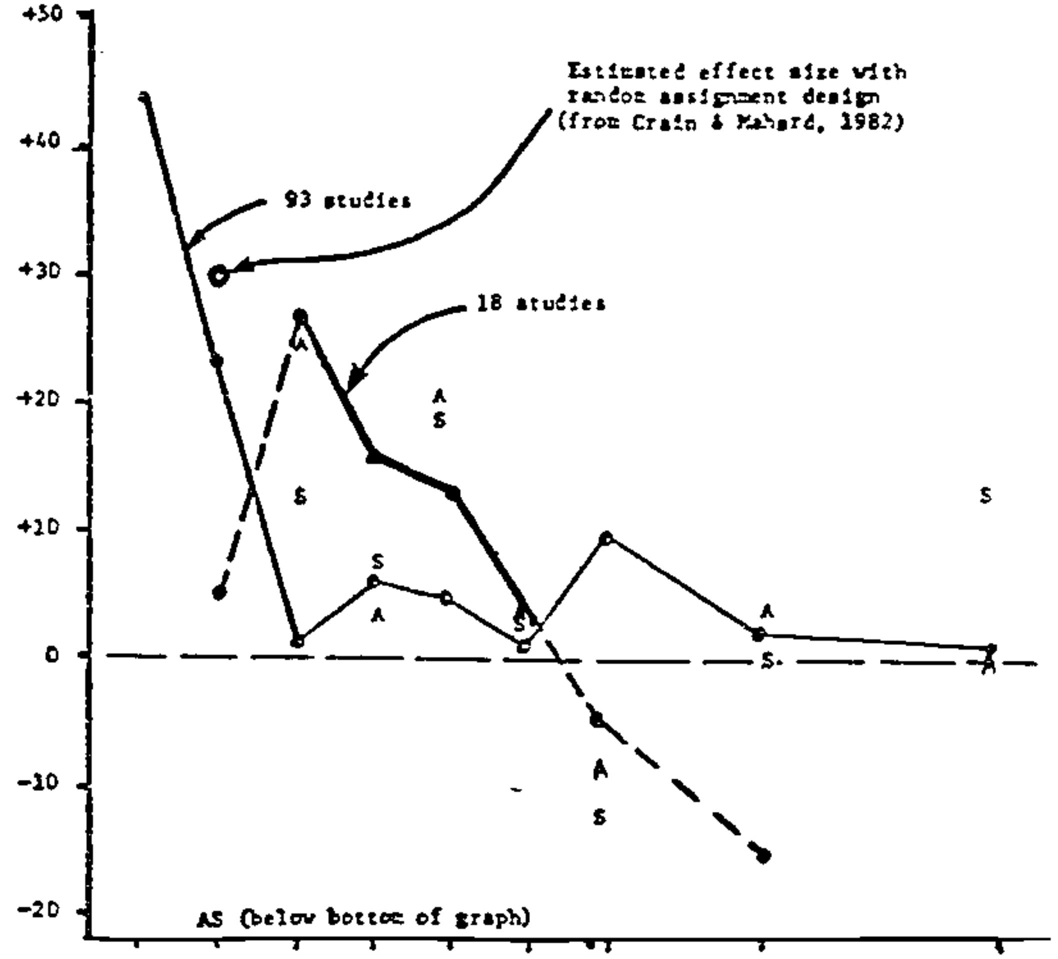
Table 3: Direction and Size of Treatment Effect, by Type of Control Group

		direction of effect				effect size	
	Design	+	0	-	(n)	d	(n)
1.	randomized	86	5	10	(21)	.235	(15)
2.	longitudinal	55	20	25	(141)	.083	(116)
3.	cross-sectional	62	13	26	(39)	.130	(34)
4.	cohort	53	16	31	(64)	.084	(53)
5.	white controls	33	8	58	(12)	.058	(12)
6.	norm controls	34	11	54	(44)	-.030	(39)
	total sample	54	16	30	(321)	.080	(269)

The problem with the research panel's approach is that by excluding supposedly inferior studies by one criterion, they have managed to exclude most of the experiments and all of the studies (except for Carrigan) in which students were desegregated in kindergarten or first grade. Figure 1 shows a plot of the effect sizes estimated by Mahard and Crain for 28 samples of students in the eighteen evaluations selected by the panel. This is shown as a heavy line, which changes to a dashed line where it joins dots based only on one or two samples of students.

The effect sizes for the entire group of 295 samples in the 93 studies we reviewed are shown as a light solid line. In grades 2 through 5 (where the bulk of the samples studied by the panel began desegregation), our estimates of effect size for the panel's studies is considerably higher than our estimate for the larger set of studies. The graph also shows, using the letters A and S, the effect size estimates for each grade computed by Armor and Stephan. In the range from second grade through fifth, their estimates are also generally higher than our estimates for our larger sample. Thus, we again see that the more selective sample shows higher estimates, presumably because it has discarded the very weak designs which are biased toward underestimating the effects of desegregation. At the same time, the other point of this graph is that there are no data points in the panel's nineteen studies for kindergarten and only 1 data point for first grade. (The one first-grade datum is regrettably the rather untrustworthy estimate by Carrigan, which uses a 50% black school for its control group). Also shown on the graph is a circle located above first grade, at approximately +.30 standard deviations, indicating the estimated effect size predicted by our regression equation for a typical study of students desegregated at first grade using a randomized experimental design. If one were willing to assume that Armor's and Stephan's data supported the early grade effect, an extrapolation down to grade one from their date would seem consistent with the estimate. Unfortunately, given the relatively small number of cases and the rather ragged pattern in the data, it is difficult to say whether either Stephan's or Armor's calculations support the hypothesis that there are stronger effects at lower grade levels.

The problem is again made more difficult by the prior selection of studies which has reduced the number of cases so greatly that it is difficult to compute reliable correlation with the data. The best data on the question is the Crain and Mahard analysis. Table 4 presents that data, and shows a quite strong pattern. Of 55 studies of students desegregated in kindergarten or first grade, 45 (82%) show a positive desegregation effect.



	KG	1	2	3	4	5	6	7-9	10-12	TOTAL
n, 18 studies	(0)	(2)	(5)	(7)	(8)	(4)	(1)	(1)	(0)	(28)
n, 93 studies	(10)	(40)	(27)	(39)	(24)	(29)	(20)	(21)	(19)	(229)
n, Armor	(0)	(1)	(5)	(6)	(6)	(7)	(2)	(1)	(2)	
n, Stephan	(0)	(1)	(5)	(6)	(6)	(7)	(2)	(1)	(2)	

Figure 1: Effect Size, Panel and Crain-Mahard samples, by grade desegregation begun

Table 4: Direction and Size of Treatment Effect, By Grade at Initial Desegregation

	Direction of Effect				Effect Size	
grade at desegregation:	+	0	-	(n)	d	(n)
KG	100	0	0	(11)	.439	(10)
1	77	7	16	(44)	.203	(40)
2	56	8	36	(36)	.050	(32)
3	50	26	24	(54)	.080	(46)
4	53	21	26	(38)	.073	(32)
5	44	8	49	(39)	.016	(33)
6	52	8	40	(25)	.090	(21)
7-9	56	16	28	(25)	.011	(22)
10-12	48	22	30	(23)	.005	(17)
total sample	56	14	29	(295)	.079	(253)

Another way to think of the difference between the small-n and large-n meta-analyses is to say that one does the selection at the beginning of the project to narrow the focus upon the most interesting cases while the other does that selection at the end. In the analysis which Mahard and I did, we identified 20 studies as being the best. Since this selection was based upon the empirical findings of the analysis, its main consideration was that the students being studied in each case had to have been desegregated at kindergarten or grade one. Beyond that, we required that there be a control group of segregated black students but our requirements for methodology and the amount of material reported by the authors were more generous than the panel's. Whether our group of 20 is superior to the group of 19 selected by the panel is a matter for the reader to decide, of course.

The 20 "best" studies

Five of the 20 studies use a randomized experimental design:

Stanly Zdep (1971) of TES carried out an evaluation of a city-to-suburban voluntary transfer plan from Newark, NJ to suburb, Verona. Verona apparently agreed to accept 38 students, and the city held a lottery among all applicants. Zdep then used a random selection from the unchosen volunteers as his control group. He limited his analysis to students in first and second grade. The first graders were pretested with the Metropolitan Readiness Test and posttested with the Cooperative Primary Test. On the pretest, the control group tested about .1 standard deviations above the students being transported to the suburbs; on the posttest, bused students were 9.8 answers higher than the control group on a test on which the bused students had a standard deviation of 5.4 and the control group a standard deviation of 3.8. In math, the posttested scores favored the treatment group by 7.6 points (control group standard deviation 6.3) and in a subtest called listening, favored the bused students by 6.0 points (control group standard deviation 5.7). Averaging the three yields an effects size of 1.60. This study was not included in the panel's 19 studies, although Zdep's analysis of second grade students was included. Presumably the first grade data was dropped because different tests were used for the pretest and posttest. Given that the difference on the readiness test between the two groups was small, favored the control group, and most importantly that the students were selected by random assignment, the requirement that the tests be identical seems overly strict. The main problem with the Zdep analysis is that there are only 13 transported students and a control group of 14 in the first grade. (Even with the small sample size there is no problem with significance. The reading test differences yield a t of about 10, for example.)

Bruce Wood (1968) wrote his doctoral dissertation on the Project Concern voluntary city-to-suburb program in Hartford, CT. He analyzed changes in IQ scores. Two-Hundred and sixty-six students in grades kindergarten through five were randomly selected and a control of 303 students was selected, also randomly. At the pretest, the control group scored .6 IQ points higher than the experimental group. In the analysis he divided the group by grade level, combining kindergarten and first grade students. and carried out an analysis of covariance. He does not report the actual raw means, but the obtained f of 4.46 suggests that there must have been a difference of 1/3 standard deviations favoring the experimental group.

Thomas Mahan (1971) was director of the Hartford Project Concern program at the time, and conducted nis own evaluation. He used data during the second year of the project, so that presumably his results are more biased by attrition from the original random treatment and control group than are Wood's. For the second year of the project, Mahan shows an average 9-point increase in IQ for the treatment groups who entered the program in the first grade, compared to control group increases of 3 and 2 points respectively. There are also large differences favoring the treatment group tor students who entered the program in grades 2 and 3 and negative treatment effects for students who entered the program in grades 4 and 5. Mahan also reports the results of achievement testing using the Metropolitan Readiness Test which showed some significant differences for the kindergarten group favoring the bused students, and also some results from the Primary Mental Abilities Test which showed results for both kindergarten and first grade students favoring the experimental group.

Project Concern operated in several cities in Connecticut, and Joseph Samuels wrote a dissertation (1971) evaluating the New Haven program. He compared 37 students who transferred to the suburbs at kindergarten to a control group of 50 students. There are possible biases here, in that Samuel's transferred students were apparently screened after being randomly selected to drop students who "had medical or psychological reasons precluding their involvement..." He does not say how many students were omitted in this way. In addition, the control group was limited to students who remained in the same school for two years, which presumably would bias the control group upward. If there were differences between the two groups, they do not appear on the Monroe Reading Aptitude Test administered to the two groups while in kindergarten; the experimental group tested only .03 standard deviations higher. Two years later, the treatment group tested 5.5 units higher on a reading test with a standard deviation of 12. They also tested 5.6 units above a group of students in a compensatory education program in the city, both differences being significant. The Project Concern students did not test higher than the control group in either word analysis or mathematics--they were about .25 standard deviations lower on both tests.

Meanwhile, the Rochester city schools carried out a similar city-to-suburb program (Rock, et al., 1968). In each of three years, 25 experimental subjects were selected and allowed to transfer to the suburbs while 25 others were held as a control group in the central city. The experimental group scored below the control group on the pretest (the Metropolitan Readiness Test). At the end of the first year, the treatment students did not score higher on the Metropolitan Achievement Test, but did score one-half year ahead of the control group on the SRA battery. The second experimental group also scored below their control on the Readiness Test. but after one year scored about three months ahead of the control group. At the end of one year the third experimental group did not score above control in reading but did score 6 months ahead of the control group in math. In that year, the treatment group was lightly superior to the control group on the pretest, which was the New York State Readiness Test, so this result is questionable.

None of these five experimental studies were selected by the panel. Usually the reason is because the pretest and posttest were not the same. It is nearly impossible to design a study with identical tests covering the kindergarten-first grade range, since the students cannot read at the beginning of that period. Tests are notoriously unreliable for students at this age. In addition, all five of the experimental designs used analysis of covariance models, and relatively little information was provided with which to compute effect sizes. Finally, all five studies have problems with attrition, It is doubtful that the attrition problems are more severe in these studies than they are in the longitudinal studies used by the panel; but these studies are usually more detailed in describing attrition, making it harder to overlook a problem which is in fact present in the majority of longitudinal studies of education. In general, we do not think that these studies should be considered inferior to those chosen by the panel.

There are 8 other studies which use what we call "cohort" comparisons (and which others often call "historical control groups"). These studies compared scores of desegregated students in the particular grade to the scores that blacks made in the same grade before desegregation occurred. This kind of design is the only way to study desegregation in a community where all schools have been desegregated, since no segregated group of black students remains to be used as control. None of these studies have data for a large number of years which would enable one to conduct an interrupted time-series analysis. For example, the Nashville-Davidson County public schools (1979) published mean test scores for black students in each grade for the nine-year period from 1970, when the desegregation plan was adopted, to 1978. The test scores show a considerable gain over the period, ranging from .2 to .4 standard deviations. Of course, the problem is that we cannot attribute this to desegregation; it may be due to other changes in testing or educational practice in the city.

One wonders whether a school district would be anxious to publish the results if it showed negative effects. Perhaps many other school districts have the same sort of data that Nashville has but have not released it to interested researchers because it shows declines in achievement. But one example which works in the opposite direction is from Pasadena, whose school board has been adamently opposed to mandatory desegregation and released a lengthy report by Harold Kurtz (1975) showing the disastrous educational consequences of desegregation there. In 15 tests of students who were desegregated in grades 2 through 12, scores were lower after desegregation 14 times. But there were very large achievement increases for students who were in kindergarten and first grade--averaging .36 standard deviations. Thus while test scores dropped for black students throughout the district during the period of time after desegregation, test scores of the very youngest students went up. This could be a peculiarity of the testing procedure used with the youngest students, of course.

Cohort analysis is necessary when a district is totally desegregated. Total desegregation in the north came first to university communities, the largest of which was Berkeley, which desegregated in 1968. Test scores dropped that spring, about .04 standard dev‿ations in reading for first graders. By 1970, second graders were reading about .16 standard deviations above the second graders of 1968. Thus one report (Dambacher, 1971) shows essentially no change in test scores using the first year of desegregation, while a second paper (Lunemann, 1973) shows a positive desegregation effect. (In this analysis black and "other," presumably Hispanics who did not consider themselves whites, were combined in one year and separated in others. The percentage of "other" students in the district changed radically, however, suggesting that these ethnic classifications were unstable. We have combined "others" with Blacks for all years in order to avoid this problem.)

Another university town which developed a desegregation plan was Evanston. Jayjia Hsia of TES (1971) carried out a lengthy evaluation, and found that in the fall of the third grade, two years after desegregation, students were testing .01 standard deviations below students two years earlier. She found gains in only 3 out of 9 tests in the upper grades over the first two years.

Another school district which reported achievement test scores for the year after desegregation in comparison to the year before was Clark county (Las Vegas) Nevada. Test scores for black students were up .1 years.

In one southern district, George Chenault (1976) found that students who were desegregated in kindergarten scored .3 years higher in the fourth grade compared to students five years earlier.

Finally we have constructed a cohort analysis from the data provided by Patricia Carrigan (1969). The panel treated Carrigan as a Longitudinal study, but the "segregated" control school is 50% black--desegregated by most people's criteria. We ignored the data for the control school and instead compared the performance of the desegregated black students to black students at the sending school prior to desegregation. We found the integrated students scoring .05 standard deviations higher.

All the cohort studies are subject to alternative interpretations--change in curricula, in type of test, in test administration, could all affect test scores. On the other hand, cohort studies have the advantage of ᵇ ving ːelatively large sample sizes. They are also not likely to be a‿fected by complicated statistical procedures which sometimes do more harm than good. Of eight studies of students desegregated at kindergarten or first grade, we found gains in 6, the exception being Hsia's Evanston study and Dambacher's Berkeley study, whose conclusions were reversed the following year by Lunemann.*

The final group of studies of students desegregated at first grade or kindergarten are longitudinal studies with non-random assignment. These are generally the most difficult studies to draw conclusions from, because the inability to use accurate pretests with very young children makes statistical matching extremely difficult. In the two best studies, by Louis Anderson (1966) of Nashville's early freedom-of choice plan, and Louise Moore (1971) of DeKalb county, GA, the full data was provided making it possible for Mahard and me to reanalyze the data. In both cases we examined student growth during the middle of elementary school, comparing growth rates for students who had experienced desegregation from kindergarten or first grade to other students in segregated schools in earlier years. One study showed a sizeable increase in the rate of learning while the other study showed a loss after desegregation. We were reluctant to take either study seriously, since we are not sure how to relate these two studies of growth rates several years after desegregation to all the other studies, which measure growth immediately following desegregation. Five other studies pretested students at kindergarten or first grade and posttested them one or two years later. These are usually very brief reports of studies with relatively small sample sizes.

Orrin Bowman's (1973) dissertation evaluates a voluntary plan in Rochester, NY. Two experimental groups exceed the controls (both a regular class and an "enriched" class) by .18 and .32 standard deviations on a readiness test at grade 1; at grade 3 they exceed the controls on an achievement battery by .90 and .88 standard deviations. Bowman's analysis of covariance shows net effects of .75 and .70; using the panel's procedure, I get effects of .72 and .66. There are only 19 and 17 treatment subjects. Ann Danaby (1971) compared 41 volunteers for desegregation to a control group randomly chosen from a segregated school. Little raw data is provided. The author uses regression to control on the seemingly large pretest differences on the Metropolitan Readiness Test, and obtains non-significant positive treatment effects. The technique used overestimates treatment effects, however.

Robert Frary and Thomas Goolsby (1979) compare 32 desegregated first graders to 77 in segregated schools, using the Metropolitan Readiness Test as a pretest and Metropolitan Achievement Test administered at the end of first grade as a posttest. There were large differences (on the order of .7 years) favoring the desegregated students. The pretest data was used to trichotomise the sample before comparing posttest means within each group. Elmer Lemke (1979), studying Peoria, Illinois, studied 180 desegregated and 60 segregated black schools five years after desegregation began. He used the Metropolitan Readiness Test and

*A ninth study, from Jefferson County (Louisville) KY., shows an increase in black scores in the elementary grades after desegregation. See Raymond, 1980. We received it too late to include in our review.

the Iowa Test of Basic Skills, and found only one significant positive effect and no significant negative effects out of a possible ten differences; we judged the overall effect as zero. T. G. Wolman (1964) studied New Rochelle, using the MAT to pretest and posttest desegregated and segregated elementary school students and the Metropolitan Readiness test to pretest and posttest kindergarten students. He reports no significant desegregation effects on the MAT, but significant gains for kindergarten students. He reports none of the data, however. Of these five studies, only Bowman is included in panel's group of 19. The other 4 studies were rejected either because they used different tests for pretest and posttest or because insufficient statistics were provided in the write-up to permit us to compute an effect size. In my judgment none of these 5 studies should be considered of especially good quality.

Conclusions

It is stretching a point to argue that the twenty kindergarten-first grade studies are the "best" studies, given their wide range of quality. They were not selected as models of research, but because they gave what we thought were the least biased estimates of the effect of desegregation. We do believe that several of these studies are better than the average of the panel's selections, which were supposedly intended to be the "best," but we are not conducting a prize competition for best dissertation* of the last two decades. We are trying to estimate the effects of desegregation.

Our 20 "best" studies include 5 analyses of four different experimental designs, all showing relatively large positive treatment effects (the median treatment effect size of these experiments is .34 standard deviations). We also found 8 "historical control groups" studies, six of which showed a positive treatment effect and only 1 a negative effect; the median effect size was .12 standard deviations. Finally, we found 7 longitudinal studies, five of which showed positive treatment effects and only one a negative effect, with a median effect size of .24. Consistent positive outcomes on 5 analyses of randomized experiments is impressive. While the other studies are a good deal weaker methodologically, their results are also consistently positive--11 studies of 15 are positive and only 2 are negative. If the principle function of selecting a superior subgroup of studies is to find the consistency of results which is masked by error in an unselected sample of studies, we believe we did that, and that the panel did not.

*One of the 93 studies, a dissertation by Ann Linney (1979) did win a prize from the American Psychological Association; it was not included in either the panel's group of 19 or our list of 20.

References

Anderson, L.V. 1966 — The Effect of Desegregation on the Achievement and Personality Patterns of Negro children. Ph.D. dissertation, George Peabody College for Teachers (University Microfilms No. 66-11237).

Armor, David 1983 — "Standard Deviation Estimates and Other Issues." (typed)

Arror, Kenneth J. 1951 — Social Choice and Individual Values. New York: Wiley.

Bowman, O. E. 1973 — Scholastic Development of Disadvantaged Negro Pupils: A Study of Pupils in Selected Segregated and Desegregated Elementary Classrooms. Ph.D. dissertation, State University of New York at Buffalo (University microfilms No. 73-19176).

Carrigan, P.M. 1969 — School Desegregation via Compulsory Pupil Transfer: Early Effects on Elementary School Children. Ann Arbor, MI: Ann Arbor Public Schools.

Chenault, G.S. 1976 — The Impact of Court-ordered Desegregation on Student Achievement. Ph.D. dissertation, University of Iowa (University Microfilms No. 77-13068).

Clark Co. 1975 — School Dist. Desegregation Report. Las Vegas, NV: Author.

Crain, Robert L Rita E. Mahard 1982 — Desegregation Plans that Raise Black Achievement: A Review of the Research. N-1844-NIE Santa Monica: The Rand Corp.

Dambacher, A.D. 1971 — A Comparison of Achievement Test Scores Made by Berkeley Elementary Students Pre and Post Integration Eras, 1967-1970. Berkeley, CA: Berkeley Unified School District.

Danahy, A.H. 1971 — A Study of the Effects of Busing on the Achievement, Attendance, Attitudes, and Social Choice of Negro Inner City Children. Ph.D. dissertation, University of Minnesota (University Microfilms No. 72-14285).

Frary, R.B., and T.M. Goolsby Jr. — "Achievement of Integrated and Segregated Negro and White First Graders in a Southern City." Integrated Education 8, 4: 48-52.

Hsia, Jayjai 1971 — Integration in Evanston, 1967-1971: A Longitudinal Evaluation. Evanston, IL: Educational Testing Service Midwestern Office.

References (continued)

Krol. R.
1978
A Meta-analysis of Comparative Research on the Effects of Desegregation on Academic Achievement. Ph.D. dissertation, Western Michigan University (University Microfilms No. 79-07962)

Kurtz, H.
1975
The Educational and Demographic Consequences of Four Years of School Desegregation in the Pasadena Unified School District. Pasadena, CA: Pasadena Unified School District.

Lemke, E.A.
1979
"The Effects of Busing on the Achievement of White and Black Students." *Educational Studies* 9: 401-406.

Linney. A.
1978
A Multivariate, Multilevel Analysis of a Midwestern City's Court Ordered Desegregation. Ph.D. dissertation, University of Illincis - Urbana-Champaign.

Luneman, A.
1973
"Desegregation and Student Achievement: A Cross-sectional and Semi-longitudinal Look at Berkeley, California." *Journal of Negro Education* 42: 439-446.

Mahan, T.W.
1971
"The Impact of Schools on Learning: Inner City Children in Suburban Schools." *Journal of School Psychology* 9, 1:1-11.

Moore, L.
1971
The Relationship of Selected Pupil and School Variables and the Reading Achievement of Third-year Primary Pupils in a Desegregated School Setting. Ph.D. dissertation, University of Georgia (University Microfilms No. 72-11018).

Nashville-Davidson County Public Schools
1979
Achievement Performance over Seven Years. Nashville, TN: Author.

Raymond. L.
1980
"Busing: Five Years Later - Test Score Trends: Blacks Gain, Whites Hold." *Louisville Times* (May 13).

Rock. W.C., and J.E. Lang, H.R. Goldberg L.W. Heinrich
1968
A Report on a Cooperative Program Between A City School District and a Suburban School District. Rochester, N.Y. City School District.

Stephan, Walter G.
1983
"Blacks and Brown: The Effect of School Desegregation on Black Students." (typed)

References (continued)

Wolman, T.G.
1964
"Learning Effects of Integration in New Rochelle." *Integrated Education* 2, 6: 30-31.

Wood, B.H.
1968
The Effects of Busing on the Intellectual Functioning of Inner City, Disadvantaged Elementary School Children. Ph.D. dissertation, University of Massachusetts (University Microfilms No. 69-5186).

Wortman, Paul M.
1983
"Scho 1 Desegregation and Black Achievement: A Meta-analysis." (typed)

Zdep, S.M.
1971
"Educating Disadvantaged Urban Children in Suburban Schools: An Evaluation." *Journal of Applied Social Psychology* 1,2: 173-186.

School Desegregation as a Social Reform: A Meta-Analysis of its Effects on Black Academic Achievement

Norman Miller and Michael Carlson
University of Southern California

INTRODUCTION

This paper addresses the specific question of what effect school desegregation has had on the achievement test scores of black children. It is one of a common set of papers addressing this issue, all prepared for the National Institute of Education. All of the papers base their conclusions and analyses on the same set of core studies that the panel of experts, selected by NIE to perform the review task, have agreed upon as meeting certain criteria for inclusion among those to be reviewed.

Before summarizing the results of these core studies, it is important first to put the question itself into an historical context, and second, to discuss the criteria for inclusion and exclusion of studies and the procedures used in performing the analysis. Then, after presenting the findings, their meaning and policy implications will be discussed.

BACKGROUND

School desegregation was initiated to address a social inequity--the impairment of minority children's right to equal educational opportunity. The Brown decision required school desegregation as a remedy for prior discrimination, declaring separate facilities inherently unequal. It is important to note that in the view of Brown, educational outcome is not the issue. Had it been shown that blacks in segregated schools performed on standardized tests as well as did whites in segregated schools, inequality of educational opportunity would nevertheless prevail according to Brown. This is not to deny that the evidence of social scientists that was presented in the case did focus on inequalities between black and white children in their self-concepts, motivation, and academic performance. In its ruling, however, the court seem concerned primarily with the notion that segregated schooling ineluctably stigmatized blacks as a social group.

> "Does segregation of children in public schools solely on the basis of race, even though the physical facilities and other 'tangible' factors may be equal, deprive the children of the minority group of equal educational opportunities? We believe that it does...to separate Negro school children from others of similar age and qualifications solely because of their race generates a feeling of inferiority as to their status in the community that may affect their hearts and minds in a way unlikely ever to be undone...in the field of public education the doctrine 'separate but equal' has no place. Separate educational facilities are inherently unequal.

> Segregation of white and colored children in public schools has a detrimental effect upon the colored children. The impact is greater when it has the sanction of the law; for the policy of separating the races is usually interpreted as denoting the inferiority of the Negro group" (Brown v. The Board of Education, 1954).

The fact of educational separation was the problem to be cured; the cure was desegregation. In principle, this logic is simple and straightforward; it requires no other major ingredients (such as, for instance, proof that desegregation will eliminate or reduce wage inequities, or other specific differences in the outcomes of blacks and whites). Of course, when school desegregation was implemented in specific cities and school districts, the method and degree of desegregation became important issues. Presumably, in court-mandated plans, the extensiveness of a court imposed remedy should in some degree correspond to the severity or magnitude of the acts that created segregated schooling (Black, 1960; Kluger, 1977).

Americans are basically sympathetic to the plight of blacks. They know that despite the beneficial social changes for blacks that have occurred over the past decades, discrimination exists and most believe it wrong. Most believe that the full weight of the Federal government should be martialed in order to eliminate such injustice. Two decades ago, 91 percent of whites favored equal voting rights, 87 percent favored the right to a fair jury trial and non-segregated public transportation, and 72 percent favored integrated education. Despite the fact that white Americans by a margin of 2 to 1 felt in 1966 that black children would not be better educated in integrated classrooms, they had no deep aversion to black children attending the same school as their own offspring. By a margin greater than 3 to 1, they denied that the education of white children would suffer if blacks are in their classroom. Three out of four white Americans approved of the Court ruling outlawing segregation in education (Brink & Harris, 1966, p. 131). There is, of course, substantial slippage between belief and action. Despite this endorsement of the moral aspects of court rulings, most whites may not be inclined to do anything specific about helping to bring about integration in schools.

In viewing the courts' position, legal scholars have noted that the remedy or restitution (viz.. desegregation) was often imposed on parties other than either the perpetrators of segregation (for instance, the school board that created it) or on their victims (those who graduated from the segregated school system). This characteristic of legally imposed remedies has led some legal analysts to interpret the underlying legal principle or goal not as restitution to the injured party, but instead, as group protection. Child labor laws or minimum age drinking laws might be other instances of the same principal. For a discussion of this view, see Yudof's (1980) interpretation and discussion of Dworkin (1970).

Since the time of Brown, social science seems to have concerned itself with the specific effects of desegregated schooling on black academic achievement, black self-concepts, and on interracial hostility and prejudice. Although these three issues were prominent in the social science statement appended to Brown, they are not the same as racial separation and stigmatization. Among the three, the one that most closely approaches stigmatization in meaning, or is most directly related to it, is intergroup hostility and prejudice. It should be noted, however, that hostility and prejudice do not necessarily denote stigmatization. Although ingroup bias is ubiquitous in intergroup relations, not all or even most outgroups are stigmatized. We frequently encounter outgroups in our daily lives. Common examples of reciprocal ingroup-outgroup pairs might be: production and sales personnel in a particular manufacturing company; two fraternities on a university campus; two teams in a baseball little league; members of opposing political parties; etc. Yet ordinarily, none of these groups are stigmatized by each other.

The point here is that the issues that have concerned social scientists, namely, low academic achievement and poor self-concepts among black children, if not prejudice as well, are not the causes of stigmatization. As implied by Campbell's argument, even if the directions of existing difference were reversed, stigmatization would persist (Campbell, 1967). The flexibility of our evaluative terminology allows any direction of difference to be positively labeled when describing ingroup members and negatively labeled when depicting outgroups. ("We are firm; they are pigheaded.") Thus, to the extent that racial-ethnic differences in academic achievement and self-concept exist, it makes more sense to view them as consequences than as causes of stigmatization. And if they are consequences, they certainly are not the only ones. Other possible consequences are wage inequities, inequalities in employment rates, lower voter turnout among blacks, higher death and disease rates, etc.

SOCIAL SCIENCE RESEARCH ON SCHOOL DESEGREGATION

In their research on school desegregation, why have social scientists focused their attention primarily on its effects on black academic achievement and black self-esteem? Perhaps in part they took their instruction from the emphasis found in the social science statement that was appended to the plaintiffs' case in Brown, which put impairment of black children's self-concept as the most pivotal or central consequence of black stigmatization, and viewed other consequences as flowing from or being caused by this key deficiency (Stephan, 1978).

The fact that studies of the effect of school desegregation on academic achievement, however, are so much more prevalent than those of any other variable reflects two additional factors. First, it undoubtedly reflects the fact that measures of academic achievement are so routinely administered by school districts. Second, such measures are very readily seen as central to the educational mission. This makes such studies more appealing to administrators who must approve the researcher's intrusion into school activities and/or records, but also, to the public as well.

The courts, too, seem to have been responsive to this manifest connection. Despite the fact that some research suggests that education contributes relatively little to one's life outcomes (Jencks, Smith, Bane, Cohen, Gintis, Heynes, & Michelson, 1972), the California State Supreme Court (Crawford, 1975) viewed desegregated education as a means of increasing the social mobility of minorities, presumably by providing better education and higher levels of cognitive mastery to minority students. Yet, Cook (1979), who was one of the authors of the social science statement appended to Brown, states that it "nowhere predicted improvement in the school achievement of black children as a consequence of desegregation" (Cook, 1979). Nevertheless, it is clear that courts as well as social scientists, have been interested not merely in the fact of segregated schooling, but also, in the effects of desegregated schooling on minority children.

Two problems have made it difficult for social scientists to provide answers about the effect of school desegregation. The first is the ambiguity in the meaning of the term "school desegregation." The second stems from the quality and characteristics of the research designs used to study it.

The definition of school desegregation. At first thought, the meaning of the term "school desegregation" seems straightforward. An analysis of how school desegregation has been implemented in any set of communities or cities, however, reveals substantial variability. Thus, the meaning of the term is in fact vague. The only common definitional element among studies of its effects is that the ratio of minority and white students in a classroom or school has been altered. By how much? Are the whites in a classroom more or less numerous than the blacks? Is the percentage of minority students in the class or school changed from 98 percent to 45 percent? Are the changes in percentages made in all classes, or just at certain grade levels or programs within the school? Are both groups of children shifted to new schools or is just one of the groups? Is the teacher familiar to one or both groups of students or do the students have a new and unfamiliar teacher? Do both groups retain friends from the previous year in their class? To what extent have other important factors other than the ratio of white to minority students also been altered (e.g., the curriculum, the student teacher ratio, the quality of physical facilities, the quality of teaching materials, the quality of teachers, etc.)?

The problems created by an ambiguous definition can be illustrated by an analogy. Consider the question "Is eating food good for humans?" Although on first thought the answer is obviously "yes," we can quickly see that the answer will depend on what is eaten and how. If the chicken salad has "turned", or the plate it is served on is lead-contaminated, then the answer becomes, "no." If a child is fed only an ounce of food three times a day or the food is merely rubbed on the child's stomach, it will starve. It might also starve if the only food available were unpalatable (e.g., half-digested dog food taken from a dog's stomach). A nutritionally balanced high-protein drink may sustain life but also cause one's teeth to drop out. Extended hospitalization for malnutrition might give one bed sores.

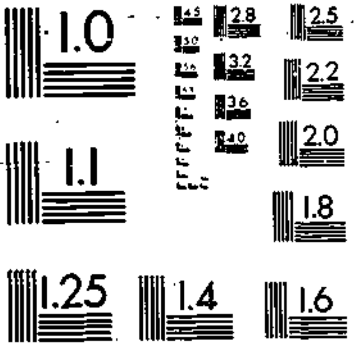
The examples above are not the "ordinary" instances of eating, But what are the "ordinary" instances of school desegregation? There are numerous circumstances in which few would expect desegrega.ed schooling to produce academic gains for blacks: e.g., when teachers, students, or principals in receiving schools are prejudiced against blacks (the food is poisoned); when there is only one or two of them in classroom, or when they are ignored in the classroom (too little food to provide nourishment); when the curriculum is not modified to match their current performance level, and consequently is not assimilated (food is rubbed on their stomach); when they are made to feel rejected and incompetent (the food is unpalatable). On the other hand, it may produce academic gains but, simultaneously, as a consequence of exposure to higher performing classmates, lower their academic self-concepts (bed-sores).

Americans may feel it is better or more moral to ship government overstocks of potatoes to an undernourished third-world country than to dump them in the ocean. As we have learned in the past, however, shipping food to people is not the same as nourishing them. Potatoes won't help if they arrive rotten, or if the receiving country lacks adequate mechanisms for distributing them. Nor will they help if protein deficiency is the problem. But nevertheless, despite our failure to achieve the goal of nourishing a famine-plagued third world country, we might feel righteous about our efforts.

Simply put, many factors are relevant to school outcomes. Those factors that go hand in hand with desegregation in one setting may not in the next. Consequently, the meaning of the term varies from one study to the next, and often, in ways that are important but not well documented.

Research designs in studies of school desegregation. As indicated, a second problem in assessing the effects of school desegregation is that researchers have rarely used a methodology that permits inferences about what it was that caused some observable d_fference between comparison groups (segregated and desegregated students). This issue is quite separate from the previous one, which pointed to the variation in the meaning of the term desegregation and covariation of other factors with implementation of a change in the ratio of blacks to whites in a school. It refers instead to the fact that children, classrooms, or schools are almost never randomly assigned to comparison conditions. As a result, one cannot know whether initial differences between the groups account for (or cause) the differences found after the treatment (desegregated schooling).

Experts are agreed that attempts to select out from, (a) those students who continue to have segregated schooling and (b) those students who change to desegregated schooling, two subsets of children that are matched (or on the average equal) on key variables on which they were originally matched, they will again differ from each other in the direction in which they initially differed.[1] Similarly, they will also differ on variables correlated with the variable on which they were matched. Consequently, if, for instance, a high IQ implies better ability to learn, and if prior to their desegregation the average IQ of the desegregated students exceeded that of those who remained segregated, they might well perform better after desegregation. Such a difference might just as readily be attributed to the initial



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS
STANDARD REFERENCE MATERIAL 1010a
(ANSI and ISO TEST CHART No. 2)

difference in IQ as to the difference in type of schooling. Why might students with higher IQ's naturally appear more frequently in the desegregated group? Parents and children who are brighter may be more motivated to seek out better schools. If they believe desegregated education to be superior, they will push to be in that program, to be included sooner in the desegregated group, or to be assigned to the desegregated school, etc., (e.g., Gerard & Miller, 1975).

METHODOLOGICAL CONSIDERATIONS FOR SUMMARIZING THE NIE SET OF STUDIES

PROCEDURES FOR COMBINING THE RESULTS OF STUDIES

Several different methods exist for summarizing the outcomes of a group of studies. Recently these procedures have come to be called meta-analysis (Glass, 1976). One procedure is simply to tally the number of studies giving positive versus negative effects. This box score or voting approach is crude because it fails, for instance, to acknowledge differences among studies in the strength or magnitude of difference between comparison conditions. Almost no experts now advocate the voting method alone (Hunter, Schmidt, & Jackson, 1982). Furthermore, the voting or box score method can lead to erroneous conclusions due to "'false' conflicting results" in the literature (Hunter et al. p. 132).

The z-score method provides an alternative procedure for representing the size of the relationship between the treatment variable and the dependent measures in a given study. It requires computing the exact *p* of the statistic employed by the original researcher (and dividing it in half if a two-tailed test was employed) and then converting each *p* value to an exact *z*-score, based on the normal probability distribution. The sum of these *z*-scores across studies is then divided by the square root of the number of findings included to generate an overall *z*-score and its associated probability level. This provides an estimate of overall statistical significance, assessing the likelihood that the results of the entire pool of studies reflect chance outcomes. (This particular procedure typically understates significant effects because many authors do not include specific *t*, *F*, or x^2 values in their research reports, and as a result, nominal rather than exact *p* values have to be entered into the analysis.) With this method, a fail-safe *n* can be calculated to determine the number of additional studies with summed *z*-scores that total to zero which would be needed before the probability value associated with the overall *z* would exceed the .05 level.

The effect size method is the most preferred method and the one used for this paper. In this method, the difference between the means of pairs of treatment conditions in each study is divided by the within-group standard deviation of the outcome measure employed, thus yielding a standardized mean difference score (Glass, 1977). These difference scores can then be averaged across studies in order to generate an overall effect size estimate.

EVALUATING THE STRENGTH OF RESEARCH DESIGNS

Apart from generating summary estimates of overall effects, meta-analysis procedures can in principle be utilized to assess whether characteristics of research design and/or program implementation features are related to program effectiveness. For this purpose, characteristics of subjects, studies, and programs must be coded and then entered as predictors in multiple regression analyses, with estimates of size of effects as the dependent variable. Examples of such predictor variables might be factors such as age of program recipients, nature of the experimental design employed in the study, the extent of parental involvement in the program, etc. In general, the search for such predictor or moderator variables is highly prone to capitalization on chance unless the number of studies is very large. In the present case, many statistical experts might judge the number of studies as too few to justify application of this procedure.

The study selection criteria imposed by the panel attempted to eliminate particularly weak studies from consideration. This does not mean that all or even most studies that survived the weeding out imposed by application of the minimum procedures are strong studies. They are not. And typically, studies with weak research designs show stronger or more positive effects than do those with stronger designs. For instance, in a meta-analysis of the larger body of school desegregation research concerned with achievement test performance, Krol (1978) found an average effect size of +0.21 among studies with weak designs, whereas among those with stronger designs, the effect was reduced by half (+0.10). While the effects of several design factors (threats to validity) have been found to be negligible in some educational contexts (Walberg, 1981), their influence nevertheless should be assessed whenever meta-analyses are undertaken in any new research arena. By imposing the selection criteria that we did, however, most of the variation in strength of design found in the total set of nineteen studies on school desegregation and academic achievement has been eliminated.

As indicated above, in addition to analyses involving research design considerations, it is ordinarily important to separate studies in terms of variables associated with the strength of program implementation. For this purpose, studies ideally should be rated or classified on implementation variables independently of knowledge of their outcomes. Unfortunately, the studies analyzed for this paper do not provide much information on correlates of (or strength of) the implementation of desegregation. Moreover, it is not even clear what "strength of implementation" means with respect to school desegregation.

VARIATION IN NUMBER AND TYPE OF DEPENDENT MEASURE

In the subset of studies analyzed for this report, the specific dependent measure varies from one study to the next. Not only do studies use different measures of verbal achievement, but within the same study the measure used prior to the implementation of desegregation may differ from that used later. In addition, some studies also include measures of achievement in mathematics, science, and other subjects, as well as verbal achievement.

Does it make sense to try to summarize studies whose measures of verbal achievement differ from one study to the next? It depends on the situation or problem. Although, for instance, it may make perfect sense to distinguish between vocabulary mastery and reading comprehension for some studies of educational success, in the present case there is little or no theoretical reason to expect school desegregation to differ in its impact on the two. In other words, with respect to the issue of whether school desegregation affects black academic achievement, different measures of verbal performance are conceptually interchangeable in that they all tap some aspect of the verbal component of the academic curriculum.

For the same reason, the distinction between measures of verbal achievement and mathematical (and/or other academic areas such as science) can also be ignored, being merely another instance of the same issue; again, there appears to be little theoretical reason to think desegregation might affect the several areas of mastery differently. This line of reasoning argues that a single effect size be computed across studies regardless of variation across studies in the particular dependent measure (e.g., vocabulary, reading comprehension, mathematics, social studies, etc.).

In addition to variation among studies in their dependent measure, many studies report outcomes for several dependent measures. In this case, we are not dealing just with variation across studies in their dependent measure, but with multiple outcomes on the same set of children. Here, the ideal procedure would convert the two sets of scores on each child (math and verbal achievement test score) to standard scores which would then be averaged for each child. The effect size for each study would then be computed on these averages. This results in each study contributing one value to the meta-analysis and at the same time minimizes error of measurement. Unfortunately, in the present instance this cannot readily be done because the raw score information is not available. To ignore the issue and treat the separate outcomes in math and verbal performance obtained in a single study as separate entries in the meta-analysis ignores the fact that these outcomes are not independent. Although not perfectly ideal, the best solution is to average the two effect sizes. This assures that studies with more measures are not given greater weight than those with few (or none).

MULTIPLE SUBJECT GROUPS

The same logic applies to the analysis of subgroups of multiple groups with the same study. The ideal procedure is to use an overall test across all subgroups. If this is not provided by the individual researcher, then the best alternative is to average the effect sizes computed for each subgroup.

CRITERIA FOR INCLUSION

Appendix A lists the criteria agreed upon by the NIE panel as a basis for inclusion of studies to be analyzed. These yielded a core sample of 19 studies. Only studies included in the NIE core sample were considered appropriate for meta-analysis. This requirement provides the first entry in Table 1, which details additional inclusion criteria for the present study. Given this set of core studies, a further criterion is that the proportion of blacks in the segregated control group must exceed 50%. This provision serves conceptually to tighten the notion of "segregation", and insures that the proportion of control group non-blacks in some studies will not approach the experimental group non-black proportions which are represented in others. The studies by Carrigan (1969) and Thompson & Smidchens (1979) were excluded from the analysis by this criterion.

The second part of Table 1 provides the guidelines for including the various segregated-desegregated comparisons which are contained within the 17 selected studies. The first restriction is that the Ns for both segregated and desegregated pre-and post-tests must be at least 10. This sets at least a moderate lower bound on the reliability of the estimates of sample means and standard deviations, as the precision of such estimates increases with sample size. Very small samples occasionally yield standard deviations which are only a fraction of the population value, and thereby are capable of producing highly misleading effect size estimates. A second inclusionary restriction on the particular comparisons concerns segregated control groups exposed to "enriched" or other novel types of curricula. Such control groups are not used because the resultant effect size estimates inversely reflect the efficacy of the particular special treatment employed in the "control" group. Such a situation fails to produce an acceptable test of the effects of desegregation on black achievement.

As indicated earlier, standardized achievement and ability tests of specialized content areas (e.g., social studies, science), as well as verbal and mathematical achievement, were included in the analysis. IQ comparisons were eliminated on the grounds that, in theory, a student's level of intelligence should not be especially sensitive to classroom experiences. Additionally, tests of "work study skills" were excluded because they do not correspond to any major academic content area. A further restriction noted in Table 1 is that the pretest and posttest had to measure an identical construct (e.g., "vocabulary", "arithmetic concepts"). Usually, this meant use of the same standardized tests (e.g., IOWA, Stanford, etc.--corresponding to the appropriate grade levels) for both the pretest and the posttest. However, cases in which the pretest and posttest differed, but nonetheless assessed the same construct, were also included, with the pretest means being adjusted to correspond to the posttest scale.

As noted in a preceding section, in studies of school desegregation, researchers are rarely able to assign children randomly to experimental and control conditions. The selection effects that occur sometimes result in higher test score means and larger standard deviations in experimental than in control groups prior to the onset of desegregated schooling. Therefore, it is important to attempt to correct post-measured differences so that they dc not simply reflect the initial

Table 1

Inclusion Criteria

A. Criteria for inclusion of studies:

1. Study must be included in NIE core list.

2. Segregated control group must be over 50% black.

B. Criteria for inclusion of comparisons within studies:

1. *Ns* must be larger than 10 for both segregated and desegregated conditions.

2. Segregated control group must not receive any special treatments which extend beyond the typical classroom experience (e.g. "enriched" control classes are excluded).

3. Dependent variable must consist of a verbal, math, or "other" (e.g. science, social studies) achievement or ability test which corresponds to a major content area (excluded are IQ tests and "work study skills" tests).

4. Pretests and posttests must measure an identical construct.

5. Either:

a. Posttest standard deviations (or reliable estimates from national norms or a comparable study), along with pretest to posttest mean differences for segregated and for desegregated conditions, must be present; or

b. An ANCOVA table (with pretest differences as a covariate) which reports a *t* or an *F* value for segregated vs. desegregated posttest score differences must be present.

inequivalence of the comparison groups, but instead, reflect the effect of desegregated schooling.

In order to arrive at pretest-adjusted estimates of effect size, it is necessary to possess the following information: (1) an estimate of differential experimental vs. control group pretest/posttest gain scores; and (2) an estimate of the population standard deviation. Thus, the final criterion for inclusion listed in Table 1 is the presence of these two pieces of information. These numbers typically were furnished in the form of tables containing pretest and posttest means and standard deviations for both segregated and desegregated groups. Analysis of covariance summary tables (with pretest differences as a covariate) provided an acceptable alternative source of such information. Finally, in the absence of the above sources of information, a comparison could still be included if the pretest and posttest means were reported and if the standard deviation could be estimated from either national norms or from a comparable study using the same test for the same grade-level.

COMPUTATION OF EFFECT SIZE

The calculation of effect size estimates for the included comparisons was achieved via the following formula:

$$ES_1 = \frac{\bar{X}_{E(post)} - \bar{X}_{C(post)}}{\sqrt{\frac{(N_E-1)s^2_{E(post)} + (N_C-1)s^2_{C(post)}}{N_E + N_C - 2}}} - \frac{\bar{X}_{E(pre)} - \bar{X}_{C(pre)}}{\sqrt{\frac{(N_E-1)s^2_{E(pre)} + (N_C-1)s^2_{C(pre)}}{N_E + N_C - 2}}}$$

E = Experimental (Desegregated) Group
C = Control (Non-Desegregated) Group

Effect size is defined here as the posttest desegregated vs. segregated difference in means (as expressed in pooled posttest standard units) minus the pretest desegregated vs. segregated difference in means (as expressed in pooled pretest standard units). For the estimation of population pretest and posttest standard deviations, a pooled figure is used (in preference to Glass' recommendation of using only the control group standard deviation) in order to increase the reliability of such estimates. The soundness of using a population estimate based on a pooled figure lies in the fact that preliminary tests indicated that among the NIE core studies, no overall significant difference was present between the standard deviations of the desegregated and segregated groups at either the time of the pretest or the posttest.

Fan-Spread. It is important to note that the present effect size estimation procedure eliminates any interpretative problems stemming from the "fan-spread hypothesis." According to the fan-spread notion, a widening of the difference between group means over time will be accompanied by an increase in the within group standard deviations. This implies that the difference between two group means may grow over time in the absence of any increment in the correlation between the treatment and the dependent variable (Kenny, 1975). The effect size formula used in this study, by separately standardizing the difference between means at times T_1 and T_2, permits a determination of the extent to which desegregation is associated with improvement in academic achievement over and above mere fan-spreading. The computational procedure is identical to that used by Armor (1983) for those cases in which he judges fan-spread to be present. In other cases, however, a difference arises, in that Armor pools the four estimates of standard deviation in instances in which he judges that fan-spread does not exist.

Amor's procedure contains two problems. First, fan-spread is a matter of degree. What criteria should be used to make a dichotomous judgment of "present" or "absent" and how can such a dichotomous decision be justified? A statistical test of whether standard deviations differ in a particular instance is not a satisfactory criteria, in that it sensibly could be argued that correction should also be made when differences fall just short, or somewhat short, etc., of statistical significance.

A second problem is that Armor's procedure may systematically place undue weight on pretest differences. If it assumed that fan-spread effects do not occur, (or do not all of the time), and further that the distribution of pretest vs. posttest standard deviation differences is associated with a certain degree of sampling variance (which is particularly likely here due to small sample sizes), then sampling error alone will produce a set of instances in which the pretest standard deviation is below the posttest standard deviation. This suggests that Armor's procedure may be susceptible to a bias in which only pretest standard deviations that happen to be low will be used to specifically scale pretest mean differences, while those that are higher (relative to the posttest standard deviation) will be averaged in with the posttest estimates. The net result is that pretest differences may be given a disproportionately high weighting across cases. Because the desegregated group usually shows a higher pretest mean than the segregated control group, Armor's procedure consequently can be expected to produce a lower overall estimate of effect size than the formula that we use.

In order to assess the extent to which a consideration of fan-spreading, however, is important in accounting for the results of the current sample of desegregation studies, effect size estimates were also calculated by using an alternative formula:

$$ES_2 = \frac{(\overline{X}_{E(post)} - \overline{X}_{E(pre)}) - (\overline{X}_{C(post)} - \overline{X}_{C(pre)})}{\sqrt{\dfrac{(N_E-1)S^2_{E(post)} + (N_C-1)S^2_{C(post)}}{N_E + N_C - 2}}}$$

E=Experimental (Desegregated) Group

C=Control (Non Desegregated) Group

In this formula, the desegregation vs. segregation pre-post gain score difference is divided by an estimate of standard deviation that is based on the pooled posttest figures. If the pretest standard deviations tend to be low relative to those of the posttest, and if the desegregation group tends to possess a higher mean than the control group at the time of the pretest (as is the case when the fan-spread hypothesis holds), then this formula should produce larger estimates of effect size than should the first formula. This is true because the typical pretest advantage for the desegregated students, which is subtracted from the standardized posttest difference, will be weighted more heavily in determining effect size estimates.

Effect size estimates based on analysis of covariance. For cases that only reported an ANCOVA (Analysis of Covariance) summary table, in which pretest scores served as the covariate, the following transformation procedure was used to estimate the effect size:

$$ES = t \frac{2}{\sqrt{N}} (.633)$$

where N is the combined sample size. Multiplying by .633 serves to correct for the fact that the variance of change scores tends to be lower than the variance of raw sample scores:
($S^2_{change} = 2S^2(1-r)$ as reported by Armor), with the difference being greatest for cases involving high pretest-posttest reliabilities. For the present purposes, a fairly high reliability estimate ($r=.8$) was assumed, which algebraically leads to the modification of effect size noted above.

Sample size. Some experts (e.g., Hunter, et al.) argue that a summary statistic of the effect sizes computed for the sample of studies (viz., mean effect size) should be weighted by the sample size of each study. Though there often may be good reasons to adopt this procedure, especially when summarizing experimental studies, for several reasons, it will not be used here. In experimental research, the manipulations are designed to correspond to a theoretical variable. Researchers almost routinely use manipulation checks to assess whether or not the independent variable theoretically postulated to affect the dependent measure has in fact been manipulated by the experimental operations that were employed, and if so, to assess whether it was manipulated "strongly enough." If, in a particular study, the manipulation check failed to confirm appropriate variation of the independent variable, and in addition, there were no treatment effects, no sensible scientist would want to include the study in the meta-analysis.

In contrast, as argued above, it is not clear what, if any, theoretical variable corresponds to or is conceptually linked to a change in the ratio of black and white children in a classroom (or school) and consequently, might be responsible for black achievement gains. Indeed, as indicated later in this paper, research seriously impugns any positive role for the one theoretical process postulated in the past to cause academic gains for minority students. Not knowing what underlying theoretical variable is relevant to academic gains for blacks, it makes perfect sense that such manipulation checks simply are not found in desegregation research. Consequently, one cannot know whether or not in any particular study the desegregated groups were exposed to the "key ingredients." If a study with a very large sample fails to contain these ingredients (or contains other features which produce losses in black achievement), and if this study outcome were weighted by its sample size, it might more than counterbalance the effects of other studies, which with smaller samples, produced positive effects. (In this regard, it is noteworthy that sample sizes among studies in the NIE core set vary by a margin of fifty to one.) Stating this another way, extraneous factors related to sample size, which may or may not be causal, may be correlated with effect size.

Anticipating the results, analyses show that: (1) sample size is indeed negatively correlated with effect size (r= -.404) and (2) the observed variation among effect sizes exceeds that to be expected from sampling error, suggesting that moderator variables are in fact operating. Taken together, these considerations argue strongly for the decision to weight study outcomes equally, rather than by sample size.

Correction for unreliability. In the current analysis, each effect size estimate was corrected for unreliability (following the procedures of Hunter et al., 1982). Measurement unreliability has the effect of artificially inflating the variability of scores, thereby leading to larger standard deviations and, hence, lower absolute values of effect size estimates. The unreliability correction procedure advanced by Hunter, et al., divides the estimated effect size value by the square root of the reliability coefficient of the dependent measure. In some of the cases comprising the NIE core studies, reliability coefficients were either reported directly or were readily available from national norms. For the remainder, a conservatively high reliability estimate of .95 was automatically assumed for each test. The net result of correcting for unreliability was to increase the absolute value of the particular effect size estimate by about 1.5% to 3%.

RESULTS

The results of the meta-analysis are summarized in Table 2. For each study, a mean was calculated (when possible) for each of the three types of dependent variable categories (i.e., verbal, math, and "other"). Next to each mean, in parentheses, is the number of different tests that were averaged in arriving at the figure.

Table 2

Effect Size Estimates

Study	Year	Miller and Carlson (1)[a] Verbal	Math	Other	Miller and Carlson (2) Verbal	Math	Other	Armor Verbal	Math	Stephan Verbal	Math	Wortman Verbal	Math
Anderson	1966	+.711 (1)[b]	+.669 (1)	---	+.764(1)	+.671(1)	---	+.09 (1)	+.54 (1)	+.42	+.24	+.95 (1)	+.51 (1)
Beker	1967	+.400 (2)	-.203 (1)	+1.020(1)	+.265(2)	-.101(1)	+.942(1)	+.25 (2)	-.16 (2)	+.125	-.24	+.095(2)	+.205 (2)
Bowman	1973	+.066 (1)	-.160 (2)	---	+.011(1)	-.325(2)	---	-.26 (2)	-.25 (2)	+.325	+.055	+.02 (1)	-.06 (1)
Carrigan	1969	-	---	---	---	---	---	-.068(6)	---	-.04	---	-.04 (6)	---
Clark[e]	1971	+.039 (2)	-.132 (1)	---	d	d	---	-.01 (1)	+.32 (1)	+.00	-.24	e	e
Evans	1973	+.040 (2)	+.004 (2)	---	d	d	---	+.015(2)	+.07 (2)	+.07	+.03	e	e
Iwanicki and Gable	1978	-.051 (1)	---	---	d	---	---	0.0 (1)	---	---	----	e	e
Klein[e]	1967	+.359 (1)	+.111 (1)	+.353(1)	d	d	d	0.0 (1)	-.00 (1)	+.25	+.31	e	e
Laird and Weeks	1966	+.257 (5)	+.074 (1)	---	+.341 (5)	+.127(1)	---	-.121	-.06 (1)	+.11	+.061	e	e
Moorefield	1967	+.202 (3)	-.120 (5)	---	+.439 (3)	-.225 (3)	---	---	---	+.22	-.051	+.497(3)	-.310 (1)
Savage	1971	-.080 (5)	-.091 (3)	+.344 (1)	-.077 (1)	-.034(1)	+.010(1)	+.15 (1)	-.08 (1)	+.06	-.04	+.14 (1)	-.05 (1)
Sheehan and Marcus[c]	1978	-.100 (2)	-.361 (3)	---	d	d	---	.16 (1)	+.21 (1)	-.07	-.00	-.16 (1)	-.36 (1)
Slone[c]	1968	+.091 (2)	+.291 (1)	---	d	d	---	+.27 (1)	+.47 (1)	+.19	+.22	e	e
Smith	1978	-.077 (1)	+.330 (1)	---	-.308 (1)	+.256(1)	---	-.06 (1)	+.13 (1)	-.01	+.02	-.05 (1)	+.10 (1)
Syracuse	1966	+.695 (1)	---	---	+.693 (1)	---	---	+.175(2)	---	+.75	---	e	--
Thompson and Smith-Adams	1979	---	---	---	---	---	---	---	---	-.15	+.04	e	e
Van Every	1969	-.166 (2)	+.542 (2)	--	-.257 (2)	+.559(1)	---	.46 (1)	+.55 (2)	-.32	+.14	-.40 (1)	+.55 (1)
Walberg	1971	-.049 (1)	---	---	-.040 (1)	---	---	-.10 (4)	---	+.055	-.02	+.032(1)	?
Zdep	1971	+.671 (2)	.151 (1)	---	+.662 (2)	-.157(1)	---	+.53 (1)	-.17 (1)	+.66	-.15	+.65 (1)	-.35 (1)
		N = 17 $\bar{x}$ = +.161* SD = .297	N = 10 $\bar{x}$ = +.077 SD = .284	N = 3 $\bar{x}$ = +.520 SD = .576									

For Verbal and Math:	For Verbal, Math, and Other:

n=31
$\bar{x}$=.121*
SD=.289
	n=34
	$\bar{x}$=.159**
	SD=.326

Notes:
a. See text for formulas #1 and #2.
b. Numbers in parentheses are the number of effect size comparisons.
c. Uses estimates based on ANCOVA.
d. Estimates from formulas #1 and #2 are identical due to use of ANCOVAS
e. Not pretest adjusted.
* $p < .05$; ** $p < .01$.

BEST COPY AVAILABLE

Using formula (1), the overall effect size is +.159 (see bottom of column 1, Table 2). This estimate weights results within each study equally and weights each study equally. The fact that formula (2) gives an outcome of +.155, which is essentially equivalent to that obtained with formula (1), confirms the view, presented earlier, that fan-spread is not a problem in these data.

For purposes of comparison, the effect size computations of Armor (1983), Stephan (1983), and Wortman (1983) are reported in the adjacent columns of Table 2 (columns 3, 4, and 5). Table 3 summarizes the findings of all four researchers, reporting their mean effect sizes, separately for verbal and math tests, for each study. Pooling the outcomes across researchers and studies, the effect size of +.156 for verbal tests is significant ($t=2.26$, $p < .05$), as is the pooled verbal and math effect size of +.119 ($t=2.40$, $p < .05$). The effects of desegregation on mathematics tests is smaller than that found on verbal tests (though not significantly so) and when tested separately, does not yield a significant effect size (see columns 1 and 2, and see Table 3).

Sources of Disparity in the Effect Size Estimates for Individual Studies

Comparisons of our own effect size computations with those of Armor, Stephan, and Wortman for each study reveal that they agree fairly well; the correlations, using estimates based on formula (1) are +.87, +.76 and +.74 with Armor, Stephan, and Wortman, respectively.

The correlations were computed by treating the mean verbal effect size per study and the mean math effect size per study as separate entries. The fact that the verbal and math effect size estimates are not based on independent samples is irrelevant for this computation in that it seeks to assess the comparability of effect size computations performed by independent investigators. There is little reason to think that computations performed within a study are less independent than those between studies. Despite the high correlation between estimates, the fact that these correlations are less than perfect, as well as the fact that inspection of effect sizes across the rows of Table 2 reveals variation, makes it clear that computational differences exist.

The following paragraphs, on a case by case basis, examine all instances in which our estimates differed from the mean estimate of Armor, Stephan, and Wortman by more than .1 of a standard deviation.

Anderson (Math)

Our estimate is slightly higher (+.669) than those of Armor (+.54) and Wortman (+.53), mainly as a result of discrepancy between the mean of the raw pretest segregated math scores contained in Table 26 (45.093, p. 138) and the mean he presents in his pretest summary table (43.82, p. 144). We used the mean of the raw scores, which led to a higher effect size estimate due to the inclusion of a larger segregated group pretest figure.

Table 3
Mean Effect Size Estimates

Study	Verbal	Math	
Anderson	+ .75	+ .49	
Baker	+ .22	- .08	
Bowman	+ .01	- .09	
Carrigan	[+ .049	---]	
Clark	+ .04	- .16	
Evans	+ .03	+ .06	
Iwan. & Gable	+ .03	---	
Klein	+ .13	+ .19	
Laird & Weeks	+ .24	+ .03	
Rentsch	+ .31	- .10	
Savage	+ .07	- .07	
Sheeh. and Marcus	- .14	- .15	
Stone	+ .18	+ .33	
Smith	- .05	+ .10	
Syracuse	+ .61	---	
Thompson & Smid	[- .15	+ .04]	
Van Every	- .30	+ .43	
Walberg	+ .02	- .02	
Zdep[a]	+ .63	- .16	
			Combined V&M
N	17	15	32
$\bar{X}$	+ .156[b]	+.053	+.108[c]
SD	.284	.215	.255

[a] Entries combine the computations of Miller (#1), with those of Armor, Stephan, and Wortman. Excludes Carrigan, Thompson and Smidchens.

[b] $t_{(16)} = 2.26, p < .05$

[c] $t_{(31)} = 2.40, p < .05$

BEST COPY AVAILABLE

Beker (Verbal)

The major reason for our higher estimate seems to be our inclusion of a wider array of tests (spelling, word meaning, language, and vocabulary) which demonstrated larger positive effects than did paragraph meaning. Wortman's estimate is additionally lower due to his exclusive use of the "refused transfer" controls instead of the "requested transfer" group.

Klein (Math)

Our estimate for math agrees with that of Stephan (+.33), but is substantially higher than Armor's (-.08). The reason for the discrepancy is that we used only the "random" control group, while Armor used only the "matched" control group. The matched controls were excluded from the present analysis because the corresponding ANCOVA summary table mixes the data for the segregated and desegregated blacks along with that of the white students.

Rentsch (Verbal)

Our verbal effect size estimate, though quite close to Stephan, is lower than that of Wortman. This is primarily due to Wortman's use of the "abnormally low" pretest standard deviations (see in particular the control group). His use of Glass' formulas creates this outcome. Our own formula #2 outcome, which lacks sensitivity to temporal changes in standard deviations, yields, as expected, a result much closer to Wortman's.

Savage (Verbal)

Our estimate for verbal achievement (-.08) is both lower than and in the opposite direction of the mean of the estimates of Armor, Wortman, and Stephen (+.117). The sole reason for this appears to be our inclusion of STEP Writing (+.048) and STEP Listening (-.437) in arriving at a verbal effect size estimate. Our figure for Reading (+.150) agrees perfectly with Armor's estimate and differs from Wortman's by only .01.

Slone (Verbal)

Our estimate of .091 is somewhat lower than that of both Armor (+.27) and Stephan (+.19). This is because in addition to Reading (+.242, which is fairly close to the other estimates) we included the Language Skills test (-.061).

Syracuse (Verbal)

Our figure for the Syracuse report (+.691), while relatively close to Stephan's estimate (+.75), is much higher than Armor's (+.375). The reason is that Armor includes a second comparison (which we excluded because of missing standard deviations) in which the effect size was essentially zero.

Van Every (Verbal and Math)

Our estimate for verbal achievement (-.166) is somewhat less negative than the estimates of Armor (-.46) and of Wortman (-.44). This is because they only consider Reading (which we estimated to be -.468), whereas we additionally included Language Arts (+.137).

Our math estimate is nearly identical to those of Armor and Wortman, and differs significantly only from Stephan's figure. Stephan's lower estimate most likely stems from his use of Glassian formulas, in conjunction with his correction procedure for the amount of time elapsing between the pretest and the posttest.

Walberg (General Note)

Due to problems in the legibility of our copy of this report, we were unable to calculate a verbal effect size estimate for the 10-12th grade group, as well as any estimates for math achievement.

Sources of Disparity in Overall Effect Size Estimates

Among the three NIE panel members' computed effect size estimates, Armor's overall effect size estimate of +.077 is most discrepant from our own. Consequently, his computations were chosen as a basis for estimating sources of discrepancy.

Table 4 presents an analysis of the disparity. It shows that correction for unreliability in the dependent measures is not a major contributor to our higher estimate. In part, this is due to the fact that conservatively high reliability estimates (viz.,.95) were assumed for the studies for which no reliability was reported. Reliability estimates provided by test publishers do not report separate reliability estimate for blacks, but were they available, they are likely to be lower than those reported for whites. In sum, a less conservative and more realistic correction for unreliability would yield a larger, more positive overall effect size estimate.

The factor responsible for the largest portion of the difference (approximately 45%) was our inclusion of results on achievement tests on content other than verbal skills and mathematics. It is worth noting that although only three studies report such results, the mean effect size (and its standard deviation) is substantially larger than that of effect sizes based on verbal and mathematics tests.

Moderator Variables

Ordinarily, with such a small set of studies, it is hard to justify a search for variables that explain the relation between the independent (school desegregation) and dependent (academic achievement) variables. A simple set of computations, however, can suggest whether such a search will be fruitful. The variance of the effect sizes over the sample studies can be computed and corrected for sampling error. If the effect sizes are really identical and vary only because of sampling error (i.e., they are simply random deviations from the true mean value), then the "true variance" of the effect sizes would be zero. Hunter, et al., provide formulas for computing the variance of an array of effect sizes, corrected for sampling error. When sampling variability (σ^2_{error}) is

Table 4

Analysis of Discrepancy Between Effect Size Estimates of Armor and Miller and Carlson (#1)[a]

Source	Contributions
Inclusion of Reliability Correction	+ .005
Inclusion of Rentsch	- .008
Inclusion of "other" category data	+ .0358
Averaging in of extra tests excluded by Armor	+ .002
Calculational differences on same non-Ancova cases	+ .006
Calculational differences on cases where we estimated from Ancova	- .006
Different comparison groups used in same study (Klein)	+ .0172
Armor~~'s inclusion of Carrigan Study~~	+ .005
Cases within studies included only by Armor	+ .022
Total: =	+ .079
(Miller and Carlson + .159) - (Armor + .077) =	+ .082
Unaccounted difference =	+ .003

Note:

a. Table entries are based on overall means of Miller and Carlson's Verbal, Math, and "Other" tests.

removed from the computed variance among obtained effect sizes (σ^2_{ES}) there should be no residual (viz. $\sigma^2_{ES} - \sigma^2_{error} = 0$) if, in fact, the effect size is really the same across studies. If, on the other hand, the residual vsriation is large, especially if large in comparison to the mean value, a search for moderator variables should be made.

In the present case, our effect sizes for verbal achievement tests were used to assess this issue. When sampling variability is removed, the residual variance does not approximate zero.

$$(\quad \sigma^2_{ES} = .038; \sigma^2_{error} = .012 \quad)$$

These results show that 68% of the variance in the computed effect size scores (weighted by sample size) is unexplained by sampling error.

$$\text{Proportion of variance which is unexplainable on the basis of sampling error.} = \frac{\text{Variance ES - Variance error}}{\text{Variance ES}} = \frac{.026}{.038}$$

These results argue strongly that variation among study characteristics and not mere sampling fluctuation is responsible for the observed variation in the computed effect sizes.

Given these results, three potential moderator variables were examined: year of study, region (North vs. South), and percentage of black students in the desegregated class. Prior to computing the correlation between effect size and each potential moderator variable, we averaged our own effect size estimates with those of Armor, Stephan, and Wortman, separately for verbal and math achievement. Pooling gives a more stable estimate. Although earlier in the chapter we argued that the different content domains of academic performance should be considered indices of a common underlying construct, separate treatment of verbal and math effects is justified by the low correlation between these two effect size estimates within each study ($r= +.29$; $r^2 = + .084$; df = 12; $p > .05$), and the fact that Stephan provides a theoretical rationale for different outcomes on verbal and math tests. When the verbal and math effect sizes of Armor, Stephan, and Wortman are pooled with our own, the correlation between them is even smaller ($r = +.15$; $r^2 = +.023$; df= 12; $p > .05$).

Interestingly both verbsl and math effect size estimates correlate negatively with year of study ($r_V = -.554$ and $r_M = -.559$, $p < .05$ respectively. Region is unassociated with effect size (point biserial: $r_V = +.104$; $r_M = +.04$, north positive, $p > .05$).

There is some suggestion, however, that petcentage of blacks in the classroom is important and that it has different effects on verbal and math achievement. The correlation between percentage of black students in the class and verbal effect size is -.281. In contrast, no such effect is found for math achievement; in fact, the correlation between percentage black and math achievement, though not significant, is opposite in sign (+.310). When year of study is partialled out, the above correlations for verbal and math are equal to -.339 and +.422

respectively; the difference between them is significant ($p < .05$, one-tailed).

These results provide some support for Stephan's (1983) interpretation of his own computed effect size differences for verbal and math achievement, showing desegregation to produce essentially no benefit for the latter. He interprets the gain in black verbal achievement that is found with desegregated schooling to be a consequence of increased exposure to white speech style, syntax, grammar, etc. If this interpretation has merit, it makes sense that percentage of blacks in the classroom should be inversely related to such gains. The fewer the number of other blacks in the classroom, the more likely it is that the desegregated black child must interact with white children and the less likely it is that he or she would find a within-race peer support group in which black speech is practiced and reinforced.

Correction of Effect Size Estimates for "Overall School Improvement"

The analyses presented above examine the achievement gains of desegregated black children but ignore changes among their white classmates. It is important to examine the latter, however, because when both groups gain (or lose), it suggests that it is not desegregation per se that is responsible for the effect, but instead, some other factor that has affected the school or school district as a whole, thereby improving the academic performance of all of its students. Such factors might be: influx of new funding; improved curriculum materials; a new principal; renewed teacher enthusiasm; increased emphasis on preparation for state-mandated testing; or whatever.

Those sympathetic to the idea of desegregation might contend that when school changes such as those cited above appear hand in hand with desegregation, they should not be viewed as confounding effects, that is, as factors other than desegregated schooling that explain the observed minority gains. Instead, they should be thought of as natural covariates of desegregation, that is, as part of the meaniug of the term. In other words, according to this line of thought, whenever one desegregates a school or school district these simultaneous changes (whatever they are, and however unspecified they must remain) can be expected to co-occur with the change in the ratio of black and white students. And as long as they regularly or naturally co-occur with desegregation, their academic benefits to minority children can be attributed to desegregation. In this view, if whites gain along with blacks, all the better.

There are two problems with this line of thought. One lies in the validity of the assumption that these school changes can be expected to co-occur routinely with desegregation in the future (or in other unsampled districts). For instance, today, in an era of minimal availability of increased state and federal funding for schools, some of these mediating factors (e.g., new or improved curriculum and/or text materials, or lower pupil-teacher ratios) may no longer be readily available to desegregating districts. Similarly, 15 years ago teachers and principals may well have been more inclined to expect positive outcomes as a consequence of desegregation than they do today. Such expectancies have often been found to be self-fulfilling for one reason or another. If present then, but not today, outcomes would again differ depending on whether one included or excluded such factors in one's definition and implementation of desegregation. The strong negative correlations reported above between year of study and positivity of both verbal and math effect size estimates argues strongly that one cannot rely routinely on the natural occurrence of these beneficial ingredients.

A second problem lies in one's definition of academic benefit. Some scholars argue that benefit should be defined in an absolute sense. If desegregation produces academic gains for blacks, and does not produce losses for whites, it is beneficial. In this view, it does not matter if the gains of white children equal or exceed those of blacks. An alternate view focuses instead on the closing of the academic achievement gap. Consequently, it defines desegregation as beneficial only if the gains of black children exceed those of whites.

Three studies in the NIE core set, Beker (1967), Clark (1971), and Laird and Weeks (1966), provide data that permits analysis of the effects of desegregation on white as well as black children. All seven available cases of the mean verbal, math, or "other test" effect size per study can be compared by using the following formula:

$$\left(\begin{array}{c} \text{Desegre-} \\ \text{gated} \\ \text{blacks} \end{array} : \frac{X \text{ post} - X \text{ pre}}{\text{pooled pre + post DD}} \right) - \left(\begin{array}{c} \text{Receiving} \\ \text{School} \\ \text{whites} \end{array} : \frac{X \text{ post} - X \text{ pre}}{\text{pooled pre + post ED}} \right)$$

The resulting difference in effect sizes is -.379, (N=7, $p > .05$, S.D.=.894). Although not significant with only seven cases, the direction of effect shows that the gains of white children in the receiving schools of these studies substantially exceeded those of black children, which were roughly of the same positive magnitude as the gains found for the entire sample of blacks. That is, the mean effect size for blacks in these three studies (weighting tests equally) was +.15, (compared to the entire sample effect size of +.159), whereas the effect size for whites was +.52. In other words, the achievement gains of white children in these three studies were more than three times as large in standard units as those of their black classmates.

In summary, on the basis of this extremely small subsample, it appears that black gains relative to white gains were small. In terms of the preceding discussion, these data suggest that the observed gains of desegregated black children are not attributable to the presence of white classmates per se. Instead, they appear due to more general improvements in schools or districts that occur during the implementation of desegregation.

DISCUSSION

Interpretation of the Obtained Effect Size

How does one interpret a mean effect size of +.159? In magnitude, it approaches the +.20 effect size that Walberg (1983) states is "average" for various educational interventions. Thus, on this basis the effects of desegregation are relatively similar to other attempts to improve educational outcomes. Two points, however, bear reiteration with respect to this conclusion. First, as argued earlier, desegregation is not an educational program in the sense, for instance, that are many of the interventions examined in the Michigan group's quantitative summaries (Kulik, Shwalb, and Kulik, 1983; Cohen, and Ebeling, 1980; Kulik, Kulik, and Cohen, 1979; Kulik, Kulik, and Cohen, 1980). Computer-based instruction, individualized instruction, open classrooms, tutorial programs, Bloom's mastery learning, etc., all presumably improve educational performance as a consequence of identifiable independent variables that comprise the program. The same cannot be said for school desegregation. At this point in time, we have not yet identified an underlying social psychological process which, as a result of a change in the ratio of black and white students in a classroom or school, will augment minority scholastic achievement. Second, as implied by our analyses pointing to moderator variables and as suggested by our analyses of white student outcomes, when benefit to black students is found, it is not attributable to desegregation per se, but instead, to other school or district factors that accompany its implementation.

Factors Affecting Academic Outcomes in Desegregated Settings

As stated above, there is little good theoretical understanding of how desegregated schooling might improve the academic performance of minority children. Much past theorizing has not withstood the test of data. The next section briefly discusses an array of factors, some of which were thought in the past to be relevant and some of which continue to appear important.

Anxiety and threat. The fact that high anxiety impairs performance on complex or difficult tasks fits with common sense and is one of the better established findings of psychology. In his review of variables that affect black performance on cognitive tasks, Katz (1968) summarized substantial evidence showing impairment when performing under the scrutiny of higher status whites. The administration of standardized achievement tests to black students by a white teacher in a white dominated setting, such as a desegregated classroom, structurally parallels the situations studied and cited by Katz as impairing black performance. The fact that standardized achievement tests are administered with time limits acts to further raise anxiety. Some evidence suggests that one-way busing of blacks to white receiving schools will increase their anxiety in general, at least during the initial phases of desegregation (e.g., Gerard & Miller, 1975). Mussen (1953) found that black children perceive more hostility or threat in their environment than do whites. Baughman (1971) interprets the heightened level of worry and anxiety that black children attribute to their characters when asked to make up stories as confirming Mussen's results.

Taken together, such data implies that measured black performance is likely to be an underestimate of true mastery; it implies that the obtained effect sizes for black academic achievement do not reflect true level of achievement. But if adult black intellectual activity is performed in a white world, aren't such depressed scores in fact legitimate scores? Perhaps, but in work settings, performance is rarely under the constant scrutiny of a white supervisor.

Self-concepts and aspirations. In the social science statement appended to Brown, scholars argued that segregated schooling lowered the self-concept of the minority child and that this in turn produced a sense of defeatism, self-doubt, and lack of aspiration that interfered with effective learning. Although the argument appears credible, it has not withstood empirical analysis. Not only has the interpretation of Clark's (1937) original doll preference data on which the argument was based been questioned (Brand, Ruiz & Padilla, 1974; Banks, 1976), but recent reviews of self-esteem research that employs direct self-report measures consistently show either higher levels of self-esteem among black children than among white children or no consistent effects (Epps, 1979, Porter & Washington, 1979, St. John, 1975, Stephan, 1978, Wylie, 1979). Furthermore, if school desegregation does affect the self-esteem of black children, its effects, at least initially, are more likely adverse than positive (Porter & Washington, 1979).

Measures of aspirations present a similar picture. Black children in segregated schools typically report higher aspirations than do white students (Epps, 1975; Proshensky & Newton, 1968; Weinberg, 1975). And black adults seem to value education more strongly than do whites (Wilson, 1970). The effect of desegregated schooling on the motivation of black students remains unclear, some studies showing higher black aspirations in desegregated schools (Curtis, 1968; DeBord, Griffen, & Clark, 1977; Fisher, 1971; Knapp & Hammer, 1971, Reniston, 1973), others showing an opposite effect (St. John, 1966; White & Knight, 1973; Wilson, 1959), and still others showing little difference between black children who attend segregated or desegregated schools (Curtis, 1968; Falk, 1978; Hall & Wiant, 1973). Two points must be made with respect to this issue. First, most experts today would agree that level of aspiration per se is not as meaningful or important an indicator of a healthy personality as is a level of aspiration that is in line with one's level of performance and one's obtained outcomes. Second, the nature or design of these studies does not allow causal interpretation of whatever differences are found.

Finally, although the theorizing of social scientists at the time of Brown allowed for circular feedback loops (or bi-directional or reciprocal causations) among self-esteem, motivation and aspiration, intergroup acceptance, and academic performance, their arguments clearly emphasized a causal pattern in which personality variables (self-concept and achievement motivation) caused subsequent changes in academic performance. If there is any preponderent direction of causal effect, researchers today would emphasize the impact of school outcomes (academic performance and achievement) in forming personality or creating changes in it, rather than a causal pattern in which changes in personality cause subsequent shifts in performance (Gottfredson, 1980; Miller, 1982; Rubin, Maruyama, & Kingsly, 1979; Scheirer & Kraut, 1979).

Peer Comparison. Students know who is smart and who is not (Lippit & Gold, 1959; Hoffman & Cohen, 1972). Differences in opportunity to perform, when coupled with a narrow range of valued abilities, act to create widely shared perceptions of competence (Simpson, 1981; Rosenholtz & Rosenholtz, 1981). When black children attend desegregated rather than segregated schools, social comparisons between their own academic performance and that of white students will reveal disparities that might be expected to lower performance. If such effects occur, they should be greater at higher grade levels in that, on the average, the academic disparities between black and white students increase as they progress through school.

On the other hand, other data suggests that black children primarily compare themselves to other black children (Baughman, 1971). To the extent that the desegregation plan provides enough black children in each class to form the basis for a within-race comparison group, the debilitating effects of comparison with white children should be lessened. Moreover, children, like the rest of us, are self-protective and adaptive. They find ways to ignore self-disparaging comparisons and, as evidence on black children's self-esteem and aspirations shows, if anything, these children show high levels of self-regard and expectation in their self-reports. Whether or not these high levels are "defensively high" as suggested by Entwisle & Hayduk, (1982), and Miller, (1982), and reflect a negative consequence of peer comparison remains unclear.

Expectations. As indicated above, expectations often create self-fulfilling cycles. Expectations to perform poorly cause behavior that subsequently confirms the expectation. But expectations are intimately linked to actual behavior. Rehearsal of academic information and content improves performance on subsequent testing of the mastery of this information. It is the better student who volunteers the answer when the teacher calls for a response, who leads the discussion in peer tutoring or small work group exercises, and who the teacher routinely gives more opportunities to respond (Good, 1970). Thus, it is the better student who gets the benefit of overt rehearsal at the expense of less capable peers, thereby further improving the performance of the better student. The social dominance of whites when in interaction with blacks is well documented. Even when the resources and knowledge brought to the problem by black and white children is equivalent, the white child will initiate verbal comments more often than the black and will dominate the interaction, with the black child taking a more subordinate role (Cohen, 1982). Apparently, generalized status differences are implicit in the distinction between races. Even when black students are primed with correct information that makes them a more superior source of knowledge than the white children, the generalized status difference between blacks and whites nevertheless results in continued verbal dominance by the white children (Cohen & Roper, 1982; Tammivaara, 1982).

Peer relations. Some social scientists believed that the peer environment of the desegregated school would be critical in producing academic gains (Coleman et al. 1965; Crain & Weissman, 1972; Pettigrew, 1969). This belief rested on the assumptions that (a) the student body of a desegregated receiving school is more likely than that of a segregated school to be of middle class family background; (b) middle class students are more strongly oriented toward achievement and thereby create a normative structure that emphasizes it; and (c) provided that the number of white students in the receiving school exceeds the number of incoming minority students, the latter group will adapt to the prevailing norm structure of the middle class whites. This argument, spelled out in detail by Katz (1964), rests on the additional assumption that minority children will be accepted or befriended by white children.

The latter assumption is at best, less true than one might wish. Resegregation is common in desegregated classrooms (e.g., Rogers & Miller, 1980; Rogers & Miller, 1981; Schofield, 1980), and when white children accept minority students, it is a consequence of the minority students' good academic performance rather than a cause of it (Maruyama & Miller, 1979; Maruyama & Miller, 1983). Thus, it is not the peer system that provides a critical normative influence. Instead, as discussed in more detail below, it is provided by the teachers and administrators.

School effects. Recent research, Jencks et al. (1972) notwithstanding, shows that schools can exert powerful educational effects on students (Heyns, 1978) and differ in the extent to which they educate them (Edmonds, 1976). These effects are system or organization effects, produced in concert by principals, teachers, students, neighborhood, parents, and all having reciprocal influence on one another. This is not to argue that one cannot find, for instance, within-school differences among teachers both in their background and their approach to education, or differences among students. It startles no one when a low social class background is found to be related to a student's academic performance (Hauser, 1978). Nor does it elicit much more surprise to learn that the quality of teachers' education affects the academic outcomes of their pupils (Heim, 1970; Summers & Wolfe, 1977). More interesting, however, are the substantial differences in academic outcomes found among schools whose students are basically similar in social class background and/or race. Although some authors have argued that such school effects are small (e.g., Sewell, Haller, & Portes, 1969), the studies on which such conclusions are based all use high school samples. By high school age, self-fulfilling characteristics of background, expectations, and scholastic outcomes have homogenized schools, not unexpectedly leaving them similar in their educational impact, and consequently, leaving the false impression that the type of school attended cannot make a difference. At earlier ages, however, the homogenization process is not completed. Interestingly, studies of elementary schools do show striking differences among schools.

Two recent studies dramatically illustrate the powerful differences among schools in their effects on students (Brookover, Beady, Flood, Schweitzer, Wisenbaker, 1979; Entwisle & Hayduk, 1982). Both are very substantial in terms of their breadth and the array of measures they employ. The Brookover et al. study is based on data from over 11,000 students in the fourth and fifth grades in over 90 schools drawn by random from the entire State of Michigan. Among those, 30 are majority black schools. This exceeds the totals of students and schools in the entire array of the nineteen NIE sample desegregation studies by a margin of about 3 to 1. Entwisle and Hayduk (1982) studied approximately 1,500 children over a three-year period from first to third grade. Approximately one-third, respectively, attended a white middle class school, an integrated lower class school, and a black lower class school. Although much smaller in terms of the number of schools studied, this study measured an even broader array of variables than the Brookover et al. study and on each, took multiple (longitudinal) measurements on each child over the three-year course of the study, thereby enabling study of the temporal changes in the measured variables. It is only with temporal spacing of repeated measures on the same child that one can begin to establish the causal connection between variables. Thus, the two studies differ substantially in the characteristics of their research designs. Nevertheless, as will be indicated below, their results converge in identifying key aspects of the process of education, as well as showing that schools can produce very different outcomes for children.

Teachers. Earlier work demonstrated that teachers exert powerful effects on minority student outcomes (Johnson, Gerard, & Miller, 1975; Fraser, 1981). When desegregated minority children are imbedded in the classes of prejudiced teachers, their academic performance worsens, whereas in the classes of unprejudiced teachers, it improves (Johnson, Gerard, Miller, 1975). Furthermore, these effects can be traced to clear differences in the way in which these two types of teachers conduct their classes and interact with minority students (Frazer, 1981). This conclusion is supported by Brookover et al. and by Entwisle and Hayduk. In some lower class black schools the teachers (and the principal) have given up on the students. They do not view their students as capable of learning; attributing their poor academic outcomes to their backgrounds and not demanding good and consistent work from them. It is important to emphasize here, that it is not merely teachers' expectations that produce these effects, but instead, it is their behavior. In lower class black schools that produce poor academic outcomes, students are not expected to perform up to grade level, and demands requiring them to do so are not placed on them. When teachers judge their students to be incompetent, they do not attempt to cover as much academic material (Beez, 1970).

Teachers in most lower class schools also fail to voice concrete achievement goals. Instead, these children are often reinforced for incorrect performance, hearing the teacher say, for instance, "good try" when the answer is very clearly wrong, or not receiving immediate re-instruction when their response is incorrect (Brophy & Good, 1970). Academic norms of high academic achievement are recognized in high-achieving lower class black schools, whereas such norms and a commitment to academic mastery are missing in the low-achieving schools. In the high-achieving schools, teachers spend most of the day instructing their students, reinforcing them discriminantly rather than indiscriminantly. In these schools, teachers do not highly differentiate among students and, in the process, write off a large segment of them as unteachable.

Students. Although many factors may contribute to the greater sense of control over their outcomes in life seen in middle class as opposed to lower class children (Coleman et al. 1966), the schools they attend seem to contribute to this observed difference. The students in low-achieving schools show a legitimate sense of futility, and with reason. It is difficult for them to know what to expect, and the messages they get confuse and demoralize them. The teacher says, "Good, you're trying hard"; or "OK"; but they receive C's and D's on their report card. Consequently, their expectations are not responsibly modified by their obtained grades. In contrast to a sense of mastery and control of their academic outcomes, these students feel the system is whimsical and "stacked against them." In contrast, children in high-achieving middle class schools increasingly come to forecast their school outcomes accurately. Their expectations more closely correspond to the grades they receive, with most students predicting their marks correctly (Entwisle & Hayduk, 1982). Brookover et al. (1982) argue that a sense of control over school outcomes is one of the essential ingredients for high student achievement.

Implications of Academic Achievement Results in the Context of Educational Goals

What does one make of the moderate positive effect of desegregation on the academic achievement of black children? Although not a strong clarion for desegregation in its own right, it certainly is not a deterrent to the continuation of desegregation as a national policy. More important, however, is the fact that other valuable educational goals cannot be met without desegregated schooling. Although cognitive development and academic mastery are obviously appropriate educational goals, they are not the only ones. Despite some recent signs of increased interest in "fundamental" education, all school curricula to some degree attend to dimensions other than verbal and mathematical skills. Indeed, many components of the standard educational curriculum attend to dimensions that have little or no direct relevance to cognitive mastery (e.g., physical education; music, art, and aesthetic development; mechanical, shop, and home skills; industrial, business, and other vocational training; etc.).

In some sense all agree that schools must prepare children to function effectively in their adult lives. Thus, some view with despair the tracking of students within performance levels and in qualitatively different academic programs because it functions to prepare students for occupational and social roles that reflect their socioeconomic origins (Bowles & Gintis, 1976); and students within the different tracks do display attitudes and patterns of interpersonal behavior that are complementary to these future roles (Oakes, 1982).

Similarly, few would argue against the view that interpersonal skills are relevant to accomplishment and success in adulthood. In a multi-ethnic society, constructive modes of interethnic interaction, as well as interethnic acceptance and trust, are valuable attributes. It is both appropriate and feasible for schools to develop children's strength and facility in these directions. But schools cannot do so if children lack day-to-day contact with children whose racial-ethnic identities differ from their own. The point here is not that contact per se can be counted on to produce interethnic acceptance. Recent studies show clearly that racial-ethnic boundaries function to organize patterns of social interaction in desegregated school settings (Singleton & Asher, 1979). Furthermore, racial-ethnic encapsulation is more prevalent among girls than boys (Rogers & Miller, 1981; Schofield & Francis, 1982), and hostility is manifested more overtly on the playground than in classrooms (Rogers & Miller, 1981). The list of boundary conditions under which contact is likely to increase interethnic acceptance has grown increasingly longer (Cook, 1983; Stephan & Stephan, 1983). On the other hand, and perhaps in response to the growing realization that they are needed, social scientists have begun to develop educational technologies that successfully promote increased interethnic acceptance (Aronson et al. 1978; Cohen & Roper, 1972; Cook, 1982; DeVries, Edwards, & Slavin, 1978; Johnson, 1975; Rogers, Hennigan, & Miller, 1981; Sharan & Sharan, 1976; Slavin, 1978; Serow & Solomon, 1979). Though these procedures differ in their details, the common thread among them is their use of structured cooperative interaction in small groups, whether in conjunction with the curriculum or on the playground. Meta-analyses of their use not only show consistent and substantial benefit to interethnic acceptance, but improved academic mastery when coordinated with academic curriculum materials (Johnson, Maruyama, Johnson, Nelson, & Skon, 1981; Johnson, Johnson, & Maruyama, 1983).

In summary, it is appropriate for schools to be concerned with children's development of effective and constructive interpersonal skills. The capacity for interethnic acceptance, respect, and trust is an important aspect of intrapersonal development and requires the existence of desegregated schools. Among the various goals that might be achieved by desegregated schooling, increased interethnic acceptance most directly addresses the central concern of Brown, namely, the stigmatization of blacks. Thus, we would argue that even if on the average the effect of desegregated schooling on academic achievement was shown to be zero, desegregated schooling is required if the issue of interracial acceptance is to be addressed.

Conclusion

Taken together, the desegregation studies that meet the NIE minimal criteria show some moderate academic benefit to black children when they attend desegregated schools. Although one reviewer finds a larger margin of benefit among studies with stronger designs (Crain & Mahard, 1978), most reviewers find that the magnitude of effect is smaller in studies with better research designs (e.g., Krol, 1978; St. John, 1975). Our calculation of the magnitude of these effects translates into the rather trivial increase of about twenty points on the typical SAT college entrance test which has a mean of 500 and a standard deviation of 100. Most studies of desegregation assess the effects of only a year of desegregated schooling. The likelihood, however, that twelve years of desegregated schooling will translate into an average gain of over 200 points (two standard deviations) on an SAT-type of test seems low. Our own longitudinal data from Riverside, California certainly argue against such a view (Gerard & Miller, 1975). On the other hand, the high likelihood that the same level of performance is evaluated more favorably by the external world if a black student attends a desegregated, as opposed to a segregated, school must be added to this picture. Given equal grade point averages or achievement test scores, the black student from a desegregated school is likely to be viewed as more capable and promising than his or her peer from a segregated school.

Our analyses of these and other data argue that the ratio of black and white students per se is probably not a direct causal factor in producing the small positive effect that is found. The fact that the magnitude of benefit is greater in studies conducted in the sixties than in those of the seventies supports this interpretation. The higher expectations and greater resources available in the earlier era should have generated increased morale and greater disruption of the status quo, thereby breaking the system effects that ordinarily depress the academic mastery of black children. Thus, we argue that whatever the academic effects found, they are due to teachers and schools and only attributable to changes in the percentages of black and white students to the extent that such changes concomitantly change teachers and schools.

Given the school effects that have been described in earlier sections, one could argue that such results essentially argue against the desegregation of schools. Implying as they do that lower class minority schools can be effective, education administrators should simply make the changes necessary to see that all such schools function effectively. Such a suggestion is not without merit, but is not easy to implement. When new teachers are brought into such schools to replace old ones, the normative structure exerts its influence on them, making them similar in outlook and practice to those they replaced. Such systems of norms can continue to show their effects, even when all the persons in the system have one by one been replaced (Jacobs & Campbell, 1961). As new persons come into the system they too adopt the old norms, and in turn, transmit them to still newer replacements.

For these reasons, a change in the black child's school environment is more easily achieved by moving him or her to a more middle class school, than by attempting to change the school currently being attended. Middle class schools, being more likely to be high-achieving schools, are less likely to have these debilitating systems of norms. Such a change can also give the minority student a sense of a fresh start.

In conclusion, the fact that school desegregation does not depress the academic performance of black children, but instead is moderately positive in its effect, (and as revealed in other reviews, does not adversely affect the academic performance of white children), means that if there are other compelling reasons to desegregate schools, consideration of academic achievement provides no deterrence. Because racially mixed schools are necessary if effective programs for increasing intergroup acceptance are to be applied, school desegregation should be encouraged.

Footnotes

1. Technically termed regression, this effect is due to the fact that the measuring instruments (tests) do not tell us each person's true score; there is a component of error in each score.

2. In determining whether or not the amount of variability across the studies exceeds that which would be expected on the basis of sampling error, it is necessary to weight the effect size estimates by sample size. Because smaller sample sizes are associated with increased imprecision of effect size estimate, it is important to assign such cases less weight so as not to overestimate the extent of variability that occurs over and above sampling error (i.e. to avoid overstating the case for the operation of moderator variables). It should be noted, however, that although taking a nonweighting approach normally will increase the likelihood of falsely concluding that moderators are present, this same procedure, which is the one that we do use for estimating the correlation between moderator variables and effect size, is conservative in this latter regard. The reason for this is that cases involving increased attenuation (via the imprecision of small samples) are given equal weight in determining the amount of correlation.

References

Armor, D.J., The evidence on desegregation and black achievement. Paper, commissioned by the National Institute of Education. Washington, D.C., 1983.

Aronson, E., Bridgeman, D., & Geffner, R., "Interdependent interactions and prosocial behavior." A Journal of Research and Development in Education, 1978, 12, 16-27.

Banks, W.C. "White preference in Blacks: A paradigm in search of a phenomenon." Psychological Bulletin, 1976, 83, 1179-1186.

Baughman, E.E. Black Americans: A psychological analysis. New York: Academic Press., 1971.

Beez, V.W. "Influence of biased psychological reports on teacher behavior and pupil performance." In M. Miles & W.W. Charters (eds.) Learning in Social Settings. Boston: Allyn and Bacon, 1970.

Black, C. "The lawfulness of the segregation decisions." Yale Law Journal, 1960, 69 421-430.

Bowles, S. & Gintis, H. Schooling in capitalist America. New York: Basic Books, 1976.

Brand, E.S., Ruiz, R.A., and Padilla, A.M. "Ethnic identification and preference." Psychological Bulletin, 1974, 81, 860-890.

Brink, W. & Harris, L. Black and White. New York: Simon and Schuster, 1969.

Brookover, W., Beady, C., Flood, P., Schweitzer, J., & Wisenbaker. School social systems and student achievement: Schools can make a difference. New York: Praeger, 1979.

Brophy, J.E. & Good, T.L. "Teachers' Communication of differential expectations for children's classroom performance." Journal of Educational Psychology, 1970, 61, 367-374.

Campbell, D.T. "Stereotypes and perception of group differences." American Psychologist, 1967, 22, 817-829.

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

Clark, N. "Development of consciousness of self and the emergence of racial identification in Negro pre-school children." Journal of Social Psychology, 1939, 10, 591-599.

Cohen, E. & Roper, S. "Modification of interracial interaction disability: An application of status characteristics theory." American Sociological Review, 1972, 37, 643-657.

Cook, S. "Motives in a conceptual analysis of attitude-related behavior." In W. Arnold and D. Levine (Eds.), Nebraska Symposium on Motivation. Lincoln, Nebraska: University of Nebraska Press, 1969.

Cook, S.W. "Social science and school desegregation: Did we mislead the Supreme Court?" Personality and Psychology Bulletin, 1979, 5, 420-437.

Cook, S.W. "Cooperative interaction in multi-ethnic contexts." In N. Miller & M.B. Brewer (Eds.), Groups in Contact: Desegregation. New York: Academic Press (in press).

Crain, R.L., & Mahard, R.E. "School racial composition and black college attendance and achievement test performance." Sociology of Education, 1978 51, 81-101.

Crain, R.L. & Weisman, C.S. Discrimination, personality, and achievement: A survey of northern Blacks. New York: Seminar Press, 1972.

Crawford v. Board of Education, L.A. No. 30485, 1976, 17 C3d 280-310.

Curtis, B. "The effect of segregation on the vocational aspirations of negro students." Dissertation Abstracts, 1968, 29, 772.

DeBord, L.W., Griffin, L.J., and Clark, M. "Race and sex influences in the schooling processes of rural and small town youth." Sociology of Education, 1977, 50, 85-102.

DeVries, D.L., Edwards, K.J., & Slavin, R.E. Biracial learning teams and race relations in the classroom: Four field experiments on teams-games-tournament. Report #230, Center for Social Organization of Schools, Johns Hopkins University, 1977.

Dworkin, R. "Social sciences and constitutional rights - the consequences of uncertainty." In R.C. Rist & R.J. Anson (Eds.), Education, Social Science, and the Judicial Process. New York: Teachers College Press, 1977.

Edmonds, R.R. Search for effective schools: The identification and analysis of the schools that are instructionally effective for poor children. Unpublished manuscript, Harvard University, 1976.

Entwisle, D.R. & Hayduk, L.A. Early schooling: cognitive and affective outcomes. Baltimore: Johns Hopkins University Press, 1982.

Epps, E.G. "Impact of school desegregation on aspirations, self-concepts and other aspects of personality." Law and Contemporary Problems, 1975, 39, 300-313.

Epps, E.G. "The impact of school desegregation on the self-evaluation and achievement orientation of minority children. Law and Contemporary Problems, 1979, 43, 57-76.

Falk, W.W. "School desegregation and the educational attainment process: Some results from rural Texas schools." Sociology of Education, 1978, 51, 282-288.

Fisher, J.E. An exploration of the effects of desegregation on the educational plans of Negro and White boys. Dissertation Abstracts, 1971, 31, 5548.

Fraser, R.W. Behavioral and attitudinal differences between teachers in desegregated classrooms. Unpublished doctoral dissertation, University of Southern California, 1981.

Gerard, H.B., & Miller, N. School desegregation. New York Plenum, 1975.

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." Review of Research in Education, 1977, 5, 351-379.

Good, T.L. "Which pupils do teachers call on?" Elementary School Journal, 1970, 70, 190-198.

Gottfredson, D.C. Personality and persistence in education: A longitudinal study. Paper presented at the Annual Meeting of the American Psychological Association, Montreal, Canada.

Hall, J.A. & Wiant, H.V. "Does school desegregation charge occupational goals of Negro males?" Journal of Vocational Behavior, 1973, 3, 175-179.

Hauser, R.M. "On 'A reconceptualization of school effects.'" Sociology of Education, 1978, 51, 68-72.

Heim, J.M. What research says about improving student performance. Albany: The University of the State of New York, the State Education Department, Bureau of School Programs Evaluation, March, 1973.

Heyns, B. Summer Learning. New York: Academic Press, 1978.

Hoffman, D. & Cohen, E.G. An exploratory study to determine the effects of generalized performance expectations upon activity and influence of students engaged in a group simulation game. Paper presented at the Annual Meetings of the American Educational Research Association, Chicago, 1972.

Hunter, J.E., Schmidt, F.L., & Jackson, G.B. Meta-analysis: cumulating research findings across studies. Beverly Hills: Sage, 1982.

Jacobs, R.C. & Campbell, D.T. "The perpetuation of an arbitary tradition through several generations of a laboratory microculture." Journal of Abnormal and Social Psychology, 1961, 62, 649-658.

Jencks, C., Smith, M., Acland, H., Bane, M.J., Cohen; D. Gintis, H., Heyns, B., Michelson, S. Inequality. New York: Basic books, 1972.

Johnson, D.W. "Cooperativeness and social perspective taking." Journal of Personality and Social Psychology, 1975, 31, 241-244.

Johnson, E., Gerard, H., & Miller, M. "Teacher influences in the segregated classroom." In H.B. Gerard & N. Miller (Eds.), School Desegregation, New York: Plenum Press, 1975.

Johnson, D.W., Johnson, R.T., and Maruyama, G. "Effects of cooperative learning: A meta-analysis." In N. Miller & M.D. Brewer (Eds.), Group in Contact: Desegregation. New York: Academic Press (in press).

Johnson, D.W., Maruyama, Johnson, R., Nelson, D. & Skon, L. "Effects of cooperative competitive, and individualistic goal structures on achievement: A meta-analysis." Psychological Bulletin, 1981, 89, 47-62.

Katz, I. Factors influencing Negro performance in the desegregated school. In M. Deutsch, I. Katz & A.R. Jensen (Eds.), Social Class, Race, and Psychological Development. New York: Holt, Rinehard, and Winston, 1968.

Kenny, D.A. "Cross lagged panel correlations: A test for spuriousness." Psychological Bulletin, 1975, 82, 887-903.

Kluger, R. Simple Justice. New York: Vintage, 1977.

Knapp, N. and Hammer, E. "Racial composition of southern schools and adolescent educational and occupational aspirations and expectations." Paper presented at annual meeting of the Association of Southern Agricultural Workers, Memphis, January 1971.

Krol, R.A. "A meta-analysis of comparative research on the effect of desegregation on academic achievement." Unpublished doctoral dissertation, Western Michigan University, 1978.

Kulik, J.A., Cohen, P.A., & Ebeling, B.J. "Effectiveness of programmed instruction in higher education." Educational Evaluation and Policy Analysis, 1980, 2, 51-64.

Kulik, J.A., Kulik, C.-L.C. & Cohen, P.A. "Effectiveness of computer based college teaching." Review of Educational Research, 1980, 50, 525-544.

Kulik, J.A., Kulik, C.-L.C. & Cohen, P.A. "Research an audio-tutorial instruction." Research in Higher Education, 1979a, 11, 321-341.

Lippitt, R. & Gold, M. "Classroom social structure as a mental health problem." Journal of Social Issues, 1959, 15, 40-49.

Maruyama, G. & Miller, N. "Reexamination of normative influence processes in desegregated classrooms." American Educational Research Journal, 1979, 16, 273-284.

Maruyama, G. & Miller, N. The relation between popularity and achievement: A longitudinal test of the lateral transmission of value hypothesis. Unpublished paper, 1983.

Miller, N. "Changing views about the effects of school desegregation: Brown then and now." In M.B. Brewer & B.E. Collins (Eds.), Scientific Inquiry and the Social Sciences. San Francisco: Jossey-Bass, 1981.

Mussen, P.H. "Differences between the TAT responses of Negro and White boys." Journal of Consulting Psychology, 1953, 17, 373-376.

Oakes, J. "Classroom social relationships: Exploring the Bowles and Gintis hypothesis." Sociology of Education, 1982, 55, 197-212.

Pettigrew, T. Social evaluation theory: Convergences and applications. In D. Levine (Ed.), Nebraska Symposium on Motivation. (Vol. 15), Lincoln, Nebraska: University of Nebraska Press, 1967.

Porter, J.D.R., and Washington, R.E. "Black identity and self-esteem: A review of studies of Black self-concept, 1968-1978." Annual Review of Sociology, 1979, 5, 53-74.

Proshansky, H., and Newton, P. "The nature and meaning of Negro self-identity." In M. Deutsch, I. Katz, and A.R. Jenson (Eds.) Social Class, Race, Psychological Development. New York: Holt, Rinehard and Winston, 1968.

Reniston, E.G. Levels of aspiration of Black students as a function of significant others in integrated and segregated schools. Dissertation Abstracts, 1973, 33, 7020-7021.

Rogers, M. & Miller, N. Quantitative and qualitative differences in peer selection among desegregated school children. Paper presented at the Annual Meeting of the American Psychological Association, Montreal, Canada, 1980.

Rogers, M. & Miller, N. The effect of school setting on cross racial interaction. Paper presented at the Annual Meeting of the American Psychological Association, Los Angeles, 1981.

Rogers, M., Miller, N., Hennigan, K. "Cooperative games as an intervention to promote cross-racial acceptance." American Educational Research Journal, 1981, 18, 513-516.

Rosenholtz, S. & Rosenholtz, S.H. "Classroom organization and the perception of ability." Sociology of Education, 1981, 54, 132-140.

Rubin, R.A., Maruyama, G., & Kingsbury, G.G. Self-esteem and educational achievement: A causal model analysis. Paper presented at the Annual Meeting of the American Psychological Association, New York, 1979.

Scheirer, M.A. & Kraut, R.E. "Increasing educational achievement via self-concept charge." Review of Educational Research, 1979, 49, 131-150.

Schofield, J. W. "Complementary and conflicting identities: Images and interaction in an interracial school." In S. Asher & J. Gottman (Eds.), The Development of Children's Friendship. New York: Cambridge University Press, 1980.

Serow, R.C. & Solomon, D. "Classroom climate and students intergroup behavior." Journal of Educational Psychology, 1979, 71, 669-676.

Sewell, W.H. Haller, A.O., & Portes, A. "The educational and early occupational attainment process." American Sociological Review, 1969, 34, 82-92.

Sharan, S. and Sharan, Y. Small-group teaching. Englewood Cliff: Educational Technology Publications, 1976.

Simpson, C. "Classroom structure and the organization of ability." Sociology of Education, 1981, 54, 120-132.

Singleton, L.C. & Asher, S.R. "Racial integration and children's peer preferences: An investigation of developmental and cohort differences." Child Development, 1979, 50, 936-941.

Slavin, R.E. "Student teams and comparison among equals: Effects on academic performance and student attitudes." Journal of Educational Psychology, 1978, 70, 532-538.

St. John, N.H. "The effect of segregation on the aspirations of Negro youth." Harvard Educational Review, 1966, 36, 284-294.

St. John N.H. School Desegregation outcomes for children. New York: Wiley, 1975.

Stephan, W.G. "School Desegregation: An evaluation of predictions made in Brown Vs. Board of Education." Psychological Bulletin, 1978, 85, 217-238.

Stephan, W.G. Blacks and Brown: The effects of school desegregation on black students. Paper commissioned by the National Institute of Education. Washington, D.C., 1983.

Stephan, W.G. & Stephan, C.W. "The role of ignorance in intergroup relations: Increasing knowledge of ethnic outgroups in multi-ethnic classrooms." In N. Miller & M.B. Brewer (Eds.), Groups in Contact: Desegregation. New York: Academic Press (in press).

Summers, A.A. & Wolfe, B.L. "Do schools make a difference?" American Economic Review, 1977, 65, 639-52.

Tammivaara, J.S. "The effects of task structure on beliefs about competence and participation in small groups." Sociology of Education, 1982, 55, 212-222.

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, 1979.

Walberg, H.J. "What makes schooling effective?" Contemporary Education Review, 1982, 1, 1-34.

Walberg, H.J. Desegregation and educational productivity. Paper, commissioned by the National Institute of Education, Washington, D.C., 1983.

Weinberg, M. "The relationship between school desegregation and academic achievement: A review of the research." Law and Contemporary Problems, 1975, 39, 240-270.

White, K. & Knight, J.H. "School desegregation, SES, SEX, and the aspirations of Southern Negro adolescents." Journal of Negro Education, 1973, 42, 71-78.

Wilson, A.B. "Residential segregation of social classes and aspirations of high school boys." American Sociological Review, 1959, 24, 836-845.

Wilson, W. "Rank order of discrimination and it's relevance to civil rights priorities." Journal of Personality and Social Psychology, 1970. 15, 188-224.

Wortman, P.M. School desegregation and black achievement: A meta-analysis. Paper, commissioned by the National Institute of Education, Washington, D.C., 1983.

Wylie, R.C. The Self-Concept: Theory and Research on Selected Topics (Vol. 2), Lincoln: University of Nebraska Press, 1979.

Yudof, M.G. "Nondiscrimination and beyond: The search for principle in supreme court desegregation decisions." In W.G. Stephan & J.R. Feagin (Eds.), School Desegregation: Past, Present, and Future. New York: Plenum, 1980.

Appendix A

1) Type of Study

   a) non empirical
   b) summary report

2) Location

   a) outside USA
   b) geographically non specific

3) Comparisons

   a) not a study of achievement of desegregated blacks
   b) multi-ethnic combined
   c) comparisons across ethnics only
   d) heterogeneous proportions minority in desegregated condition
   e) no control data
   f) no pre-desegregation data
   g) control measures not contemporaneous
   h) majority black in a segregated condition (unless the reviewer provides specific justification)
   i) varied exposure to desegregation (unless the reviewer provides a specific justification demonstrating that the vatiation in exposure time is not meaningful)

4) Study Desegregation

   a) cross-sectional survey
   b) sampling procedure unknown
   c) separate non-comparable samples at each observation

5) Measures

   a) unreliable and/or unstandardized instruments
   b) test content and/or instrument unknown
   c) dates of administration unknown
   d) different tests used in pretests and posttests
   e) test of IQ or verbal ability

6) Data Analysis

   a) no pretest means
   b) no posttest means, unless the author reported pretest scores and gains
   c) no data presented
   d) The following will be rejected dependent upon the amount of information available for the reviewer to estimate values
      1. no pretest standard deviations
      2. no posttest standard deviations
      3. no significance tests
      4. N's not discernable

It was decided that "excessive attrition" and "groups that are initially non-comparable" would not be used as criterion for rejection. In each case it was argued that the point at which the problem became an issue was extremely vague. It was felt that the project is better served by including studies exhibiting attrition and comparability problems and allowing individual reviewers to articulate these limitations. Using this criteria, 19 studies were studied which were deemed acceptable for inclusion in the project. These are:

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

Baker, Jerome. A study of integration in racially imbalanced urban public schools. Syracuse, New York: Syracuse University Youth Development Center, Final Report, May 1977.

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.

Carrigan, Patricia M. School desegregation via compulsory pupil transfer: Early effects on elementary school children. Ann Arbor, Michigan: Ann Arbor Public Schools., 1979.

Clark, El Nadel. Analysis of the difference between pre- and post-test 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.

Evans, Charles L. Short term desegregation effects: The academic achievement of bused students 1971-1972. Fort Worth, Texas: Fort Worth Independent School District, 1973. (ERIC, No. Ed 086 759)

Iwanicki, E.F., & Gable R.X. 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 doctcral 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.

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

Savage, L.W. *Arithmetic achievement of black students transferring from a segregated junior high school to an integrated junior 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, 165-182.

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

Syracuse City School District. Study of the effects of integration -- Washington Irving and Host Pupils. Hearing held in Rochester, New York, September 16-17, *U.S. Commission on Civil Rights.*

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.W. *Effects 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)

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

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

Blacks and Brown: The Effects of School Desegregation on Black Students*

Walter C. Stephan
New Mexico State University

The Effects of Segregation and Desegregation

It is important to put the question of the effects of desegregation on Black achievement in historical context. To do this I would like to quote from social scientists and other expert witnesses who testified in the Brown (1954) trial. It is clear from their testimony that the social scientists believed that segregation had a negative impact on Black achievement in at least three ways.

First, the fact that segregated Black schools were inferior to White schools in terms of the quality of the facilities and per pupil expenditures was thought to lead to low levels of achievement. Prior to Brown it was not uncommon for Southern states to allocate from 2 to 5 times as much money per pupil for White students as was allocated for Blacks (Ashmore, 1954; Thompson, 1975). Also, Black schools in the South had teachers who were less well trained and who were paid about half as much as teachers in White schools (Ashmore, 1954). Conditions in Black schools were often appalling. Consider the findings of Matthew Whitehead who testified about the schools in Clarendon County, South Carolina, during the Briggs vs. Elliot (1951) case.

> "The total value of the buildings, grounds, and furnishings of the two white schools, that accommodated 276 children was four times as high as the total for the three Negro schools that accommodated a total of 808 students. The white schools were constructed of brick and stucco; there was one teacher for each 28 children; at the colored schools, there was one teacher for each 47 children. At the white high school, there was only one class with an enrollment as high as 24; at the Scott's Branch high School for Negroes, classes ranged from 33 to 47. Besides the courses offered at both schools, the curriculum at the white high school included biology, typing, and bookkeeping; at the black high school, only agriculture and home economics were offered. There was no running water at one of the two outlying colored grade schools and no electricity at the other one. There were indoor flush toilets at both white schools but no flush toilets, indoors or outdoors, at any of the Negro schools--only outhouses, and not nearly enough of them." (Kluger, 1976, p. 332)

*The author wishes to thank Deanna Nielson for her assistance in preparing this article.

Second, it was thought that the "badge of inferiority" that segregation represented led Black students, and their teachers, to have low expectations regarding their capacities to achieve. These low expectations were believed to lead to low achievement. This argument can be traced through the testimony of several social scientists. David Krech said:

> "Legal segregation, because it is legal, because it is obvious to everyone, gives....environmental support for the belief that Negroes are in some way different from and inferior to white people." (Kluger, 1976, p. 362)

In another trial Horace English testified that:

> "If we din it into a person that he is incapable of learning, then he is less likely to be able to learn... There is a tendency for us to live up to-- or perhaps-- I should say down to social expectations and to learn what people say we can learn, and legal segregation definitely depresses the Negro's expectancy and is therefore prejudicial to his learning." (Kluger, 1976, p. 415)

Third, in addition to reducing expectancies, segregation was also thought to reduce the motivation to learn among Black students. Brewster Smith testified that:

> "Segregation is, in itself, under the social circumstances in which it occurs, a social and official insult and ... has widely ramifying consequences on the individual's motivation to learn." (Kluger, 1976, p.491)

And Louisa Holt argued that:

> "The fact that segregation is enforced... gives legal and official sanction to a policy which is inevitably interpreted both by white people and by Negroes as denoting the inferiority of the Negro group... A sense of inferiority must always affect one's motivation for learning since it affects the feeling one has for one's self as a person." (Kluger, 1976, p. 421)

In the original Brown (1951) decision this line of reasoning was sufficient to convince Judge Huxman that:

> "Segregation of white and colored children in public schools has a detrimental effect upon the colored children. The impact is greater when it has the sanction of the law; for the policy of separating the races is usually interpreted as denoting the inferiority of the Negro group. A sense of inferiority affects the motivation of a child to learn. Segregation with the sanction of law, therefore, has a tendency to retard the educational and mental development of Negro children and to deprive them of some of the benefits they would receive in a racially integrated school system." (Kluger, 1976, p. 424)

To summarize, it was because segregation was associated with inferior schools and led to low levels of expectancy and motivation in Black children that it was believed to cause low levels of achievement. At the time little or no data existed on the relative achievement levels of Blacks and Whites in segregated schools. Thus, the argument rested on reason, not fact.

Because the Brown trials were concerned with the negative effects to segregation, minimal consideration was given to the anticipated effects of desegregation. In fact, desegregation as a remedy for segregation was rarely mentioned (Kluger, 1976). The social scientists' arguments concerning the effects of segregation implied that removing the "badge of inferiority" represented by segregation would increase the academic expectancies and motivation of Blacks and that these increases, along with improved facilities and instruction, would lead to higher achievement.

Subsequent theorizing about the effects of segregation and desegregation on Black achievement has elaborated on these basic notions. For instance, the U.S. Commission on Civil Rights' study of Racial Isolation in the Public Schools suggested that:

> "Negro children suffer serious harm when their education takes place in public schools which are racially segregated, whatever the source of such segregation may be. Negro children who attend predominantly Negro schools do not achieve as well as other children, Negro and White. Their aspirations are more restricted than those of other children and they do not have much confidence that they can influence their own futures." (1967)

Jencks and his colleagues (Jencks, Smith, Aclard, Bane, Cohen, Bintis, Heyns and Michelson, 1972, pp. 97-98) offered four reasons why desegregation should improve Black achievement. First, they cited the anticipated positive effects of improvements in school and teacher quality. Second, they cited the knowledge that may be acquired from White peers who have been socialized into middle class White norms--the lateral transmission of values hypothesis (for evidence that this does not occur see Miller, 1981). Third, Jencks et al. suggested that teachers in desegregated schools may expect more from Blacks and this may lead Blacks to learn more. Fourth, desegregation may lead Blacks to expect that they have a better chance of making it in society which may motivate them to work harder and learn more (for a synthesis of many of these arguments see Linsenmeier and Wortman, 1978).

Achievement Tests

All of the studies to be considered in this analysis of the effects of desegregation on Black achievement employed standardized achievement tests. Any understanding of the results of these studies requires that some consideration be given to the nature of these tests. Achievement tests were developed to measure what students have learned. They consist of items that sample the general body of knowledge that schools are expected to teach. The items that are selected are those that discriminate best between students who have learned a great deal and those who have not. Items which sample knowledge that everyone learns are not included. This restricts the type of knowledge sampled to that which is not always learned or taught.

The tests usually take one to three hours to complete. During this period students at the junior high school level attempt to answer approximately 85 multiple choice questions per hour. The content areas covered most thoroughly (and the only ones reported in most desegregation studies) are math and verbal skills. Some tests deal with science and social studies, but use less extensive coverage for these topics. Thus, these tests examine only a very restricted domain of achievement. This domain, verbal and math skills, is clearly important, but so too are other domains of achievement that are not measured. Among these other domains are knowledge of our political, economic, and legal systems, and knowledge of the history of our society and other countries.

Scores on these tests correlate reasonably well from year to year and they correlate reasonably well with tests designed to measure aptitude and intelligence (Jencks et al., 1972, p. 60; Wallach, 1976). However, neither achievement tests nor those designed specifically for the purpose are especially good at predicting college grades or later success in life (Jencks et al., 1972, p. 57).

The test that has been most extensively scrutinized in this regard is the Scholastic Aptitude Test (SAT) developed by the Educational Testing Service. More than 2,000 studies have examined the ability of this test to predict future academic performance. The results indicate that the SAT correlates about .30 to .40 with first year college grades (Lord and Campos, cited in Linn, 1982). SAT scores do not correlate as well with overall college grades (Humphreys, 1968) nor do they predict whether or not students will finish college (Astin, 1970). Also, there is little relationship between SAT scores (or similar measures such as the GRE) and later success after college (Marston, 1971; McClelland, 1971). In sum, the SAT and most standardized achievement tests have high content and construct validity, but only low to moderate predictive validity.

We must be extremely cautious in interpreting the meaning of achievement scores. They reflect the amount of standard curriculum materials in the domain of math and verbal skills that students have learned. Thus, achievement scores may serve as an indicator of the quality of the math and verbal skills programs at the schools the students are attending, although the same material may be acquired in the home, from peers, or from the mass media. To the extent that desegregation has an effect on achievement scores, it may be caused by changes in the quality and amount of instruction in math and verbal skills, changes in the quality of the student body, or changes in the students' motivation to learn. The changes that do occur probably should not be interpreted as an indication that the students will subsequently be more or less successful in institutions of higher education or in economic terms.

I do not mean to imply that test scores are not important, but I believe they are often important for the wrong reasons. Scores on achievement tests are used as criteria to determine what tracks students will be assigned to and whether students will be admitted to college. They are also important because students and teachers perceive the scores as an indication of ability and individual worth. In this way, these tests may place inappropriate limits on the aspirations and self evaluations of low scoring students and they may lead teachers to have low expectations for low scoring students (For evidence that teachers have low expectations see Mercer, Iadicola, and Moore, 1980).

Because these tests measure what students have learned, anything that affects how much material they are taught or their capacity to assimilate what is presented will affect achievement test scores. Curriculum changes, differences in styles of presentation and testing, and disruptions that influence the capacity of teachers to teach or students' ability or desire to learn are likely to have a negative impact on what students learn. Because many of the studies reported in the literature cover only the initial phases of school desegregation they are very likely to be affected by these factors. ~~In particular,~~ the learning environment is apt to differ from the students' previous experiences, especially for minority students. Some of these differences may be beneficial in the long run such as more demanding teachers, more competitive classmates, and greater diversity in the student body, but these factors may initially have negative effects on achievement. Other factors such as tension and conflict between groups, negative comparisons with better prepared students who are often higher in social class, and dealing with teachers who have little experience teaching minority group students probably have a negative impact and continue to do do.

Although achievement tests are designed to measure what students have learned, scores on these tests are also affected by other factors. Most important among these other factors is the situation in which the tests are administered. In particular, high anxiety levels have a negative effect on performance, except for the very best students. It is possible that Black students taking these tests in desegregated schools experience more anxiety than Blacks in segregated schools. This is likely to be the case to the extent that achievement is emphasized in desegregated schools and the Black students feel academically inferior to or threatened by the White students.

Achievement tests are "speeded" which means that students have a time limit that is too short for many of them to finish all the items. This too may create anxiety; it also means that a premium is placed on motivation and attentiveness. Students who are not motivated to do well or who do not try hard will not score well on these tests. Lapses of attention that amount to 5 minutes during the testing hour will mean failing to answer about 7 questions (at the junior high level). This could affect the outcomes by more than 50 points (on tests that have a range of 200-800 with an average of 500). The tests are most likely to yield accurate results when the conditions of testing do not elicit high levels of anxiety and the students are motivated to do well and are attentive.

While these factors would be expected to influence measures of achievement both before and after desegregation, it would not be surprising to find that they had a more negative impact after desegregation.

The race of the examiner can also affect test performance. Blacks often perform better when the examiner is Black rather than White (e.g., Katz, Roberts, and Robinson, 1965). It is frequently the case that as students move from segregated to desegregated schools the race of the examiners changes from Black to White. Regrettably, we have no information on the degree to which such factors actually have affected the results of the studies we are reviewing, but they should lead us to be cautious about interpreting these studies.

The Studies in the N.I.E. Study Set

Anderson

This early study examines an unusual early desegregation plan in which students in the numerical minority in a given school could transfer to schools in which their group was in the majority. Thus, students could transfer from desegregated to segregated schools. The study was done in Nashville in 1963. It followed students from the 2nd to the 4th grade. The Metropolitan Achievement Tests were used to measure reading and math achievement. The sample size was adequate (N=34 in the desegregated group), but not large. It is possible that some of the students in the desegregated group were exposed to one year of desegregation prior to being pretested in the second grade. It appears from the report that this problem probably affected less than one-sixth of the students in this group.

Beker

Like most early studies, the desegregation that was examined in this study (1964) consisted of voluntary transfers. The study was done in a large Northern city. Two grade levels were included (grades 2 and 3). The sample sizes were very small and may yield unreliable results (N = 7 -25). The study is a Fall-to-Spring comparison of reading and math abilities done during the first year of desegregation (measured with the Stanford Achievement Test).

Bowman

This is one of the longer studies in the set. It runs from 1967 to 1970. A group of students was followed from grades 1 to 3 and another group from grades 3 to 5. The sample sizes were of moderate size (around 50 total at each grade level), but adequate. The students participated in the program voluntarily and it took place in a medium-sized Northern city (Syracuse). Different tests, the Iowa Tests of Basic Skills and New York State's Tests, were used to measure achievement at the pretest and the posttest levels which makes changes in test scores somewhat difficult to interpret.

Carrigan

I did not calculate effect sizes for this study because I believe the control group cannot be used to assess the effects of desegregation. In this stud; the control group was attending desegregated schools (50% Black). Since this control group had already received the "treatment" of desegregation, they provide a check primarily for maturation effects. Any changes in this group may be a consequence of ongoing exposure to desegregation, which means that the differences occurring in this group are not a proper control for the differences in the "desegregated" group. Also, the "desegregated" group actually started out in a somewhat desegregated school (80% Black), so this is not an optimal group to measure the effects of desegregation.

Clark

This is one of the small number of studies in the set that was done in the South. It is a study of a majority-to-minority transfer program that took place in 1969-1970. The sample size is adequate (N = 108 for desegregated group), but the duration of the study is brief, extending from Fall to Spring. This is the only study in the set that includes rural students. It covers only the sixth grade and provides both a test of reading and math (SCAT).

Evans

This study was done in Fort Worth during the 1971-1972 school year. The Iowa Tests of Basic Skills were given to 4th and 5th grade students in the Fall and Spring of that year. The court-ordered desegregation plan involved clustering elementary students and busing Black students (in grades 3-5) to achieve a degree of racial balance. The sample sizes were larger than in most of the other studies in this set (N = 179-393).

Iwanicki and Gable

I excluded this study because the "predesegregation" group had already been attending desegregated schools for a full academic year at the time of the "pretest." Thus, the predesegregation comparison is actually a cross-sectional comparison between a segregated control group and a group of students that has been desegregated for one year. This means that the measure of the effects of desegregation is a measure of the effects of the second year of desegregation. Since all of the other studies that I have included measured the first year of desegregation, including this study with the others may yield an inaccurate picture of the effects of desegregation. This would be particularly true if desegregation had a greater impact on achievement during the first year than during subsequent years.

Klein

This is a Fall-to-Spring examination of the effects of desegregation done in a small city (35,000) in the South. The students were in the tenth grade. The sample size was adequate (N = 38 in the desegregated group), but not large. The study was done in 1965. The desegregation plan was a voluntary one involving Black students who transferred from segregated Black schools to White schools. The tests used were the Math and English Cooperative Exams.

Laird and Weeks

This is an early study of the effects of desegregation (1964). It was done in a large Northern city (Philadelphia) over a 1½-year time span. Desegregation was brought about by overcrowding in a segregated Black school. Parents in this school could request to transfer their children to White schools so desegregation was voluntary. Students in grades 4-6 were tested on the district's own verbal and math tests. The sample size at each grade level is modest (22-39), but acceptable.

Rentsch

This study was done on a voluntary desegregation plan in Rochester, New York, and covers a 2-year time period. There were adequate sample sizes (N = 27 to 33) to calculate effects in grades 3-5. The students were tested on reading and math skills (apparently using a test developed by the District). The students who attended the desegregated schools had previously attended schools that were 90% minority. Attrition was fairly high in this group (56%). Although this study provided analyses of both matched and unmatched samples of segregated and desegregated students, I decided against using the analyses of the matched groups because the sample sizes were small (N = 9-13).

Savage

This study covered a longer time period than many of the others, 2 years, and it is one of the minority of studies that were conducted in the South (Richmond, Va.). Also, it is one of the relatively small number of studies examining senior high school students. The sample size is adequate (N = 42 in the desegregated group) to calculate reliable means for math and reading achievement on the Sequential Educational Progress Test. The study was conducted between 1969 and 1971 and examined a voluntary desegregation plan involving minority-to-majority transfers.

Sheehan and Marcus

This study was done in Dallas, Texas, and covers a 1½-year period. It involves court ordered busing and it was done recently (1976-1978). In these regards it is more representative of urban desegregation programs than most of the other studies in the set. The fourth grade students were measured with the Iowa Test of Basic Skills. The sample size is very large (nearly 2,000). One drawback is that the degree of desegregation varied considerably within the desegregated sample (from 5% to 65% Black).

Slone

This is a study of the second year of school desegregation. Desegregation occurred during the 1963-1964 school year. The first measure of achievement was gathered in April 1965 and the second in March 1966. The predesegregation school was multi-ethnic (90% minority, but only about 70% Black) and thus this study differs from the other studies of desegregation. Also, the "segregated" control group wes attending a school that was 40% White. Since the predesegregation levels of achievement cannot be determined, the effects of desegregation cannot be evaluated.

Smith

This is a long-term study, covering 3 school years. It was conducted in Tulsa, Oklahoma. The students were pretested in seventh grade and posttested in ninth grade. The sample size is larger than in most studies (N = 274). The Stanford Achievement Tests were used to measure math and verbal skills. The desegregated students were attending naturally integrated junior high schools. Unfortunately, no information was provided on the degree of segregation in Tulsa's elementary schools, but it is probably reasonable to assume a high level of segregation given that the study began in 1965.

Syracuse

This study of fourth grade students measured reading achievement (Stanford Achievement Test) in the Fall and Spring of the 1965-1966 school year. The number of students in the desegregated group was small, but adequate (N = 24). The type of desegregation program the students participated in is not specified in the report.

Thompson and Smidchens

This study of natural desegregation in the elementary schools of Ann Arbor was eliminated from the analyses because the students had been attending desegregated schools for 2 years before the predesegregation measures were obtained. Thus, this study lacks a true predesegregation measure. In addition, the "segregated" control group was 58% White.

Van Every

This study was done in Flint, Michigan, and involves desegregation produced by locating a low-cost housing project in a previously all White neighborhood. The study covers a 2-year period, following students from the fourth to the sixth grade. The sample size is somewhat small (desegregated group = 22). The study was completed in 1969. The Science Research Associates tests for reading and math were used. Research Associates tests for reading and math were used.

Walberg

This is a study of the Boston Metro Project in which urban Black students at all grade levels were voluntarily bused to suburban White schools. The performance of these Black students on the Metropolitan Achievement Tests for reading and math were compared to the performance of their siblings who remained in segregated Black schools. The study was conducted during 1968-1969. The sample sizes for the desegregated groups are moderate (N = 61-144), those for the segregated groups are smaller (N = 14-53), but still reasonably adequate.

Zdep

This is a study of a voluntary transfer plan in which urban Blacks could attend suburban schools. The students were very young (grade 2). The Metropolitan Readiness Test was used to measure reading and math ability in the Fall and during the Spring of the first year of desegregation. The study was done in 1968. The sample size was quite small and may not yield reliable results (N = 12 in the desegregated group). The report does not indicate where the study was done.

In summary, the desegregation in these studies was typically voluntary (66% of the cases), the cities it occurred in were generally medium to large, the region was more often the North than the South, the schools the students attended were more frequently elementary schools than secondary schools ($\bar{X}$ grade level = 5.5), Blacks were very much in the minority in most of these schools, and most of the studies were conducted prior to 1970 ($\bar{X}$ = 1968).

Effect Sizes

The principal measure of interest to be extracted from these studies is the size of the effects of desegregation on the verbal and math achievement of Black students. To calculate these effect sizes the formulas proposed by Glass (1977) were employed. In calculating these effect sizes I have taken into consideration the duration of the study.

All of the studies included in the study set employ quasi-experimental designs in which one group of students is tested before and after desegregation. The results for these students are compared to those of a group of students who remain in segregated schools and who are pretested and posttested at the same time as the desegregated group. The generic formula to obtain effect sizes in standard deviation units for this design is to calculate the difference between the desegregated and segregated groups at the pretest and divide this score by the standard deviation for the segregated group.

$$1) \quad \frac{\bar{X}_1 - \bar{X}_2}{S.D._2} = \text{pretest difference}$$

This score indicates the degree of pretest equality between the two groups. A similar score is then obtained for the posttest scores.

$$2) \quad \frac{\bar{X}_1 - \bar{X}_2}{S.D._2} = \text{posttest difference}$$

To derive an overall effect size the pretest difference (1) is subtracted from the posttest difference (2). This formula yields an index of the magnitude of the effects of desegregation in units that can be compared across studies.

The use of the standard deviation of the control group (the segregated group in this case) to calculate effect sizes was proposed by Glass (1977). It would be possible to use in place of this standard deviation a pooled standard deviation comprised of the average of the standard deviations of the experimental and control groups on the assumption that this would yield a more stable estimate of the standard deviation. This more complex approach would be justified if the standard deviations of the experimental and control groups differed substantially. This appears not to have been the case in the present set of studies. In no instance (on the pretest or the posttest) were there significant differences between the mean standard deviations of the segregated and the desegregated groups. Thus, it seemed reasonable to employ the simpler formula advocated by Glass.

In this set of studies the duration of desegregation varies considerably. In order to obtain an index of the effects of desegregation during the first year of desegregation I first divided the effect size (E) by the duration (D) of the study to yield an effect size per month. In calculating the duration of the study I used the total number of months the study covered and subtracted 3 months for each summer vacation period that was included. Thus, the duration measure reflects only the number of months the students actually spent in school. Next, I multiplied effect size per month by 8 to obtain an index of the effect size per year.

$$\frac{E}{D} \times 8 = \text{effect size per year}$$

The primary value of this index of effect size is that it avoids including together in subsequent analyses studies that vary in duration from 4 to 36 months. These scores were calculated separately for verbal and math achievement to determine if desegregation had differential effects on the two basic areas covered by achievement tests. Since some studies included more than one grade level, I calculated effect sizes for each grade and for each study as a whole so that comparisons could be made using grade or study as the unit of analysis. The effect sizes for grade are presented in Table 1.

Using this procedure for calculating effect size per year assumes that desegregation has linear effects over time, at least over the first 3 years of desegregation. This is the easiest and, I believe, the most defensible assumption to make in dealing with the effects of desegregation over the first few years of desegregation. There are other plausible relationships, however. For instance, it might be predicted that if desegregation had positive effects, most of the benefits would accrue to the students during the initial year or two of desegregation after which little additional benefit would be derived. Alternatively, desegregation might be expected to have negative effects on achievement initially because of the negative conditions under which it so frequently occurs. Later, after adjustments have been made, desegregation might be predicted to have beneficial effects. The curvilinear nature of these predictions makes them difficult to apply to the present studies. In this set of studies the assumption of linearity appears to be reasonable in the case of math where the correlation between the duration of the study and the effect size was marginally

Table 1

Effect Sizes

Study	Grade	Effect for Reading	Effect for Math
Anderson	4	.42	.24
Beker	2	.19	-.31
	3	.06	-.17
Bowman	3	.25	.21
	5	.00	-.10
Carrigan[a]	1	-.41	
	2	-.02	
	3	.30	
	4	-.13	
	5	.33	
	6	-.31	
Clark	6	.08	-.24
Evans	3	.02	.03
	4	.02	.03
	5	.02	.03
Iwanicki &	3	.00	
Gable[a]	5	.00	
Klein	10	.23	.33
Laird & Weeks	4	.22	.18
	5	.31	.18
	6	-.14	-.17
Rentsch	5	.07	.02
	6	.26	-.08
	7	.33	-.10
Savage	12	.06	-.04
Sheehan & Marcus	4	-.07	-.08
Slone[a]	5	.19	.22
Smith	9	-.01	.02
Syracuse	4	.55	
Thompson & Smidchens[a]	5	-.15	.04
Van Every	6	-.12	.14
Walberg	4	.15	.07
	6	.05	-.53
	8	.17	.24
	11	-.15	.14
Zdep	2	.66	-.15

[a] Excluded from analyses

significant (r = .48, p $<$ .10). In the case of reading, the correlation was not significant (r = - .17, ns). Krol's (1978) study of effect sizes for achievement is consistent with the assumption that the effects over time are linear.

The manner in which the results of these studies are presented is highly variable. In some studies the means and standard deviations necessary to calualate effect sizes using the generic formula are reported, but in others the effect sizes must be calculated using F tests, T tests, analyses of difference scores or analyses of covariance. Strictly speaking none of the latter calculations is precisely comparable to the generic formula, since the derived standard deviations are calculated from the overall variance. In cases where only covariance analyses are available, the effect sizes are almost certainly overestimated. This means that the average effect sizes across this group of studies are only approximate estimates.

Using studies as the unit of analysis, the average effect size for the first year of desegregation (8 months) was .17 verbal achievement, while the average effect size for math achievement was .00 (Table 2). Using the effect size for each grade as the unit of analysis, the effects are .15 for reading and .00 for math. Dropping the four studies from the sample set that I excluded has little effect on the results. Using studies as the unit of analysis, the mean effect size for verbal achievement including all the studies in the set is .14 and for math it is .04. These results appear to indicate that verbal achievement improves somewhat, but math achievement shows little effect as a result of desegregation. The difference between the X for reading achievement and the $\bar{X}$ for math achievement is marginally significant (t = 1.96, p $<$ .08, Table 4).

One way to convey the magnitude of these effect sizes is to consider what it would mean in terms of a test, such as the SAT or the GRE, that has a $\bar{X}$ of 500 and a standard deviation of 100. The effect for verbal achievement would translate into a 17 point increase as a consequence of the first year of desegregation. The math effect would translate into no improvement. Another more approximate way of thinking about these figures would be to consider what the effects of desegregation are on the average percentile ranking of Black students on a standardized test. If desegregation improved verbal achievement .17 standard deviation units, this would raise the average percentile rank of Blacks about 5 percentage points during the first year of desegregation. For math there would be no changes in percentile rank due to desegregation.

Why would desegregation affect the reading achievement of Blacks and not their achievements in math? One possibility is that reading achievement may be improved by direct exposure to the language usage and vocabulary of White students and teachers. Learning middle-class vocabulary and syntax may aid test performance. Such an improvement would not be due to any changes in the quality of teaching, or changes in expectancies or achievement motivation, but simply to being able to understand the tests and the content of the questions better. Similar improvements would not be expected for math because there is no parallel to this type of indirectly learned information in the case of math. Here no improvement

Table 2

Means for
Uncorrected Effect Size and
Effect Size Corrected for Duration of Study

Using Classes as the Unit of Analysis

	Reading	Math
Uncorrected		
$\bar{X}$	.24	.04
S.D.	.39	.34
Corrected		
$\bar{X}$	.15	.00
S.D.	.22	.20

Using Studies as the Unit of Analysis

	Reading	Math
Uncorrected		
$\bar{X}$	.24	.06
S.D.	.35	.25
Corrected		
$\bar{X}$	.17	.00
S.D.	.22	.16

Table 4

Reading vs. Math*

	Reading Effect Size	Math Effect Size	t	df	p
Uncorrected (Studies)	.21	.06	1.33	13	ns
Corrected (Studies)	.14	.00	1.96	13	.08
Uncorrected (Classes)	.21	.03	2.27	24	.04
Corrected (Classes)	.12	.00	2.52	24	.02

*The Syracuse study is excluded from this analysis because it did not include math achievement.

would be expected unless there were changes in the quality of instruction or the students' expectancies or achievement motivation increased.

In this set of studies. the magnitude of the effect sizes is unrelated to the region in which the studies were done, the size of the cities in which the studies were done, and the size of the samples (Table 3). There is a marginally significant negative correlation between the grade the students were in when they were desegregated and the size of the effect for reading achievement ($r = -.33$, $p < .10$). The relationship between grade and effect size is not significant for math ($r = .22$, ns). For reading this suggests that younger students benefited more than older students from desegregation. One explanation for this relationship is that exposure to White students (and in some cases, White teachers) may benefit students who have had little previous direct or vicarious contact with Whites. This benefit probably consists of exposure to the type of vocabulary that achievement tests measure. Older students who have had more direct and vicarious contact with Whites may benefit less from exposure to Whites in desegregated schools because they have had more exposure to White middle-class language usage and vocabulary.

The correlation between the year the study was done and the size of the effect for reading is also marginally significant ($r = -.49$, $p < .10$, using studies as the unit of analysis). The correlation between the year the study was done and math achievement is not significant ($r = -.32$). It is not clear why this effect exists for reading. One possibility is that the early studies tended to be of voluntary desegregation where only select students participated. These desegregation programs may have made special efforts to help the incoming students and these students were probably highly motivated to succeed. In contrast, students in mandatory desegregation programs and later voluntary programs may have received less special treatment and may not have been as motivated to learn. However, the effects of special treatment would be expected to affect both reading and math, and there was no relationship for math, although the direction of the correlation is the same.

It also appears that the effect size for reading was larger in school districts where the desegregation was voluntary rather than mandatory ($\bar{X} = .21$ voluntary, $\bar{X} = -.03$ mandatory). While this difference is statistically significant ($t = 3.15$, $p < .05$, using studies as the unit of analysis and the corrected effect sizes as the dependent measure), the number of districts in which desegregation was mandatory is so small ($n = 2$) that these results may not be reliable. The effect for math was not significant ($t = .25$, ns). The most likely explanation for these effects is that the students who participated in desegregation voluntarily were more motivated to get to know other students. This informal contact would have enabled them to acquire verbal skills that could have affected their test performances, but it would not have enabled them to acquire math skills that affect test performance.

I would like to argue that none of the relationships regarding effect size, grade, year, city size, region, or type of desegregation should be regarded as conclusive because the effect sizes themselves are

Table 3

Correlations of Corrected Achievement Scores with Grade, Year, City Size and Sample Size

	Reading		Math	
	By Classes	By Studies	By Classes	By Studies
Grade	-.33**		.22	
Year	-.42*	-.49**	-.10	-.32
City Size	-.18	-.21	-.18	.19
Sample Size	-.28	-.38	-.05	-.17

*p <.05
**p <.10

unreliable. Even the overall effect sizes that were obtained may not be meaningful. Given the variability in the effect sizes in these studies, the confidence limits are rather broad. The 95% confidence limits (the range within which the true population $\bar{X}$ is likely to fall, with only a 5% probability of being mistaken) for verbal achievement are .04 to .30, and the 95% confidence limits for math achievement are -.09 to +.09. Thus, in the case of reading achievement we can be reasonably confident that desegregation has an effect, although it may be very small indeed. In the case of math, desegregation appears to have no effects.

There are other reasons why the average effect sizes should be regarded with more than a little caution. In those studies involving multiple grades it is possible to examine fluctuations in the standard deviations of the students' achievement scores. For instance, in Rentsch's study the range in standard deviations for the verbal scores is 9.57 to 13.14, and the range for math scores is 6.52 to 13.37. Obviously, when these standard deviations are used to calculate effect sizes (using the generic formula) the magnitude of the effect size will depend on the standard deviation that is used. If the standard deviations are unstable, then the effect sizes will be correspondingly unstable. The lack of stability in standard deviations tends to be a problem with the studies where the sample sizes are small.

One reason that the studies with small samples have variable standard deviations consists of sampling problems (e.g., non-random sampling). Fluctuations in standard deviations within studies may also occur as a consequence of variable conditions during test administration. Anyone who has given tests to elementary students is aware of how difficult it is to maintain standardized procedures. Large sample sizes compensate somewhat for this variability in testing conditions, but most of the studies reviewed here did not use large samples.

Even if the standard deviations were stable, the small sample sizes of many of these studies would result in means that may not be accurate. In order to be accurate to within .5 standard deviation units of the true population $\bar{X}$, a sample size of 15 is required. To be accurate to within .1 standard deviation units, requires a sample of 384. Thus, the mean values reported in the studies with small sample sizes are not likely to be measured accurately enough to provide reliable effect sizes. If there were a sufficient number of these samples, the errors of measurement would cancel each other out, but the number of samples is not large enough in this set of studies to lead to confidence in the summary figures concerning effect sizes. Also, the substantial variability in effect sizes suggests that the mean effect size may be distorted by extreme scores and indeed the effect size for verbal achievement is lowered to .13 if the median is used as a measure of central tendency rather than the mean. If the effect sizes were corrected for the unreliability of the achievement tests this would also lower the estimate of the verbal achievement effect size.

Another reason that the average effect sizes should be viewed with caution concerns methodological problems with the studies. While these studies were chosen because they are the best ones available, they are not without their defects. The list of potential defects is a long one. Threats to internal validity include those already mentioned, small sample sizes, non-random samples, and fluctuations in standard deviations (suggesting unreliability of measures). In addition, the quality of the measures of achievement varies (some use measures developed within the district, others use tests standardized on White populations), attrition varies considerably across studies and threatens the validity of studies where it is high, and the segregated control groups are often of uncertain comparability to the desegregated groups.

Threats to external validity are comprised primarily of concerns with the non-representativeness of these samples of Black students and of this group of studies. Only students who are in desegregated schools at the end of the study are included in the posttest and often in the pretest $\bar{X}$'s. Usually students whc stay in the program are not compared to those who drop out to determine if they are different. Thus, we cannot be confident that the samples of desegregated students in these studies are representative of Black students generally. Also, the studies are mostly of voluntary desegregation in medium to large northern cities. The degree to which it is appropriate to generalize these results to maudatory desegregation in other regions of the country or to small cities and rural areas is unclear.

Glass (1977) in discussing meta-analyses as a research method suggests that "Respect for parsimony and good sense demands an acceptance of the notion that imperfect studies can converge on a true conclusion" (p.356). His argument relies on an example in which a set of studies are "similar in that they show a superiority of the experimental over the control group" (p.356). However, this argument may not apply as forcefully to a set of studies, such as those on the effects of desegregation on Black achievement, in which the results are variable rather than similar. Under these circumstances, the variability in results may be interpreted in terms of methodological problems as parsimoniously as in terms of more substantive causes.

A Basic Problem in Evaluating Desegregation

Perhaps the most fundamental oversight of the social scientists involved in the *Brown* trial was in not giving due consideration to the manner in which segregation would be eliminated. They were not alone in this oversight, even the lawyers for the NAACP did not consider this problem in detail until after the first *Brown* decisiou in 1954. The Justices of the Supreme Court were vague in their recommendations saying in the second *Brown* decision in 1955 only that segregation should be ended with "all deliberate speed" (Kluger, 1976, pp. 714-747). When desegregation began to be implemented 10 years after *Brown*, the forms it took were as varied as the communities in which it took place. I believe it is this complexity more than any other factor that accounts for the diverse results that have been observed in studies of the effects of desegregation on achievement. The diversity of desegregation programs is so great as to render the word without a precise meaning.

Let me be specific about this complexity, although it is familiar to anyone who has studied the problem. Each community starts with its own unique history of relations between the races including when Blacks and Whites settled there, the origins of members of these groups, the social class structure of the groups, the degree of residential segregation and so on. The communities vary along such potentially important dimensions as size, region of the country, ratio of majority-to-minority group members, presence of suburbs and private schools to which Whites may flee, and funding for public schools. The desegregation programs implemented in these communities have their own unique history of litigation and decision making by school boards and other public officials. The programs themselves vary in the techniques used to create desegregation, some programs are voluntary but most are not, the plans may involve voluntary cross-district busing, pairing, the use of magnet schools, the closing of some (usually Black) schools, and the mandated busing of students (usually Black students). The desegregation of teachers may or may not accompany the desegregation of students and the amount of preparation teachers are given for desegregation is variable. Additional curricular changes may occur at the same time as desegregation, the age of the students included in desegregation plans varies, the speed with which a plan is implemented varies, community opposition varies as does the amount of White flight, the ratio of majority-to-minority students differs from community to community as do the social class backgrounds of the students and the quality of their predesegregation educational experiences. As long as this list seems, it is surely incomplete. What these differences mean is that comparing the effects of desegregation across communities is extraordinarily difficult. It is possible to use quantitative measures to examine the effects of some of the factors in this list, but the majority are more difficult to study and compare.

The Effects of Desegregation on Self Esteem and Race Relations

The social scientists who participated in the Brown trials believed that segregation has negative effects on the self esteem of Black students and on relations between the races, as well as having negative effects on achievement. One of the clearest presentations of their views comes from the statement that 35 social scientists filed as an Amicus Curiae brief in the Brown trial.

> " Segregation, prejudices and discriminations, and their social concomitants potentially damage the personality of all children ... Minority group children learn the inferior status to which they are assigned ... they often react with feelings of inferiority and a sense of personal humiliation ... Under these conditions, the minority group child is thrown into a conflict with regard to his feelings about himself and his group. He wonders whether his group and he himself are worthy of no more respect than they receive. This conflict and confusion leads to self-hatred ...
>
> Some children, usually of the lower socio-economic classes, may react by overt aggressions and hostility directed toward their own group or members of the dominant group." (Allport et al., pp. 429-430)

The social science brief and testimony in the individual trials leading up to Brown indicate that it was anticipated that ending segregation would remove the stigma of inferiority that was forced on Black children.

Self esteem. The cffects of desegregation on self esteem appear to be less favorable than the effects of desegregation on achievement. In my earlier review (Stephan, 1978), I found that desegregation led to decreases in the self esteem of Black students in 5 of 20 studies and that there were no studies indicating that desegregation increased Black self esteem. As was true for the studies of the effects of desegregation on achievement, the majority of these studies have been concerned with the effects of desegregation over a period of 1 year or less. One study that examined the effects over a longer period of time found ~~that while Black self esteem initially dr~~opped, it rebounded to predesegregation levels during the second year (Gerard and Miller, 1975). Subsequent studies of Black self esteem, including my own (Stephan and Rosenfield, 1978), have not changed this picture much. My conclusions regarding the effects of desegregation on the self esteem of Black students are consistent with those of other investigators (e.g., Banks, 1976; Epps, 1975; Gordon, 1977; Shuey, 1966).

It appears that the social scientists who participated in Brown used an invalid assumption as a basis for their argument that desegregation would increase the self esteem of Black students. Undoubtedly segregation stigmatizes Black students, but this stigma is not reflected in the self esteem of Black students. Studies of segregated Blacks and Whites show that Black students have self esteem levels that are similar to or higher than White students in more cases than they have lower self esteem (see Porter and Washington, 1979, and Stephan and Rosenfield, 1979, for reviews). These studies have employed questionnaire measures of self esteem rather than indirect measures such as the doll tests upon which the social scientists' statements in Brown were based. The indirect measures may have been tapping attitudes toward Blacks and Whites as ethnic groups. There is considerable evidence indicating that young Black children have less favorable attitudes toward Blacks than toward Whites (Williams and Morland, 1976).

If segregated Black students do not have low self esteem, there is little reason to expect that desegregation would increase self esteem. In fact, their are several compelling reasons why decreases in self esteem might be expected. For instance, social comparison with White students who are academically better prepared than Blacks could lead Blacks to evaluate themselves negatively. Likewise, the loss of status and power that occurs when Blacks represent a minority of the student body in desegregated schools could also lower the self esteem of Black students. In addition, negative evaluations by ethnocentric White students could adversely affect the self esteem of Blacks.

Attitudes. The social scientists in their brief were also hopeful that contact within the schools would improve intergroup relations.

> "Under certain circumstances desegregation ... has been observed to lead to the emergence of more favorable

> attitudes and friendlier relations between races. ...There is less likelihood of unfriendly relations when change is simultaneously introduced into all units of a social institution ... and when there is consistent and firm enforcement of the new policy by those in authority. ...These conditions can generally be satisfied in ... public schools." (pp. 437-438)

The social scientists appreciated the fact that contact alone would not be sufficient to improve intergroup relations. Their statement notes several preconditions for favorable change; equal status between the groups, and firm, thorough implementation of desegregation. It is likely that they were aware of other relevant factors such as those mentioned by Williams (1947) a half dozen years before the social science statement was drafted:

> "Lessened hostility will result from arranging intergroup collaboration, on the basis of personal association of individuals as functional equals on a common task jointly accepted as worthwhile." (Williams, 1947)

The data on the initial effects of desegregation on race relations suggest that the social scientists' caution was well founded. In an earlier review of the data, I found that desegregation increased Black prejudice toward Whites in almost as many cases as it decreased prejudice (Stephan, 1978). The results for Whites were somewhat more negative. Recent studies; including my own, which also indicated that desegregation does not improve race relations (Stephan and Rosenfield, 1978), have not led me to revise these conclusions (e.g., Bullock, 1976; Campbell, 1977; Patchen, 1982; Sheehan, 1980). The quality of these studies is not as high as the better achievement studies, and there is such a small number of them that these conclusions can only be regarded as tentative. My conclusions are, however, generally consistent with those of other investigators (Armor, 1972; Epps, 1975; St. John, 1975; Schofield, 1978; Weinberg, 1970).

In the year since <u>Brown</u> the contact hypothesis has been elaborated and refined. These elaborations are helpful in understanding why desegregation often has not has a positive effect on race relations. Here are my own most recent statements concerning the conditions under which contact improves intergroup relations.

1. Cooperation within groups should be maximized and competition between groups should be minimized.

2. Members of ingroup and outgroup should be of equal status both within and outside of the contact situation.

3. Similarity of group members on non-status dimensions appears to be desirable (beliefs, values, etc.).

4. Differences in competence should be avoided.

5. The outcomes should be positive.

6. There should be strong normative and institutional support for the contact.

7. The intergroup contact should have the potential to extend beyond the immediate situation.

8. Individuation of group members should be promoted.

9. Non-superficial contact (e.g., mutual disclosure of information) should be encouraged.

10. The contact should be voluntary.

11. Positive effects are likely to correlate with the duration of the contact.

12. The contact should occur in a variety of contexts with a variety of ingroup and outgroup members.

13. There should be equal numbers of ingroup and outgroup members. (Stephan, 1983)

Desegregation rarely occurs under conditions that would lead to improvements in race relations. Instead, desegregation often occurs after there has been considerable community opposition from parents, administrators, school boards, and teachers. Thus, institutional and normative support for the contact is frequently low; the atmosphere tends to be competitive rather than emphasizing cooperation in pursuit of common goals; the statuses of Blacks and Whites often are unequal both outside the school (due to social class) and within the school (due to unbalanced ratios of Blacks and Whites); the Black students are often not as well prepared academically as the Whites, so stereotype-confirming differences in academic competencies frequently occur; busing often limits out-of-school contact and the within-school contact that does occur is more likely to be negative or neutral than positive, and in most cases it will be superficial. Also, the contact is involuntary in the case of court-ordered desegregation.

Recent research on the use of cooperative interethnic groups in desegregated schools indicates that when the conditions specified above are met, intergroup relations and self esteem improve without any costs in terms of lowered achievement (e.g., Aronson, Stephan, Sikes, Blaney and Snapp, 1978; Cohen, 1980; Cooper, Johnson, Johnson and Wilderson, 1980; De Vries, Edwards and Slavin, 1978; Weigel, Wiser and Cook, 1975). Other intergroup relations techniques involving multiethnic curricula, discussions of race issues, and explicitly providing information about the cultures of different groups have also been found to improve intergroup relations in the majority of cases (see Stephan, 1983; and Stephan and Stephan, 1983, for reviews). What these studies demonstrate is that while simply mixing students of different groups in desegregated schools does not improve race relations, intergroup relations can be improved in desegregated schools by introducing special programs designed to achieve this goal.

Future Directions for Research in Desegregation

I would like to see research into techniques to improve achievement, race relations, and self esteem continue. In addition, there are several other areas where I thank research should also be done. One of the major problems with nearly all desegregation research is that it only covers the effects of the first year of desegregation, or at most the first two or three years of desegregation. There are almost no studies of the long-term effects of desegregation. We need to know not only what the long-term educational effects of desegregation are, but we also need to know what the non-educational effects are. And we need to know the effects not only for Whites and Blacks, but also for other ethnic groups as well. Does school desegregation reduce segregation in other realms, such as housing; do minority students who have attended desegregated schools get better jobs and do they get promoted at a faster rate than students who attended segregated schools; and is subsequent political participation increased as a result of attending desegregated schools?

Also, we need to know more about the effects of desegregation on the communities that have undergone it. For instance, how do people in communities with well-established desegregation programs feel about desegregation now; are people who have attended desegregated schools more willing to send their children to desegregated schools than people who attended segregated schools; and what differences are there in the race relations of communities with well-established desegregation programs compared to other communities?

A third set of questions concerns the factors associated with successful desegregation programs. When desegregation goes well, why does it work? One can imagine a wide variety of factors that could be relevant, some having to do with the community in which it takes place, others having to do with the way administrators and teachers respond to desegregation, and still others with the composition of the student body. The fact is that we know precious little about what differentiates successful from unsuccessful desegregation programs.

Desegregation in Perspective

It would be impossible to present a comprehensive evaluation of the effects of desegregation in this short article. Instead, I have attempted to confine myself to some of the effects of desegregation on students. However, the larger context in which desegregation occurs is of immense importance to an understanding of the meaning of desegregation.

In order to put desegregation in perspective, we must consider the role that it has played in influencing relations between the races in our society. Since 1954, vast changes in race relations have occurred; many overt forms of discrimination have been eliminated, levels of prejudice have decreased, most minority groups have made economic advances, political participation by minority group members has increased dramatically, and more minority group members are attending college.

School desegregation has played a role in these economic, political, and social changes, but it is a role that is not well understood and is little studied. Any analysis that abstracts school desegregation from its social context is necessarily incomplete. Unfortunately, we are not now in a position to perform such an analysis. Given the difficulty of answering even a limited question like the effects of desegregation on Black achievement, it doesn't seem likely to me that we will be in a position to do an adequate comprehensive evaluation of desegregation anytime in the near future.

As we acquire more information on the outcomes of desegregation, we will be in a better position to base policy decisions on data. However, for the present, it seems to me that we will have to continue to make major policy decisions about desegregation on the basis of competing values. Some of these values concern the goals of public education, in particular the degree to which the schools should concern themselves with intergroup relations and the preparation of students to participate in a pluralistic society. Other decisions that we will continue to have to make pit the importance of creating equal educational opportunities against freedom of choice and freedom of association. Perhaps most importantly we will have to decide whether we value the elimination of segregation enough to continue the 50-year battle against it. Social science may be of less value in making these choices decision than in making choices about the best ways of implementing these decisions.

References

Allpcrt, F.H., et al. "The effects of segregation and the consequences of desegregation: A social science statement." Minnesota Law Review, 1953, 37, 429-440.

Armor, D. J. "The evidence on busing." The Public Interest, 1972, 28, 90-126.

Aronson, E., Stephan, C., Sikes, J., Blaney, N., and Snapp, M. The Jigsaw Classroom, Beverly Hills, Calif.: Sage Publications, 1978.

Ashmore, H.S. The Negro and The School. New York: Van Rees Press, 1954.

Astin, A. W. "Racial Considerations in Admissions," in David C. Nichols and Olive Mills, eds., The Campus and the Racial Crisis, Washington, D. C.: American Council on Education, 1970, p. 87.

Banks, W. C. "White preference in blacks: A paradigm in search of a phenomenonx." Psychological Bulletin, 1976, 83, 1179-1186.

Campbell, B. C. "The impact of school desegregation." Youth and Society, 1977, 9, 79-111.

Cohen. E. G. A multi-ability approach to integrated classrooms. Paper presented at the American Psychological Association, Montreal, Canada, Septembet, 1980.

Coleman, J., et al. "Equality of educational opportunity," Washington, D. C.: Department of Health, Education and Welfare, 1966.

Cooper, L., Johnson, D., Johnson, R., and Wilderson, F. "The effects of cooperative, competitive and individualistic experiences on interpersonal attraction among heterogeneous peers." Journal of Social Psychology, 1980, 111, 243-252.

De Lone, R. H. Small Futures. New York/London: Harcourt Brace Jovanovich, 1979.

De Vries, D. L., Edwards, K. J., and Slavin, R. E. "Biracial learning teams and race relations in the classroom: Four field experiments using teams--fames-- tournaments." Journal of Educational Psychology, 1978, 70, 356-362.

Epps, E. G. "The impact of school desegregation on aspirations, self-concepts and other aspects of personality." Law and Comtemporary Problems, 1975, 39, 300-313.

Ford, S. F., Campos, S. Summary of Validity Data from the Admissions Testing Program cited in R. L. Linn, "Admissions testing on trial," American Psychologist, 1982, 37, 279-291.

Gerard, H. B., and Miller, N. School Desegregation. New York/London: Plenum Press.

Gordon, V. V. The Self-concept of Black Americans. Washington, D. C.: University Press of America, 1977.

Hoyt, D. P. The Relationship Between College Grades and Adult Achievement: A review of the Literature, ACT Research Report, Iowa City: ACT Research and Development Division, 1965.

Humphreys, L. G. "The fleeting nature of the prediction of college academic success." Journal of Educational Psychology, 1968, 59, 375-380.

Jencks, C., Smith, H. A., Bane, M. J., Cohen, D., Gintis, H., Heyns, B., Michelson, S. Inequality. New York/London: Basic Books, 1972.

Katz, I., Roberts, S.O., and Robinson, J. M. "Effects of task difficulty, race of administrator, and instructions on digit-symbol performance of Negroes." Journal of Personality and Social Psychology, 1965, 2, 53-59.

Kluger, R. Simple Justice. New York: Random House, Inc., 1975.

Krol, R. A. "A meta-analysis of comparative research on the effects of desegregation on academic achievement." Unpublished Ph.D. dissertation, Western Michigan University, 1978.

Linsenmeier, J. A. W., and Wortman, P. M. "The Riverside school study of desegregation: A re-examination." Research Review of Equal Education, 1978, 2, 3-36.

Marston, A. R. "It is time to reconsider the Graduate Record Examination." American Psychologist, 1971, 26, 653-655.

McClelland, D. C. "Testing for competence rather than for intelligence." American Psychologist, Vol. 28, No. 1, January, 1973, 1-14.

Mercer, J. R., Iadicola, P., and Moore, H. "Building effective multiethnic schools: Evolving models and paradigms." In W. Stephan and J. Feagin, School Desegregation: Past, Present and Future. New York: Plenum, 1980.

Miller, N. "Changing views about the effects of school desegregation." In M. Brewer and B. Collins, Scientific Inquiry and the Social Sciences. San Francisco: Jossey-Bass, 1981.

Scholfield, J. W. "School desegregation and intergroup relations." In D. Bartal and L. Saxe (Eds.), Social Psychology of Education: Theory and Research. Washington, D. C.: Hemisphere, 1978.

Sheehan, D. S. "A study of attitude change in desegregated intermediate schools." Sociology of Education, 1980, 53, 51-59.

Shuey. A. M. The Testing of Negro Intelligence, (2nd ed.). New York: Social Science Press, 1966.

Stephan, W. G. "Intergroup relations." In G. Lindzey and E. Aronson (Eds.), The Handbook of Social Psychology. Reading, Mass.: Addison-Wesley, 1983.

Stephan. W. G. "School desegregation: An evaluation of predictions made in Brown vs. the Board of Education." Psychological Bulletin, 1978, 85, 217-238.

Stephan, W. G., and Stephan. C. W. "The role of ignorance in intergroup relations." Desegregation: Groups in Contact. New York: Academic Press, 1983.

Stephan, W. G., Rosenfield, D. "Effects of desegregation on racial attitudes." Journal of Personality and Social Psychology, 1978a, 36, 795-804.

Stephan, W. G., Rosenfield, D. "The effects of desegregation on race relations and self-esteem." Journal of Educational Psychology, 1978b, 70. 670-679.

Stephan, W. G., Rosenfield, D., "Black self-rejection: Another look." Journal of Educational Psychology, 1979, 71, 708-716.

St. John, N. H. School Desegregation: Outcomes for Children. New York: John Wiley and Sons. 1975.

Thompson, E. T. Plantation Societies, Race Relations, and the South: The Regimentation of Populations. Durham, N. C.: Duke University Press, 1975.

Wallach, M. A. "Tests tell us little about talent." American Scientist, 1976, 64, 57-63.

Weigel, R. H., Wiser, P. L., and Cook, S. W. "The impact of cooperative learning experience on cross-ethnic relations and attitude." Journal of Social Issues, 1975, 31, 219-244.

Weinberg, M. Desegregation Research: An Appraisal. Bloomington, Ind.: Phi Delta Kappa, 1970.

Williams, J. E., Morland, J. K. Race, Color, and the Young Child. Chapel Hill: University of North Carolina Press, 1976.

Williams, R. M., Jr. The reduction of intergroup tensions: A survey of research on problems of ethnic, racial, and religious group relations. New York: Social Science Research Council, Bulletin 57, 1947.

Desegregation and Education Productivity

Herbert J. Walberg
University of Illinois at Chicago

The purpose of the present paper is to analyze research on the impact of school desegregation on academic achievement. More specifically, the particular emphasis of this paper is the comparison of the effects of desegregation with those of other factors in the process of school learning that have been recently synthesized.

The paper is divided into three sections. The remainder of this first section discusses techniques and guidelines for research synthesis including meta-analysis. The second section presents a summary of the statistical analyses of research reviews of the 1970's and a collection of meta-analyses of the 1980's, which reveal the consistently potent productivity factors in school learning and which further illustrate techniques and guidelines for research synthesis. The third section assesses selection criteria for studies of school desegregation and achievement, and compares the effects of desegregation--as revealed by three recent meta-analyses--with the effects of the educational-productivity factors.

Research Synthesis

The present is an extraordinary time in the history of education because research syntheses are demonstrating the consistency of educational effects and are helping to put teaching and other determinants of learning on a sound scientific basis. Research synthesis is an attempt to apply scientific techniques and standards explicitly to the evaluation and summarization of research; it not only statistically summarizes effects across studies but also provides detailed, replicable rationales and descriptions of literature searches, selection of studies, metrics of study effects, statistical procedures, and overall results as well as those that call for exception with respect to context or subjects by objective statistical criteria (Glass, 1977; Cooper & Rosenthal, 1980; Jackson, 1980; Walberg & Haertel, 1980; Glass, McGaw, & Smith, 1981; and Light and Pillemer, 1982). Qualitative insights may be usefully combined with quantitative synthesis (Light & Pillemer, 1982); and quantitative results from multiple reviews and syntheses of the same or different topics may be compiled and compared to estimate their relative magnitudes and consistencies (Walberg, 1982).

Research synthesis is not merely statistical analysis of studies. Jackson (1980) discusses six tasks comprising an integrative review or research synthesis: specifying the questions or hypotheses for investigation; selecting or sampling the studies for synthesis; coding or representing the characteristics of the primary studies; analyzing, or meta-analyzing (Glass, 1977) or statistically synthesizing the study effects; interpreting the results; and reporting the findings.

Although these tasks seem obviously necessary to encourage replication of reviews, Jackson found only 12 out of 87 recent reviews in prominent educational, psychological, and sociological journals that provided even a cursory statement of methods. The basic idea behind much good advice in Jackson's paper is that the methods of review and synthesis should be explicit to enable other investigators to attempt to replicate the synthesis.

Explicit methods concerning quantitative synthesis, however, inevitably call for statistics, and two are most often employed--the vote count or box score, and the effect size (Glass, 1977). The vote count is easiest to calculate and explain to those who are unaccustomed to thinking statistically; it is simply the number of percentage of all studies that are positive, for example, in which the experimental exceed control groups or the independent variable correlated positively with the dependent variable.

The effect size is the difference between the means of the experimental and control groups divided by the control group standard deviation; it measures the average superiority (or, inferiority, if negative) of the experimental relative to the control groups (for cases in which these statistics are unreported, Glass (1977) provides a number of alternate estimation formulas). If education had uniform ratio variables, such as time and money as in economics, or physical measures in natural sciences such as meters and kilograms, effect sizes would be unnecessary; it could be said, for example, that the experimental groups grew .42 comprehension units in reading history on average, and the control group grew .22 units without crude post hoc standardization for comparability required in meta-analysis.

Effect sizes permit a rough calibration of comparisons across tests, contexts, subjects, and other characteristics of studies. The estimates, however, are affected by the variances in the groups, the reliabilities of the outcomes, the match of curriculum with outcome measures, and a host of other factors, whose influences in some cases can be estimated specifically or generally. Although effect sizes are subject to distortions, many of which may counterbalance one another, they are the only means of comparing the size of effects in primary research that employs various outcome measures on non-uniform groups. They are likely to be necessary until an advanced theory and science of educational measurement develops ratio measures that are directly comparable across studies and populations.

Generalizability

The generality of the results of the synthesis can be divided into questions of extrapolation and interpolation: Do the synthesized results generalize to other populations and conditions, particularly to those that have not been studied or for whom the results are unpublished? And, do the results generalize across populations and conditions for which results are available? Extrapolation may be invalid beyond published studies because journal editors favor positive, significant studies. Smith (1980) estimates from several syntheses that mean effect sizes in unpublished work, mainly doctoral dissertations, are occasionally larger but average about a third smaller than those in published studies.

Rosenthal (1980), on the other hand, shows that given the great statistical significance of collections published studies, the probability of null effects being established by unpublished studies is minimal. Furthermore, both the low reliability of educational measures and low curricular validity (correspondence of what is taught and what is tested on outcome measures) diminish the estimates of relations between educational means and ends. Less than optimal reliability and validity, which leads to underestimates of effects, probably more than compensate for publication bias; but more empirical and analytic work is needed on these factors to determine their general and specific influences on synthesis results.

Interpolation

The interpolation problem can be readily solved by additional calculations. The most obvious questions in quantitative synthesis concern the overall percentage of positive results and their average magnitude. But the next questions should concern the consistency and magnitude of results across student and teacher characteristics, educational treatments and conditions, subject matters, study outcomes, and validity factors in the studies. These questions can be answered by calculating separate results for classifications or cross-classifications of effects.

The results may be compared by objective statistical tests (such as T, F, and regression weights in general linear models). They permit conclusions on such matters as the overall effectiveness of treatments as well as their differential effectiveness on categories of students in various conditions and different outcomes. Notwithstanding the frequent claims by reviewers for differential effects on the basis of results of a few selected studies, most research syntheses yield results that are robust and roughly consistent across such categories. Such robustness is scientifically valuable because it indicates parsimonious, law-like findings; it is also educationally valuable because educators can apply robust findings more confidently and efficiently rather than using complicated, expensive procedures, tailor-made on unproven assumptions to special cases.

A number of useful methodological writings are available. Glass (1977) providea a concise introduction to statistical methods; and Glass, McGaw, and Smith's (1981) book presents a comprehensive treatment. Jackson (1980) and Cooper (1982) discuas tasks and criteria for integrative reviews and research syntheses. Light and Pillemer (1982) describe methods for combining quantitative and qualitative methods. Walberg and Haertel (1980) present a collection of eight methodological papera by Cahen, Cooper, Hedges, Light, Rosenthal, Smith and othera and thirty-five substantive papers mostly on educational topica. In forthcoming work, Larry Hedges of the University of Chicago and Barry McGaw of Murdoch University (Australia) offer firmer statistical and psychometric footings for quantitative synthesis. Important guidelines for research synthesis that may be found in these works are further discussed and illustrated in the remaining sections.

Educational Productivity Factors

A Review of Reviews of Teaching Effects

The year 1980 marked a transitional period when investigators recognized the shortcomings of the traditional review and the advantages of more objective, explicit procedures for evaluating and summarizing research. Yet reviews still have a place, and much can be learned from them. Waxman and Walberg (1982) examined 19 reviews of teaching process-student outcome research published during a recent decade that critically reviewed at least three studies and two teaching constructs; they described their methods, compared their conclusion, synthesized them, and pointed out the implications for future reviews, syntheses, and prior research.

The 19 reviews reflect the inexplicit, varied, and vague standards revealed by Jackson's (1980) analysis of 87 review articles in prominent educational, psychological, and sociological journals. None of the reviews, for example, described their search procedures, and only one stated explicit criteria for inclusion and exclusion of primary studies. Comparative analysis of the studies, moreover, revealed that the reviewers failed to search diligently enough for primary studies or to state the reasons for excluding large parts of the research evidence. Among the five reviews that covered positive reinforcement such as praise and feedback in teaching, only six studies were covered in the most comprehensive review in contrast to the 39 listed in Lysakowski and Walberg's (1981) synthesis. Such arbitrary selection of small parts of the evidence, of course, leaves the reviews open to systematic bias and means that the reviews and their conclusions cannot be replicated in a strict sense because their methods are undescribed.

Although the reviews purported to be critical, their coverage of the 33 standard threats to methodological validity (Cook & Campbell, 1979) was spotty and haphazard. In 95.4 percent of the possible instances, the reviews ignored specific threats. External validity (interaction of teaching treatments with selection, setting, and history) was relatively well covered, perhaps reflecting the search and claims for aptitude-treatment interactions of the 1970's; but the serious problems of internal validity, such as reverse and exogenous causes in correlational studies, were almost wholly ignored. Indeed, there appeared an odd tendency to select correlational studies rather than experiments for review.

Despite these problems, however, a statistical tabulation of the conclusions of the reviews shows substantial and statistically-significant agreement that five broad teaching constructs--cognitive cues, motivational incentives, engagement, reinforcement, and management and climate--are positively associated with student learning outcomes (see Table 1). These tabulations, moreover, are in close agreement with quantitative syntheses of large, systematic collections of primary studies discussed in a subsequent section.

Current Research Syntheses

To characterize quantitative syntheses of educational research completed since 1979, sixteen were found in 1982 by scanning publications of the American Educational Research Association and writing to the members of "the invisible college" of about 100 scholars that meet annually to present and discuss research on teaching. A more systematic search in late 1982 ısing Dissertation Abstracts, Social Science Citation Index, Education Index, computer retrieval, and references in recent publications indicates that these syntheses plus those discussed in subsequent sections of this chapter represent about three-fourths of those completed in education thusfar in the 1980s. (An analysis of a more complete corpus is underway by the present author and colleagues, but the increasing number of syntheses makes exhaustive coverage an elusive goal.)

Table 2 suggests a number of instructive points for both educational practice and research synthesis. It provides, for example, an empirical answer to the coincidence of vote counts and effect sizes. Every mean effect size that was positive also had a vote count greater than 50 percent; every negative effect size had a vote count less than 50 percent. Thus, as may be expected from normal distributions, consistently positive findings will yield positive average results (the next section shows that much of the variance in effects can be predicted by regression from counts). The likely explanation for the uniform association is that strong causes produce results consistent in sign. Indeed, the only cases in which the association can be reversed are skewed distributions in which a few very strong positive results are sufficient to pull the mean above zero from a cluster of small effects, more than half of which are negative (or vice versa).

The first two syntheses grouped under Teaching Strategies in Table 2 show fairly close agreement with respect to the consistency of cooperative learning. Johnson and others (1981) categorized their results by comparisons of four treatment variations (cooperative, competitive, group competitive, and individualistic), whereas Slavin (1980) categorized his results by outcomes. Cooperative learning obviously produces superior results; but it would be useful if journal editors would allow research synthesis space to report average results by more standard classifications of independent and dependent variables and study conditions to facilitate comparisons of replicated syntheses such as these two.

Table 1

Conclusions of 19 Reviews and 2 Quantitative Syntheses of Research on Teaching

	Stimulation				
	Cognitive Cues	Motivational Incentives	Engagement	Reinforcement	Management and Climate
Number of Reviews Covering Construct	19	5	10	13	15
Number of Reviews Concluding Relation to Learning is Positive	17	5	10	9.5	13.5
Probability of an Even Split	.01	.10	.01	.10	.01
Mean Effect Sizes from Quantitative Synthesis	1.28		.88	.94	1.17
Probability of Evidence Assuming Zero Population Effect	.01			.01	

Table 2

Selected Post-1979 Quantitative Syntheses

Author	Number of Studies	Independent and Dependent Variables	Mean Correlation or Effect	Percent Positive	Comments
Teaching Strategies					
Johnson, Maruyama, Johnson, Nelson, and Skon (1981)	122	Effects of cooperation, intergroup and interpersonal competition, and individual goal effects on achievement and productivity	.00	54	Cooperative vs. group competitive
			.78	76	Cooperative vs. competitive
			.37	68	Group competitive vs. cooperative
			.76	83	Cooperative vs. individualistic
			.59	81	Group competitive vs. individualistic
			.03	47	Competitive vs. individualistic
Slavin (1980)	28	Effects of educational programs for cooperative learning		81	Curriculum-specific tests
				78	Standardized tests
				* 95	Race relations
				65	Mutual concern
Becker & Gersten (1982))	1	Effects of Direct Instruction Follow Through on later achievement (7 sites on 2 occasions, fifth and sixth grades)	.23	--	Effects larger for mathematics problem solving and for fifth grade
Pflaum, Walberg, Karegianes, and Rasher (1980)	96	Effects of different methods teaching reading on learning	.60	76	Although Hawthorne effects could be discounted, experimental groups generally did substantially better than controls; sound-symbol blend was one standard deviation higher than other treatments.

Table 2 (page 2 of 3)

Author	Number of Studies	Independent and Dependent variables	Mean Correlation or Effect	Percent Positive	Comments
Teaching Skills					
Luiten, Ames, and Anderson (1980)	135	Effects of advance organizers on learning and retention	.23	--	Effects larger on 20+ days retention higher achievers, college students and when presented aurally
Redfield and Rousseau (1981)	20	Effects of higher and lower cognitive questions	.73	--	Higher questioning effects greater training than in skills study and in more valid studies
Wilkinson (1980)	14	Effects of praise on achievement	.08	63	Praise slightly more effective for lower socioeconomic groups; primary grades, and in mathematics
Other Studies					
Butcher (1981)	47	Effects of microteaching lessons on teaching performance of secondary and elementary education students	.84 .56 .46 .35		Secondary specific skills Secondary questioning skills Elementary specific skills Elementary questioning skills
Colosimo (1981)	24	Effects of practice and beginning teaching on self attitudes	-.29	48	Initial experience associated with greater authoritarianism and self-doubt; inner-city experience more negative
Findley and Cooper (1981)	98	Correlations of locus of control and achievement	.18	79	Correlations higher among males; for adolescents in constrast to children and adult groups; for specific control measures; and for objective achievement

Table 2 (page 3 of 3)

Author	Number of Studies	Independent and Dependent Variables *	Mean Correlation or Effect	Percent Positive	Comments
Hansford and Hattie (1982)	128	Correlation of self-concept and achievement/performance	.21	84	Correlations higher for high school students in contrast to elementary and college; higher ability student specific rather than global self-concept; and verbal achievement measures
Carlsburg and Kavale (1980)	50	Effects of special versus regular classes	-.12	--	Effects positive for learning disabled and behavior disordered and negative for slow learners and mentally retarded
Otterbacher and Cooper (1981)	43	Effects of class placement of mentally retarded students on social adjustments	.05 -.07	61 46	Special class vs. regular class Special class vs. resource class
Smith and Glass (1980)	59	Effects of class size on attitudes, climate, and instruction	.49	--	In contrast to small mean effect of .01 for achievement, moderate effect observed, which were larger on teachers than students, younger students, and for studies before 1969
Williams, Haertel, Haertel, and Walberg (1982)	23	Correlations of leisure time television and achievement	-.05	34	Effects negative at ratio of less than 5 or greater than 15 hours per week and stronger for girls and higher ability groups
Willson and Putnam (1982)	32	Effects of pretests on outcomes	.17	57	Effects greater for cognitive and personality outcomes, for treatment lasting between 2 and 30 days, and for randomized studies

The next two syntheses raise important, unresolved methodological questions. Becker and Gersten's (1982) synthesis indicated a small average effect of direct instruction in several sites, but all effect sizes came from the same study. Although teachers in the various sites may have been independent actors, methodological bias can make the effects non-independent from a statistical point of view, and independent replications by different investigators would be in order to a provide a more definitive answer. Pflaum and others (1980) found no average superiority of different reading methods but a substantial advantage in learning outcomes of experimental over control groups no matter what the reading method employed. Although Hawthorne effects could be discounted by the synthesis, the increased energy and attention devoted to tasks by teachers in experimental groups rather than putative treatments themselves may partly account for superior results in teaching-methods and other educational studies.

Table 2 includes two rough replications that indicate substantial agreement in results despite large variations in study search, selection, and numbers. Hansford and Hattie's (1982) and Findley and Cooper's (1981) syntheses of correlations of self-concept and locus of control with achievement and performance differ only slightly in the second decimal place in both the vote counts and average correlations. Carlberg and Kavale's (1980) and Ottenbacher and Cooper's (1981) syntheses agree that the effects of mainstreaming (federally-encouraged efforts in the United States to mix regular and cognitively, emotionally, and physically handicapped children in the same classes) are inconsistent and probably near zero.

Two syntheses show curvilinear effects of independent variables on educational outcomes. Smith and Class (1980) found that the benefits of reduced class size are larger at the smaller ranges of one to 10 members than they are at higher ranges; for example, the measurable cognitive and affective outcome differences between classes of 20 and 60 appear trivial. Similarly, Williams and others (1982) found decreasing achievement with departures from 10 weekly hours of leisure-time television viewing such that estimated differences in achievement between children who watch about 30 hours--an average number--and 60--a large amount--are miniscule.

Other effects are summarized in the table, and the reader is referred to the original syntheses for details that are not discussed here. Overall, the results indicate a large range of effects, which, if replicated in further primary research and syntheses, would have fairly definite implications for choosing policies and practices that seem likely to have consequential effects on raising educational outcomes.

The Michigan Program

Chen-Lin and James Kulik lead a vigorous group of research synthesists at the University of Michigan, which included Peter Cohen, now of Dartmouth. The group has been unusually productive of high-quality syntheses first in higher education and later in secondary-school research. Personal communications with the group reveal that their team approach, much like that described by Shulman and Tamir (1973) in the Second Handbook of Research on Teaching, accounts in part for the quantity and quality of work.

James Kulik kindly prepared Table 3 according to the present author's specifications. It shows the results of eleven syntheses completed by the Michigan group by the end of 1981. Like the sixteen syntheses by other investigators discussed in the last section, those in Table 3 show a number of consistent moderate to large effects that can help to put high school and college teaching on a firm scientific basis.

Kulik's results also permit an estimate of the mean size of effects from vote counts. The regression equation, ES + -.403 + .008 (% Positive), accounts for 76 percent of the variance in the effect sizes. The corresponding equation for the syntheses in Table 2 for which both indexes are available, ES = -.761 + .015 (% Positive), accounts for 59 percent of the effect-size variance (the correlational results assume both causality and a one-unit increase in the independent variable). Both equations forecast near zero effect sizes for vote counts of 50 percent; but the higher slope for the results in Table 2 forecast larger effects than do the Michigan data; at vote counts of 75 percent, for example, the respective forecasts are .36 and .20. Thus the size of the regression slope is unstable across samples, and more intensive analyses of the complete corpus of syntheses are in order.

The two data sets also permit separate empirical estimates of the distributions of vote counts and effects. The mean (and standard deviations) of Michigan and other estimates of the vote counts are respectively 67 and 64 (and 19 and 16); the mean effects are respectively .17 and .22 (and .19 and .31). Assuming normal distributions of effects, empirical norms for vote counts and effect sizes can be set forth on the basis of the averages of these statistics; for example, the middle two-thirds of the effects in the recent educational research sampled range from about -.05 to .45. It could be said that effect sizes of .20 are average, and those above .45 are large and exceed about 84 percent of those typically found in educational research. Similarly, vote counts of 67 and 85 percent might be provisionally taken as average and large. These norms are, of course, very rough and preliminary, but they are based on empirical results rather than opinion and may be useful in gauging present and future results until larger normative samples are analyzed.

Syntheses of Bivariate Productivity Studies

A group at the University of Illinois at Chicago has concentrated on synthesizing research on nine theoretical constructs that appear to have consistent causal influences on academic learning: student age or development level, ability (including prior achievement), and

Table 3

Major Results from Quantitative Syntheses Conducted at the University of Michigan's

Center for Research on Learning and Teaching

Report	Independent Variable	Dependent Variable	Studies		Effect Size		Comments
			Number	% Positive	Mean	SD	
Bangert, J. Kulik, & C. Kulik (1981)	Individualized vs. conventional secondary teaching	Achievement on final examination	49	65	0.10	0.38	
		Attitude toward subject matter	14	94	0.14	0.21	
Cohen (1980)	Midsemester rating feedback to teacher vs. no feedback	Change on final ratings.	22	91	0.38	0.41	Effects were greater when teachers received consulting help along with rating feedback.
Cohen (1981)	Class rating of instructor quality	Class achievement on final examination	67	88	0.43	0.13	Correlations were higher when teachers were faculty (not teaching assistants), when all tests were graded by a common grader, and when students rated teachers after receiving grades.
Cohen, Ebeling, & J. Kulik (1981)	Visual-based vs. conventional college teaching	Achievement on final examination	65	57	0.15	0.41	Achievement effects were stronger in more recent studies, in studies from universities, & when different teachers taught visual-based & control classes.
		Student rating of course quality	16	38	-0.08	0.68	
		Course completion	10	30	-0.05	0.23	

Table (Continued)

Report	Independent Variable	Dependent Variable	Studies		Effect Size		Comments
			Number	% Positive	Mean	SD	
J. Kulik, Cohen, & Ebeling (1980)	Programmed vs. conventional college teaching	Achievement on final examination	56	71	0.24	0.52	Achievement effects were stronger in more recent studies.
		Course completion	9	61	+0.08	0.27	
J. Kulik, C. Kulik, & Cohen (1979a)	Personalized System of Instruction vs. conventional college teaching	Achievement on final examination	61	94	0.49	0.33	Achievement effects differed by subject and were stronger when different teachers taught PSI and control classes, and when control classes contained PSI features.
		Course completion	27	37	-0.10	0.30	
		Rating of course quality	11	91	0.46	0.65	
J. Kulik, C. Kulik, & Cohen (1979b)	Audio-tutorial vs. conventional college teaching	Achievement on final examination	42	65	0.20	0.43	Achievement effects were stronger in studies found in journals.
		Course completion	22	53	-0.10	0.27	
		Rating of course quality	6	50	0.13	0.53	
J. Kulik, C. Kulik, & Cohen (1980)	Computer-based vs. conventional college teaching	Achievement on final examination	54	69	0.25	0.61	Achievement effects were stronger when different teachers taught computer-based and control classes.
		Course completion	13	46	0.01	0.30	
		Rating of course quality	11	73	0.24	0.53	

motivation; amount and quality of instruction; the psychological environments of the class, home, and peer group outside school; and exposure to the mass media (Walberg, 1981). The group first collected available vote counts and effect sizes in the review literature of the 1970's and then conducted more systematic syntheses directly on the nine factors. This section summarizes both efforts.

Synthesis of reviews of the 1970's. Walberg, Schiller, and Haertel (1979) collected reviews published from 1969 to 1979 on the effects of instruction and related factors on cognitive, affective, and behavioral learning in research conducted in elementary, secondary, and college classes and indexed in standard sources. The vote counts for the corpus of reviews are shown in Table 4.

The vote counts should be cautiously interpreted because not only may journal editors more often select studies with positive results but also reviewers may select positive published studies for summarization. Neither editors nor reviewers ordinarily state explicit policies on these important points. Subsequent, more systematic syntheses, nonetheless, have generally supported traditional reviews; and it would be wasteful to ignore the labors of the last decade of effort, even though it may only be considered a starting point for subsequent work.

Notwithstanding the possible double bias in the vote counts (see earlier sections on counter-biases), the results in Table 4 are impressive. A majority of the variables in the table were positively associated with learning; in 48 or 68 percent of the 71 tabulations, 80 percent or more of the comparisons or correlations are positive. Although all of the variables are candidates for synthesis using systematic search, selection, evaluation, and summarization procedures, it appears that the 1970's produced reasonably consistent findings that are likely to be confirmed by more comprehensive and explicit methods of the present decade.

Syntheses of Productivity Factors. The Chicago group also carried out syntheses of the nine factors using methods discussed in previous sections of this chapter. The National Institute of Education supported the syntheses of learning research in ordinary classes, grades kindergarten through twelve. A separate grant from the National Science Foundation on science learning, grades 6 through 12, permitted more exhaustive, intensive search for unpublished work and an advisory group of science educators and research methodologists as well as a semi-independent replication of the results for several of the factors. A summary of the findings is shown in Table 5.

All of the effect sizes (including mean contrasts and correlations) are in the expected direction. The mean effects for the two samples of studies are similar in magnitude, which suggests generality or robustness of effects across more and less intensive methods of

Table 4

A Selective Summary of a Decade of Educational Research

Research Topics	No of Results	Percent Positive
Time on learning	25	95.4
Innovative curricula on		
Innovative learning	45	97.8
Traditional learning	14	35.7
Smaller classes on learning		
Pre-1954 studies	53	66.0
Pre-1954 better studies	19	84.2
Post-1954 studies	11	72.7
All Comparisons	691	60.0
Behavioral instruction on learning	52	98.1
Personal systems of instruction on learning	103	93.2
Mastery learning	30	96.7
Student-vs. instructor-led discussion on.		
Achievement	10	100.0
Attitude	11	100.0
Factual vs. conceptual questions on achievement	4	100.0
Specific teaching traits on achievement		
Clarity	7	100.0
Flexibility	4	100.0
Enthusiasm	5	100.0
Task-orientation	7	85.7
Use of student ideas	8	87.5
Indirectness	6	83.3
Structuring	9	100.0
Sparing criticism	17	70.6

Table 4 (Continued)

Research Topics	No of Results	Percent Positive
Psychological incentives and engagement		
Teacher cues to student	10	100.0
Teacher reinforcement of student	16	87.5
Teacher engagement of class in lesson.	6	100.0
Individual student engagement in lesson	15	100.0
Open vs. traditional education on:		
Achievement	26	54.8
Creativity	12	100.0
Self-concept	17	88.2
Attitude toward school	25	92.0
Curiosity	6	100.0
Self-determination	7	85.7
Independence	19	94.7
Freedom from anxiety	8	37.5
Cooperation	6	100.0
Programmed instruction on learning	57	80.7
Adjunct questions on learning		
After text on recall	38	97.4
After text on transfer	35	74.3
Before text on recall	13	76.9
Before text on transfer	17	23.5
Advance organizers on learning	32	37.5
Analytic revision of instruction on achievement	4	100.0
Direct instruction on achievement	4	100.0
Lecture vs. discussion on:		
Achievement	16	68.8
Retention	7	100.0
Attitudes	8	86.0
Student-vs. instructor-centered discussion on:		
Achievement	7	57.1
Understanding	6	83.3
Attitude	22	100.0
Factual vs. conceptual questions on achievement	4	100.0
Social-pschological climate and learning:		
Cohesiveness	17	85.7
Satisfaction	17	100.0
Difficulty	16	86.7
Formality	17	64.7
Goal direction	15	73.3
Democracy	14	84.6
Environment	15	85.7
Speed	14	35.8
Diversity	14	30.8
Competition	9	66.7
Friction	17	0.0
Cliqueness	15	8.3
Apathy	15	14.3
Disorganization	17	6.3
Favoritism	13	10.0
Motivation and learning	232	97.8
Social class and learning	620	97.6
Home environment on:		
Verbal achievement	30	100.0
Math achievement	22	100.0
Intelligence	20	100.0
Reading gains	6	100.0
Ability	8	100.0

Table 5

Correlations and Effect Sizes for Nine Factors in Relation to School Learning

Factor	Number of Studies	Results and Comment
Instruction		
Amount	51	Correlations range from .15 to .71 with a median of .40. partial correlations controlling for ability, socioeconomic status, and other variables range from .09 to .60 with a median of .35
Quality	95	The mean of effect sizes for reinforcement in 39 studies is 1.17, suggesting a 38-point percentile advantage over control groups, although girls and students in special schools might be somewhat more benefited. the mean effect sizes for cues, participation, and corrective feedback in 54 studies is .97, suggesting a 33-point advantage. The mean effect size of similar variables in 18 science studies is .81.
Social-psychological Environment		
Educational	12	On 19 outcomes, social-psychological climate variables added from 1 to 54 (median = 20%) to accountable variance in learning beyond ability and pretests: the signs and magnitudes of the correlations depend on specific scales (see Table 1), level of aggregation (classes and schools higher), nation, and grade level (later grades higher); but not on sample size, subject matter, domain of learning (cognitive, affective, or behavioral), or statistical adjustments for ability and pretests
Home	18	Correlations of achievement, ability, and motivation with home support and stimulation range from .02 to .82 with a median of .37, multiple correlations range from .23 to .81 with a median of .44, studies of boys and girls and middle-class children in contrast to mixed groups show higher correlations (social classes correlations in 100 studies, by contrast, have a median of .25). The median correlations for three studies of home environment and learning in science is .32.
Media-TV	23	274 correlations of leisure-time television viewing and learning ranged from −.56 to .35 with a median of −.06, although effects appear increasingly deleterious from 10 to 40 hours a week and appear stronger for girls and high-IQ children.
Peer group	10	The median correlation of peer group or friend characteristics such as socioeconomic status and educational aspirations with achievement-test scores, course grades, and educational and occupational aspirations is .24; correlations are higher in urban settings and in studies of students who reported aspirations and achievements of friends. The median of two sciences studies is .24.
Aptitude		
Age-development	9	Correlations between Piaget developmental level and school achievement range from .02 to .71 with a median of .35. The mean correlation in sciences is .40.
Ability	10	From 396 correlations with learning, mean verbal intelligence measures are highest (mean = .72) followed by total ability (.71), nonverbal (.64), and quantitative (.60); correlations with achievement test scores (.70) are higher than those with grades (.57). The mean ability-learning correlation in science is .48.
Motivation	40	Mean correlation with learning is .34, correlations were higher for older samples and for combinations of subjects (mathematics) and measures, but did not depend on type of motivation nor the sex of the samples The mean of three studies in science is .33.

synthesis. In particular, the syntheses of quality of instruction including cues, participation, and reinforcement of about 1.0 and .8 in general grades K-12 and in science grades 6-12 support the conclusions of the 19 reviews discussed in a previous section (see also Table 1). Despite these corroborations of findings, of course, independent replications of the syntheses as well as new and probing experimental studies are needed.

Syntheses of Multivariate Studies

The Chicago group also conducted multivariate analyses of the productivity factors in samples of from two to three thousand 13- and 17-year-old students who participated in the mathematics, social studies, and science parts of the National Assessment of Educational Progress (see, for example, Walberg, Pascarella, Haertel, Junker, and Boulanger, 1981, 1982). These survey analyses complement small-scale correlational and experimental studies in providing on representative national samples data on fairly comprehensive sets of the productivity factors, each of which may be statistically controlled for the others in multiple regressions of achievement and subject-matter interest.

Such analyses allow a simultaneous assessment of qualities and amounts of instruction and the other factors in the production of learning. Since the factor levels are reported as experienced by individual students, the analyses are sensitive to micro-variations in the multiple environments of the school, peer-group, home, and mass media to which each student is exposed.

Although the sets of variables available in the National Assessment can be used to assess possible exogenous causes because they are measured and can be statistically controlled in regression equations, the measures are cross-sectional for individuals. Therefore, they cannot effectively rule out reverse causation such as learning as a cause of motivation and more stimulating teaching. Another shortcoming of the data is that parental socioeconomic status serves as a proxy for ability and prior achievement.

As pointed out above, nonetheless, the strengths of the National Assessment data complement those of small-scale bivariate studies that typically control for only one or two of the factors. If syntheses of both data sources point in the same direction, then more confidence can be placed in the conclusions.

Table 6 shows that the factors, when controlled for one another, are surprisingly consistent in sign, significance, and magnitude across subject matters, ages, operational measures of the factors, and independent national samples. The median standardized regression weights and squared multiple correlations, shown in the last row, reveal the small to moderate effects of the factors when controlled for one another and sizable amounts of varirnce accounted for even without ability and prior achievement measures.

Table 6

Regressions of Avhievement on Productive Factors

Supet-Standardized Weights

	Age	Sample Size	Achievement	Attitude	SES	uality of Instruction	Quantity of Instruction	Education (Class)	Home	Peer	Extra-Curricular Activities	Homework	Stimulation	Media-TV	R^2
Science Achievement	13	2,346		.0111*	.0125*	.0097		.0147**	.0119**	.0069					.11
Science Acievement	17	3,049		.0111*	.0176**	.0101		.0340**	.0113**	.0126**		.0113**			.36
Math: Achievement	17	1,480		.0041*	.0051**	.0174** .0369*$^{a}_{b}$	.0143		.0141					-.0062	.57
Math Atitude	17	1,480	.125			.325**$_b$ .100$_c$			-.225**			.325**	.425**		.09
Social Science Achievement	11	2,426		.0996**	.0325**		.0152**		.0346**						.42
Social Science Attitude	11	2,426	.1478**			.0348**$_b$	.0174*		.0174*						.39
Social Science Achievement	17	2,001		.0586**	.0330**	.0220**$_a$	.0220**		.0256**					-.0110**	.37
Social Science Attitude	17	2,001	.1056**			.0248**$_b$	.0280**		.0217**		.0652**	.0217**	.0186**		.36

* $p<.05$
** $p<.01$
a Traditional instruction
b Student-centered instruction
c Most advanced course

Syntheses of Open Education Research

Open education is an elusive concept, now dismissed by many educators, but one that research synthesis now illuminates. The history of efforts to synthesize its effects is instructive about: the dangers of basing conclusions, policies, and practices on single studies; replication and improved methods of syntheses, and a shortcoming of much of the research discussed above that employs grades and standardized achievement as the sole outcomes of teaching.

From the start, open educators tried to encourage educational outcomes that reflect school-board goals such as cooperation, critical thinking, self-reliance, constructive learning attitudes, life-long learning, and other goals that evaluators seldom measure. Raven's (1981) summary of surveys in Western countries including England and the United States shows that educators, parents, and students rank these goals far above standardized test achievement and grades.

A synthesis of the relation of conventionally-measured educational outcomes and adult success, moreover, shows their slight association (Samson and others, 1982). Thirty-three post-1949 studies of physicians, engineers, civil servants, teachers, students in general, and other groups show a mean correlation of .155 of these educational outcomes with success indicators such as income, self-rated happiness, work performance and output indexes, and self-, peer-, and supervisor-ratings of occupational effectiveness. These results should challenge educators and researchers to seek a balance between continuing motivation and skills to learn and perform well on new tasks as an individual or group member on one hand and mastery of teacher-chosen, textbook knowledge that may soon be obsolete or forgotten on the other.

Perhaps since Socrates, however, arguments over student-centered and teacher-centered education have remained so polarized, polemical, and pervasive that educators find it difficult to stand firmly on the high middle ground of balanced, joint, or cooperative determination of the goals, means, and evaluation of learning. Progressive education, the Dalton and Winnetka plans, team teaching, the ungraded school, and other innovations in this century held forth this ideal but gravitated toward authoritarian teaching or permissiveness and could not be sustained. Although open education, too, faded from view, it was more carefully researched; and syntheses of it may help prepare educators for evaluating future efforts.

Three Syntheses of Open Education. Horwitz (1979) first synthesized about 200 comparative studies of open and traditional education by tabulating vote counts by outcome category. Although many studies yielded non-significant or mixed results especially with respect to academic achievement, self-concept, anxiety, adjustment, and locus of control, more positive results were found in open education on attitudes toward school, creativity, independence, curiosity, and cooperation.

Peterson (1979) calculated effect sizes for the 45 published studies. She found about -.1 or slightly inferior effects of open education on reading and mathematics achievement; .1 to .2 effects on creativity, attitudes toward school, and curiosity; and .3 to .5 effects on independence and attitudes toward the teacher.

Hedges, Giaconia, and Gage (1981) synthesized 153 studies including 90 dissertations using an adjustment of Glass's effect-size estimator which is slightly biased especially in small samples. The average effect was near zero for achievement, locus of control, self-concept, and anxiety; about .2 for adjustment, attitude towards school and teacher, curiosity, and general mental ability; and about .3 for cooperativeness, creativity, and independence.

Despite the differences in study selection and synthesis methods, the three studies converge roughly on the same plausible conclusion: Students in open classes do slightly or no worse in standardized achievement and slightly to substantially better on several outcomes that educators, parents, and students hold to be of great value. Unfortunately, the negative conclusion of Bennett's (1976) single study--prefaced by a prominent psychologist, published by Harvard University Press, publicized by The New York Times and ~~media and experts~~ that take that newspaper as their source--probably sounded the death knell of open education, even though the conclusion of the study was later retracted (Aitkin, Bennett, & Hesketh, 1981) because of obvious statistical flaws in the original analysis (Aitkin, Anderson, & Hinde, 1981).

Components of Open Education. Giaconia and Hedges (1982) took another recent and constructive step in the synthesis of open education research. From the prior effect-size synthesis, they identified the studies with the largest positive and negative effects on several outcomes to differentiate more and less effective program features. They found that programs that are more effective in producing the non-achievement outcomes--attitude, creativity, and self-concept--sacrificed academic achievement on standardized measures.

These programs were characterized by emphasis on the role of the child in learning, use of diagnostic rather than norm-referenced evaluation, individualized instruction, and manipulative materials but not three other components sometimes thought essential to open programs--multi-age grouping, open space, and team teaching. Giaconia and Hedges speculate that children in the most extreme open programs may do somewhat less well on conventional achievement tests because they have little experience with them. At any rate, it appears from the two most comprehensi' syntheses of effects that open classes on average enhance several non-standard outcomes without detracting from academic achievement unless they are radically extreme.

Synthesis of Instructional Theories

To specify the productivity factors in further theoretical and operational detail that provide a more explicit framework for future primary research and synthesis, Haertel, Walberg, and Weinstein (1983) compared eight contemporary psychological models of educational performance. Each of the first four factors in Table 7--student ability and motivation, and quality and quantity of instruction--may be essential or necessary but insufficient by itself for classroom learning (age and developmental level are omitted because they are unspecified in the models).

The other four factors in Table 7 are less clear: although they consistently predict outcomes, they may support or substitute for classroom learning. At any rate, it would seem useful to include all factors in future primary research to rule out exogenous causes and increase statistical precision of estimates of the effects of the essential and other factors.

Table 7 shows that, among the constructs, ability and quantity of instruction are widely and relatively richly specified among the models. Explicit theoretical treatments of motivation and quantity of instruction, however, are largely confined to the Carroll tradition represented in the first four models; and the remaining factors are largely neglected.

The table poses empirically researchable theoretical questions; the tension between theoretical parsimony and operational detail, for example, suggests several: Can the first four constructs mediate the causal influences of the last four? Would assessments of Glaser's five student-entry behaviors allow more efficient instructional prescriptions than would, say, Carroll's, Bloom's, or Bennett's more general and more parsimonious ability subconstructs? Would less numerous subconstructs than Gagne's eight instructional qualities and Harnischfeger's and Wiley's seven time categories suffice?

The theoretical formulation of educational performance models of the past two decades since the Carroll and Bruner papers has made rapid strides. The models are explicit enough to be tested in ordinary classroom settings by experimental methods and production functions. Future empirical research and syntheses that are more comprehensive and better connected operationally to these multiple theoretical formulations should help reach a greater degree of theoretical and empirical consensus as well as more effective educational practice.

Table 7

Classification of Constructs According to the Model of Educational Productivity

Theorist	Ability	Motivation	Quality of Instruction	Quantity of Instruction	Social Environment of Classroom	Home Environment	Peer Influence	Mass Media
Carroll (1963)	Aptitude Ability to comprehend instructions	Perserverance	Clarity of instruction Matching task to student characteristics	Opportunity to learn (time)	--			
Cooley and Leinhart (1975)	General ability Prior achievement	Motivators (internal)	Motivators (external) Structure Instructional Events Attitude toward teachers	Opportunity to learn (time)	Attitudes toward school		Attitudes toward peer	
Bloom (1976)	Prior achievement Reading comprehension Verbal IQ	Attitude toward subject matter Self-concept as learner	Use of cues Reinforcement Feedback and correctives	Participation in learning task (time)	Attitudes toward school			
Harnischfeger and Wiley (1976)	Pupil background	Intrinsic motivation	Teacher activities	Pupil pursuits (7 time categories)				
Bennett (1978)	Aptitude Prior achievement	Implicit	Clarity of instruction Task difficulty and pacing	Total active learning time Quantity of schooling Time allocated to curriculum activity				
Gagne (1977)	Internal conditions of learning	Implicit	Activating motivation Informing learner of objective Directing attention Stimulating recall Providing learning guidance Enhancing retention Promoting transfer of learning Eliciting performance and providing feedback					
Glaser (1976)	Task learnings already acquired Prerequisite learnings Cognitive style Task specific aptitudes General mediating ability	Implicit	Materials, procedures, and techniques that foster competence (e.g., knowledge structures; Learning-to-learn; Contingencies of reinforcement) Assessment of effects of instruction					
Bruner (1966)	Task relevant skills	Predispositions	Implanting a predisposition toward learning Structuring knowledge Sequence of materials Specifying rewards and punishments					

Desegregation and Educational Productivity

As the previous section has shown, sufficient empirical and theoretical syntheses have accumulated during the past five years to point more definitively than ever before to the proximal, alterable factors that affect educational achievement. Nearly all the research has been carried out in natural settings such as homes and schools, and most of its shows generalizability across student characteristics, subjects, and research methods, including randomized assignment to experimental treatments.

The large average magnitude and consistency of many of these productive factors justly provides a substantial amount of confidence about how educational achievement may be raised. Since many of the factors and techniques have already been extensively employed in ordinary schools and found successful, inexpensive, and non-controversial, it appears that educational achievement might be increased substantially by implementing a selection of the most productive of the factors, say, those with effect sizes above .3, more extensively and intensively. The purpose of this section is to compare the consistency and magnitude of such factors to the effects of school desegregation, as revealed by three recent meta-analyses--Krol (1978), Crain and Mahard (1982), and my statistical summary of the studies meeting the selection criteria of the National Institute of Education (NIE) panel of scholars.

Selection Criteria

Aside from the inclusion of data only on Black students in all three meta-analyses, Krol (1978, p. 16), Crain and Mahard (1982, p. 6) and the NIE panel (Schneider, Note 1) varied considerably in explicit criteria for study selection. Krol, for example, excluded studies that lacked achievement measures before and after desegregation and those that lack sufficient statistics to calculate effect sizes (pp. 83-84). Excluding studies without pretests turns out to be a reasonable decision because Wortman's (Note 2) research shows desegregated groups are on average advantaged on achievement before desegregation. Thus apparent posttest advantages of desegregation are in part attributable to pre-existing differences, and pretest adjustment is required for valid estimation of desegregation effects.

Crain and Mahard (1982) "excluded a large number of papers, many of which compared students in racially segregated and racially mixed schools, but gave no indication that a formal desegregation plan had been adopted" (p. 6). Because they included studies that employed ability (in contrast to educational achievement) as a dependent variable and conducted a more recent and exhaustive search, they used 93 studies for analysis in contrast to Krol's 55 (see Tables 8 and 9).

Table 8

Effects of Desegregation on Black Achievement

in Three Syntheses

		Effect Sizes		
Source	Positive Results Percent	Mean	Standard Deviation	Comments
Krol (1978)	61	.16	.41	Based on 71 comparisons in 55 studies, grade level, mathematics and verbal achievement, and program-duration differences tested and found insignificant.
Crain & Mahard (1982)	62	.10	.25	Percent calculated as sum of 173 positive and half of 50 non-significant comparisons of 321 comparisons in 93 studies; effect-size mean based on 70 studies. With studies as units, significantly larger effects in kindergarten and grade one were found.
"Acceptable Studies"	64	.13	.24	Since the pretest advantage of desegregated groups over control groups was .18, results are calculated for 11 study-weighted means of posttests adjusted for pretests.

Table 9

Inferences from Three Syntheses

About the Effects of Desegregation on Black Achievement

	Percent-Positive Studies		Average Effect Sizes	
	Significance	Magnitude	Significance	Magnitude
	(.05)	(67%)	(.05)	(.20)
Krol (1978)	?	No	?	No
Crain & Mahard (1982)	?	No	Yes	No
"Acceptable Studies"	No	No	No	No
Conclusion	No?	No	?	No

Note--The criteria for inferences are as follows: The significance required is the standard .05 level calculated for a sign test for a 50-50 split for positive vote counts, and a T test for the difference of the mean effect size from zero, when possible, on independent units of analysis, that is, studies not comparisons. The magnitude criteria are 67 percent of the studies positive and an average effect size of .20, for which the desegregated students would exceed 58 percent of the control-group students.

The NIE panel employed a number of stringent criteria for study rejection including the following: non-empirical and summary reports; studies done outside the U.S. and geographically non-specific; those that combined or compared ethnic groups, lacked contemporaneous-control or pre-desegregation data, or analyzed heterogenously desegregated groups; those with more than 35 percent attrition, majority-Black desegregated conditions, varied exposure to desegregation, and non-comparable groups; those with unknown sampling procedures, cross-sectional data, or non-comparable samples at each observation point; those with unreliable or unstandardized instruments, unknown test content or instruments, unknown test administration dates, ability tests as dependent variables, and non-equivalent pretests and posttests; and insufficient statistics (Schneider, Note 1). Application of these exclusion criteria (Wortman, Note 2) resulted in 19 "acceptable studies."

Thus, all three data sets are similar in including only studies of Black achievement. They differ chiefly in that Krol and the NIE panel, unlike Crain and Mahard (1982), exclude ability tests, and the NIE panel employed stringent methodological criteria that resulted in a selection of studies only 19 percent as large as Crain and Mahard's set (see Table 8).

The NIE panel may be right in specifying stringent selection criteria from one viewpoint: the conclusions of review articles are usually based upon methodologically acceptable studies. But, as Glass, McGaw, and Smith (1982, p. 226) point out, excluding studies by implicit or explicit selection criteria can convert empirical questions of research methodology to a priori assumptions. Excluding studies without pretests, for example, may exclude randomized experiments--possibly the best design in certain respects for probing causality and avoiding untenable convariance assumptions.

If it were to be found that randomized posttest only designs yielded the same results as pretest-posttest quasi-experiments, then greater confidence could be placed in the results than the results of either design by themselves, since the two designs are subject to different threats to methodological validity (Cook & Campbell, 1979). Because, for example, the findings on instructional research are generally robust and consistent across study features, such as research methods and student characteristics, substantial confidence can be placed in their results.

Morevoer, excluding studies on policy or substantive criteria may be useful to lighten the effort or to narrow research questions, but exclusion also restricts the inferences and comparisons that can be made and the policies that may be implied. In the Krol and NIE selections, for example, it will not be possible to determine whether desegregation has a different impact on achievement than it does on ability or other educational outcomes such as creativity, critical thinking, interest in further learning, and social perceptiveness. In none of the three sets of studies, moreover, will it be possible to compare the effects of desegregation on Asian, Black, Hispanic, and White students. At least for some parents, educators, policy makers, researchers, and others, it would be useful to have reliable information on these and other points.

None of this is to argue that all studies should be summarized in one overall vote count or mean effect size. Although that statistic and its significance are of interest, characteristics of the studies such as Cook and Campbell's (1979) 33 threats to methodological validity, student characteristics such as ethnicity and grade level, and conditions of desegregation such as voluntary and mandatory plans, should be categorized, coded, and tested for statistical significance with studies as the units to afford independence as assumed in statistical inference. (If desegregation is working generally well according to a study, then students in different grades within the study are likely do well, and their performance is correlated and not statistically independent; similarly, if students are doing poorly in another study, different grades lack independence; therefore the means for studies, not for grade levels or other units, must be taken as the units for meta-analysis, or each comparison in a study must be weighted inversely to the number of comparisons in the study. Another reason for using study means or weighting is to insure that each study is given an equal weighting of one, not a weighting based on the arbitrary number of comparisons the investigator happened to make.)

Synthesis of Three Meta-analyses

Tables 8 and 9 show what can be validly extracted as the chief findings from the three meta-analyses. Table 8 shows that three estimates of percent-positive studies vary between 61 and 64 percent. These percentages are in surprisingly close agreement considering the widely different selection criteria and numbers of studies in the three syntheses.

Table 9 shows that the statistical significance cannot be determined in two cases because the percentage of positive comparisons rather than studies are reported; and, in the NIE case, the sign test based on the number of studies is insignificant. By the norms of recent syntheses of productivity factors discussed in previous sections, the percentage magnitudes are neither large (85 percent) nor average (67 percent). The statistical significance of the percentages cannot be determined in the two previous syntheses previously reported and is insignificant in the case of the NIE selection.

The statistical significance of the effect sizes are mixed: indeterminate for Krol, because of comparison weighting; significant for Crain and Mahard; and not significant for the set of studies acceptable to the NIE panel. In none of the three cases was the magnitude of the effect large (.45) or average (.20). (Crain and Mahard's significant finding of higher effects in kindergarten and first grade are unsupported by Krol and reversed in analyses by Wortman (Note 2); and their randomized-longitudinal effect is insignificant with study as the unit. Thus, their overall average study-weighted effect size is reported in Table 8.)

The results from the three meta-analyses suggest that the vote counts fail with some uncertainty to reach conventional levels of statistical significance. By normative standards of recent syntheses of other educational factors, they clearly fail with respect to percentage results. The effect sizes as a set are indeterminate with respect to significance and certainly fail to reach criterion levels with respect to normative magnitude.

Conclusion

New techniques of research syntheses show a number of potent factors for improving educational achievement that have proven to be consistently effective in a wide variety of experimental and educational conditions. These include the amount and quality of instruction, constructive classroom morale, and stimulation in the home environment. It is in our national economic, social, and political interest to implement these factors more deeply and widely for all children (Walberg, 1983). In this effort, school desegregation does not appear to prove promising in the size or consistency of its effects on learning of Black students.

Reference Notes

1. Schneider, J.M. Personal communications. August 16, 1982; November 4, 1982.

2. Wortman, P. Personal communications. August 28, 1982; November 10, 12, 1982.

References

Aitkin, M., Anderson, D., & Hinde, J. Modeling of data on teaching styles (with discussion). Journal of the Royal Statistical Society, Series A, 1981, 144, 419-461.

Aitkin, M., Bennett, S.N., & Hesketh, J. Teaching styles and pupil progress: A re-analysis. British Journal of Educational Psychology, 1983, 51, in press.

Bangert, R.L., Kulik, J.A., & Kulik, C.-L.C. Individualized systems of instruction in secondary schools. Ann Arbor: University of Michigan, manuscript, 1981.

Becker, W.C., & Gersten, R. A follow-up of Follow Through. American Educational Research Journal, 1982, 19, 75-92.

Bennett, S.N. Recent research on teaching: A dream, a belief, and a model. British Journal of Educational Psychology, 1978, 48, 127-47.

Bennett, S.N. Teaching styles and pupil progress. London: Open Books, 1976.

Blaug, M. Economic theory in retrospect. New York: W.W. Norton & Co., 1966.

Bloom, B.S. Human characteristics and school learning. New York: McGraw-Hill, 1976.

Bruner, J.S. Toward a theory of instruction. New York: W.W. Norton & Co., 1966.

Butcher, P.M. An experimental investigation of the effectiveness of a value claim strategy unit for use in teacher education. Sydney, Australia: Macquarie University, unpublished doctoral dissertation, 1981.

Cahen, L.S. Meta-analysis: A technique with promise and problems. Evaluation in Education, 1980, 4, 37-42.

Carlberg, C., & Kavale, K. The efficacy of special versus regular class placement for exceptional children: A meta-analysis. Journal of Special Education, 1980, 14, 295-309.

Carroll, J.B. A model of school learning. Teachers College Record, 1963, 64, 723-733.

Cohen, P.A. Effectiveness of student-rating feedback for improving college instruction. Research in Higher Education, 1980, 13, 321-341.

Cohen, P. A., Student ratings of instruction and student achievement. Review of Educational Research, 1981, 51, 281-309.

Cohen, P.A., Kulik, J.A., & Kulik, C.-L. C. Educational outcomes of tutoring. American Educational Research Journal, 1983, in press.

Cohen, P.A., Ebeling, B. J., Kulik, J.A. A meta-analysis of outcome studies of visual-based instruction. Education Communication and Technology Journal, 1981, 29, 26-36.

Colosimo, M.L. The effect of practice or beginning teaching on the self concepts and attitudes of teachers: A quantitative synthesis. Chicago: University of Chicago, unpublished doctoral dissertation, 1981.

Cook, T.D., & Campbell, D.T. Quasi-experimentation. Chicago: Rand-McNally, 1979.

Cooley, W.W., & Leinhardt, G. The application of a model for investigating classroom process. Pittsburgh: University of Pittsburgh Learning Research and Development Center, 1975.

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

Cooper, H.M., & Rosenthal, R. A comparison of statistical and traditional procedures for summarizing research. Evaluation in Education, 1980, 4, 33-36.

Crain, R.L., & Mahard, R.E. Desegregation plans that raise Black achievement: A review of the research. Santa Monica, Cal.: Rand Corporation, 1982.

Dunkin, M.J. Problems in the accumulation of process-product evidence in classroom research. British Journal of Teacher Education, 1976, 2, 175-187.

Finley, M.J., & Cooper, H.M. The relation between locus of control and academic achievement. Columbia, Missouri: University of Missouri Center for Research in Social Behavior, 1981.

Gagne, R.M. The conditions of learning. Chicago: Holt, Rinehart, & Winston, 1977.

Giaconia, R.M., & Hedges, L.V. Identifying features of open education. Stanford, Calif.: Stanford University, 1982.

Glaser, R. Components of a psychological theory of instruction: Toward a science of design. Review of Educational Research, 1976, 46, 1-24.

Glass, G. V. Intergrating findings: The meta-analysis of research. Review of Research in Education, 1977, 5, 351-379.

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

Graue, M.E., Weinstein, T., & Walberg, H.J. School-based home instruction and learning: A quantitative synthesis. Chicago: University of Illinois, Office of Evaluation Research, 1982.

Graubard, S.R. (Ed.), America's School: Portraits and Perspectives, Daedalus, 1981, 110, 1-175.

Green, J.L. Research on teaching as a linguistic process: A state of the art. Newark: University of Delaware, 1982.

Hanford, B.C., & Hattie, J.A. The relationship between self and achievement/performance measures. Review of Educational Research, 1982, 52, 123-142.

Harnischfeger, A., & Wiley, D.E. The teaching-learning process in elementary schools: A synoptic view. Curriculum Inquiry, 1976, 6, 5-43.

Haertel, G.D., Walberg, H.J., & Weinstein, T. Psychological models of educational performance: A theoretical synthesis of constructs. Review of Educational Research, 1983, in press.

Hedges, L.V., Giaconia, R.M., & Gage, N.L. Meta-analysis of the effects of open and traditional instruction. Stanford, Calif.: Stanford University Program on Teaching Effectiveness, 1981.

Horwitz, R.A. Psychological effects of the open classroom. Review of Educational Research, 1979, 49, 71-86.

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

Johnson, D.W., Maruyama, G., Johnson, R., Nelson, D., & Skon, L. Effects of cooperative, competitive, and individualistic goal structures on achievement: A meta-analysis. Psychological Bulletin, 1981, 89, 47-62.

Krol, R.A. A meta-analysis of comparative research on the effects of desegregation on academic achievement. Unpublished doctoral dissertation, Western Michigan University, 1978.

Kulik, C. - L. C., & Kulik, J.A. Effects of ability grouping on secondary school students. Ann Arbor: University of Michigan, manuscript, 1981.

Kulik, C. -L. C., Shwalb, B.J., Kulik, J.A. Programmed instruction in secondary education. Journal of Educational Research, in press.

Kulik, J.A., Cohen, P.A., & Ebeling, B.J. Effectiveness of programmed instruction in higher education. Educational Evaluation and Policy Analysis, 1980, 2, 51-64.

Kulik, J.A., Kulik, C.-L. C., & Cohen, P.A. Research on audio-tutorial instruction. Research in Higher Education, 1979b, 11, 321-341.

Kulik, J.A., Kulik, C.-L. C., & Cohen, P.A. A meta-analysis of outcome studies of Keller's Personalized System of Instruction. American Psychologist, 1979c, 34, 307-318.

Kulik, J.A., Kulik, C.-L. C. & Cohen, P.A. Effectiveness of computer-based college teaching. Review of Educational Research, 1980, 50, 525-544.

Lecompte, M.D., & Goetz, J.P. Problems of reliability and validity in ethnographic research. Review of Educational Research, 1982, 52, 31-60.

Light, R.J., & Pillemer, D.B. Numbers and narrative: Combining their strengths in research reviews. Harvard Educational Review, 1982, 52, 1-26.

Luiten, J., Ames, W., & Ackerson, G. A meta-analysis of advance organizers on learning and retention. American Educational Research Journal, 1980, 17, 211-218.

Lysakowski, R.S., & Walberg, H.J. Cues, participation, and feedback in instruction: A quantitative synthesis. American Educational Research Journal, 1983, in press.

Ottenbacher, K., & Cooper, H. The effect of class placement on the social adjustment of mentally retarded children. Columbia: University of Missouri Center for Research in Social Behavior, 1981.

Peterson, P.L. Direct instruction reconsidered. In P.L. Peterson & H. J. Walberg (Eds.), Research on teaching. Berkeley, Calif.: McCutchan, 1979.

Pflaum, S.W., Walberg, H.J., Karegianes, M.L., & Rasher, S. Reading instruction: A quantitative synthesis. Educational Researcher, 1980, 9, 12-18.

Popper, K.R. The logic of scientific discovery. New York: Basic Books 1959.

Redfield, D.L., & Rousseau, E.W. A meta-analysis of experimental research on teacher questioning behavior. Review of Educational Research, 1981, 51, 237-245.

Rosenthal, R. Combining probabilities and the file drawer problem. Evaluation in education, 1980, 4, 18-21.

Samson, G., Graue, M.E., Weinstein, T., & Walberg, H.J. Academic and occupational performance: A quantitative synthesis. Chicago: University of Illinois Office of Evaluation Research, 1982.

Shulman, L.S., & Tamir, P. Research on teaching in the natural sciences. In R. M. W. Travers (Ed.), Handbook of research on teaching, Second Edition. Chicago: Rand-McNally, 1973.

Slavin, R.E. Cooperative learning. Review of Educational Research, 1980, 50, 315-342.

Smith, M.L. Publication bias and meta-analysis. Evaluation in Education, 1980, 4, 22-24.

Smith, M.L., & Glass, G.V. Meta-analysis of research on class size and its relationship to attitudes. American Education Research Journal, 1980, 17, 419-433.

Walberg, H.J. A psychological theory of educational productivity. In F.H. Farley & N. Gordon (Eds.), Psychology and Education. Berkeley, Calif.: McCutchan, 1980.

Walberg, H.J. Education, scientific literacy, and economic productivity. Daedalus, 1983, in press.

Walberg, H.J. What makes schooling effective? Contemporary Education Review, 1982, 1, 1-34.

Walberg H.J., & Haertel, E.H. (Eds.) Research Synthesis: The State of the Art, Evaluation in Education, 1980, 4, 1-142.

Walberg, H.J., & Genova, W.G. School practices and climates that promote integration. Contemporary Educational Psychology, 1983, in press.

Walberg, H.J., Pascarella, E., Haertel, G.D., Junker, L.K., & Boulanger, F.D. Probing a model of educational productivity with national assessment samples of older adolescents. Journal of Educational Psychology, 1982, 74, 295-307.

Walberg, H.J., Schiller, D., & Haertel, G.D. The quiet revolution in educational research. Phi Delta Kappan, 1979, 61 (3), 179-182.

Waxman, H.C., & Walberg, H.J. The relation of teaching and learning. Contemporary Education Review, 1982, 2, 103-120.

Waller, W. The sociology of teaching. New York: Longman's, 1932.

Williams, P.A., Haertel, E.H., Haertel, G.D., & Walberg, H.J. The impact of leisure-time television on school learning. American Educational Research Journal, 1982, 19, 19-50.

Wilkinson. S.S. The relationship of teacher praise and student achievement: A meta-analysis. Gainesville: University of Florida, unpublished doctoral dissertation, 1980.

Willson, V.L., & Putnam, R.R. A meta-analysis of pretest sensitization effects in experimental design. American Educational Research Journal, 1982, 19, 249-258.

School Desegregation and Black Achievement: An Integrative View

Paul M. Wortman
University of Michigan

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 and 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 64 (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 and 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

Table 1

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

Definition	Advantages	Disadvantages
Meta-analysis Method 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)

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 and 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 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 and 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 prior 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 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). 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 conclusions are presented in the last two sections.

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 ith study and $\underline{s}$ is the control group standard deviation.

The average effect size, , 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 the 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 and Erlebacher, 1967; Campbell and Boruch, 1975; Campbell and Stanley, 1966; Cook and 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, and 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 and 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 and Erlebacher, 1970; Cook and 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:

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

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

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 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 F (see Appendix 7 in Smith, Glass, and 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:

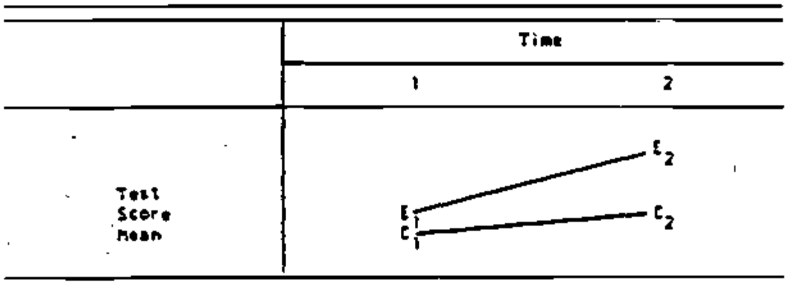
$$gain = E_2 - E_1 \text{ and } C_2 - C_1,$$

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 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 Stanley's terminology, the One-Group Pretest-Posttest (OGPP) Design. This is the NECGO 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.

Figure 1

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



E=Experimental Group
C=Control Group

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 (cf. Gehan and Freireich, 1974). In education 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 since these data are not obtained concurrently with the experimental (i.e., desegregation) data. Again extreme care is needed in interpreting these data.

Sometimes it is possible to create a cohort of students who are followed prior to the start of the program. This allows a "dry run" NECGD experiment (where there is no treatment) to be created and an estimate of the adequacy of the various adjustment procedures to be obtained (Wortman, Reichardt, and St. Pierre, 1978). Such data are rarely available, though. If repeated classes show similar effects, however, then the data are probably reliable. This variant of the "Recurrent Institutional Cycle Design" is sometimes used (cf. Teele, 1973). In general, historical controls have been found to grossly overestimate effects and thus should not be used if possible (Sacks et al., 1982). In education, for example, test scores were declining during the 1960's and 1970's so that historical controls would probably have higher scores. Such studies were not included in our analyses, but comprised 17 percent of the studies in Crain and Mahard's (1982) meta-analysis. More recently, Crain (1983) has included eight such studies among his "20 best."

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 and 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 and 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 and Gomez, 1977; Sechrest and 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 h been much discussion of this issue (Mansfield and 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.

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 and 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 and 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 and Stanley, 1963; Cook and 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

Table 2

Criteria for Selecting Studies for Meta-analysis

Criteria for Rejection	Threats to Validity			
	Internal	External	Construct	Statistical
1) <u>Type of Study</u>				
*a) Non empirical			X	
*b) Summary report insufficient detail for coding			X	
2) <u>Location</u>				
*a) Outside U S A		X		
*b) Geographically non-specific		X		
3) <u>Comparisons</u>				
*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†			X	
*j) Varied exposure to desegregation†	X			
k) Groups initially non comparable	X			
4) <u>Study Design</u>				
*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) <u>Measures</u>				
*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) <u>Data analysis</u>				
*a) No pretest means				X
*b) No posttest means				X
*c) No pretest standard deviations‡				X
*d) No posttest standard deviations‡				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 NIE Core Studies

†For the NIE Core Studies these criteria were relaxed to allow studies that provided "specific justification" for this

‡For the NIE Core Studies these criteria were combined into a single criterion, unable to calculate effect sizes

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" problems stemming from differential test reliability and low inter-test realiability. 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, studics 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 characteristic (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 and 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 above has 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.[3] 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 authors 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.

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.

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

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: O X O		Nonequivalent Control Group: O X O - - O - - O - - -		Static Group Comparison: X O - - - - - - O - - -	
	No. of Obs[a]	Mean ES x̄ (s)	No. of Obs	Mean ES x̄ (s)	No. of Obs	Mean ES x̄ (s)	No. of Obs.	Mean ES x̄ (s)
1-6	74	0.43 (0.55)	8	1.75 (2.73)	45	0.28 (0.19)	16	0.24 (0.22)
7-9	11	*1.05(1.11)	4	1.49 (0.20)	4	*0.91(1.11)	3	-0.03 (0.23)
10-12	11	0.09 (0.04)	6	*0.01(0.05)	4	0.17 (0.01)	1	-0.18
GRAND	98	0.45[b](0.68)	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(2,51)=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	11	0.17 (0.11)	1	0.01	5	0.09 (0.07)	2	0.08 (0.29)
3	8	0.22 (0.71)	1	2.15	5	0.28 (0.25)	0	- -
4	11	0.44 (0.54)	2	2.03 (1.20)	9	0.33 (0.10)	6	-0.03 (0.07)
5	27	0.51 (0.89)	3	1.54 (2.27)	15	0.38 (0.11)	3	0.17 (0.00)
6	15	0.55 (0.85)	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	0.07 (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 (x̄ = 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 Scheffé test (overall F = 11.47, df=2,91, p < .0001).

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 these 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 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.

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

<u>Mean Effect-Size for Math Vs. Reading Achievement Measures</u>

Achievement Measure	Mean Glass ES & (c^2)	<u>F</u>
Math (n=37)	0.33 (0.38)	1.86, df=1.87, $\underline{p} < .18$
Reading (n=51)	0.57 (0.94)	

<u>Note</u>--Krol found a tendency for math achievement to show a greater effect-size than reading achievement ($\underline{t}_{16}=1.90$, 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 reeults 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	NA	NA	2	4	.63	--	.95
		NA	NA	2	4	--	.39	.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	--	.95	.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	2	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	.31	--	--
		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 90 90 90 90 90	5 5 5 5 5 5	3 3 4 4 5 5	5 5 6 6 7 7	1.14 -- 1.27 -- 2.11 --	-- .95 -- .92 -- 1.40	.49 .06 .58 -.17 .78 -.22
Savage (1971)	2	100 100	NA NA	9 9	11 11	.01 --	-- .17	.14 -.09
Sheehan (1979)	2	98 98	30 30	4 4	5 5	-.29 --	-- -.21	-.18 -.16
Slone (1968)	2	60 60	NA NA	4 4	5 5	.42 --	-- .49	-- --
Smith (1974)	2	100 100	42 42	6 6	9 9	-.22 --	-- .42	-.05 .10
Syracuse School District (1979)	1	89	10	4	4	.15	--	--
Thompson & [illegible] (1979)	2	42 42	5 5	3 2	5 5	-.33 --	-- .10	-- --
Van Every (1969)	6	95 95 95 95 95 95	20 20 20 20 20 20	4 4 4 4 4 4	5 5 5 6 6 6	.18 -- -- -.25 -- --	-- .28 -- -- .36 --	.59 .11 -- -.44 .93 --
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 -.05	.11 -.24 .21 -.01
Zdep (1971)	2	NA NA	12 12	2 2	2 2	.31 --	-- -.19	.65 -.15
= 19	= 62							

BEST COPY AVAILABLE

Name of Study	# of Cases	% Black		Grade Level		Achievement Effect Size		Pretest-Adjusted Effect Size
		Seg	Deseg.	Pretest	Posttest	Reading	Math	
OVERALL MEAN[1]	(N= 62)	82.49	15.03	4.09	5.12	.28	.23	.14
MEAN FOR TREATMENTS LASTING ONE YEAR OR LESS[1]	(N= 20)	71.00	11.58	3.69	4.20	.30	.11	.13
MEAN FOR TREATMENTS LASTING MORE THAN ONE YEAR[1]	(N= 14)	95.31	17.50	4.00	5.81	.28	.39	.12

Note: *NA = Not Ascertainable

[1]Mean effect sizes, weighted by study

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 and 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 desegregation 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 and 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 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 and Bridgeman, 1979; Slavin and 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 desegregated 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 desegregation 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. Moreever, 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.

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.

REFERENCES

Aronson, E., and 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., and Gomez, H. Sensitivity, bias, and theory in impact evaluations. Professional Psychology, 1977, 8, 411-434.

Bradley, L. A., and 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., and 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., and 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 and A. A. Lumsdaine (Eds.), Evaluation and Experiment: Some Critical Issues in Assessing Social Programs. New York: Academic Press, 1975.

Campbell, D. T., and 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., and 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., and York, R. L. Equality of educational opportunity. Washington, D. C.: U. S. Government Printing Office, 1966.

Cook, T. D., and 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., and Mahard, R. E. Desegregation and Black achievement: A review of the research. Law and Contemporary Problems. 1978, 42, 17-56.

Crain, R. L., and 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., and Freireich, E. J. Non-randomized controls in cancer clinical trials. The New England Journal of Medicine, 1974, 290, 198-203.

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

Gilbert, J. P., McPeek, B., and Mosteller, F. Progress in surgery and anesthesia: Benefits and risks of innovative therapy. In J. P. Bunker, B. A. Barnes, and 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., and Smith, M. L. Meta-analysis in social research. Beverly Hills, CA: Sage Publications, 1981.

Glass, G. V., and 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., and Dawes, R. M. Psychotherapy outcome: Smith and Glass conclusions stand up under scrutiny. American Psychologist, 1982. 37, 504-516.

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

Linsenmeier, J. A. W., and 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., and Busse, T. V. Meta-analysis of research: A rejoinder to Glass. Educational Research, 1979, 6, 3.

Moskowitz, J. M. and Wortman, P. M. A secondary analysis of the Riverside School Study of desegregation. In R. F. Boruch, P. M. Wortman, and 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., and Smith, H. Randomized versus historical controls for clinical trials. American Journal of Medicine, 1982, 72, 233-240.

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

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

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

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

Staines, G. L. The strategic combination argument. In W. Leinfeliner and 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 and 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 busing 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., and Bryant, F.B. Meta-analysis of quasi-experiments: School desegregation and Black achievement. Part 1-Retrieval and coding. Ann Arbor, MI: Institute for Social Research, 1982.

Wortman, P. M., Reichardt, C. S., and 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., and Yeaton, W. H. Synthesis of results in controlled trials of coronary artery bypass surgery. In R. J. Light (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 11 -- 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., and 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., and 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. 323-326.

*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.

*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.