DOCUMENT RESUME

ED 238 999 UD 023 302

AUTHOR Crain, Robert L.
TITLE Dilemmas in Meta-Analysis: A Reply to Reanalyses of the Desegregation-Achievement Synthesis.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE 83
NOTE 35p.; For related documents, see UD 023 303-308. Paper submitted as one of a collection from the National Institute of Education Panel on the Effects of School Desegregation. Document may not reproduce well.
PUB TYPE Information Analyses (070) -- Viewpoints (120)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Gains; *Black Students; *Desegregation Effects; Elementary Secondary Education; *Evaluation Criteria; Grade 1; Kindergarten; *Meta Analysis; Outcomes of Education; Program Effectiveness; Program Evaluation; *Research Methodology; Research Reports; School Desegregation; Validity

ABSTRACT

The decision by the National Institute of Education panel on the effects of school desegregation to select (for meta-analysis) a small group of preferred studies based upon criteria chosen in advance of examining the studies was, in principle, a mistake. One usually cannot know until the data have been examined which of several competing methodological criteria are most important. In the case of the effects of desegregation on minority achievement, Crain and Mahard in their 1982 review of 93 desegregation studies found a methodological error so specific to desegregation research that it was not even recognized as an error until the review was done. The error was that studies of the effects of desegregation on minority achievement will underestimate any effects when using subjects who have not been in desegregated settings since kindergarten or Grade 1. Whereas Crain and Mahard found 20 studies of blacks in desegregated settings since kindergarten or Grade 1, the panel discarded all but one of them because they did not fit their chosen-in-advance criteria. Of the 20 studies identified by Crain and Mahard, 16 showed consistent positive outcomes and only 2 were negative. If the principal function of selecting a superior subgroup of studies is to find the consistency of results which is masked by error in an unselected sample, Crain and Mahard succeeded, and the panel did not. (CMG)

Dilemmas in Meta-Analysis: A Reply to Reanalyses of the Desegregation-Achievement Synthesis.

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 sicentists 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 science, 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

Submitted as one of the
papers from the National
Institute of Education
Panel on the Effects of
School Desegregation

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,[5] 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 scores 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 conslusions 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 scientists to dissagree about methodology.[6] 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 no 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 unquestionalby 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 or should 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 particular 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 anlysis. 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 the list and them 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 unequivocably 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. 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

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*	--

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 do 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) researchers. 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 that 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 based on official reports are likely to be reasonably accurate and one might choose to ignore those studies based on self-reported 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 which 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. Only 18%--two studies--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

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

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 metholodogical 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.

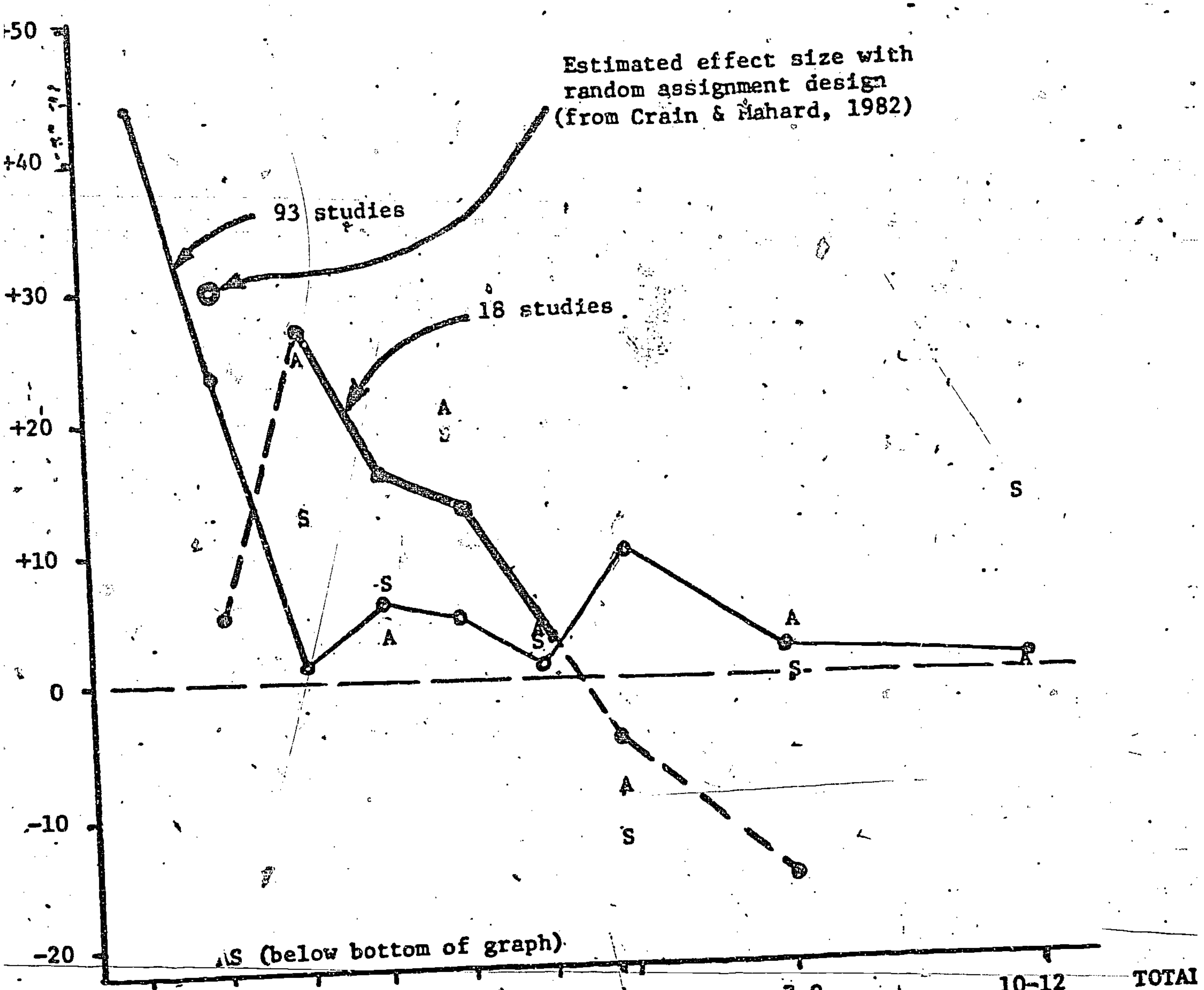
Table 3: Direction and Size of Treatment Effect, by Type of Control Group

Design	direction of effect				effect size	
	+	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)

On the other hand studies which compare desegregated blacks to segregated blacks--either in a "cohort" design (the segregated blacks are the students in that 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.

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 graphs 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



	KG	1	2	3	4	5	6	7-9	10-12	TOTAI
8 studies	(0)	(2)	(5)	(7)	(8)	(4)	(1)	(1)	(0)	(28)
3 studies	(10)	(40)	(27)	(39)	(24)	(29)	(20)	(21)	(19)	(228)
rmor	(0)	(1)	(5)	(6)	(6)	(7)	(2)	(1)	(2)	
tephan	(0)	(1)	(5)	(6)	(6)	(7)	(2)	(1)	(2)	

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

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 this estimate. Unfortunately, given the relative 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 correlations within 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.

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

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

	Direction of Effect				Effect Size	
	+	0	-	(n)	d	(n)
grade at desegregation:						
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)

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:

Stanley Zdep (1971) of ETS carried out an evaluation of a city-to-suburban voluntary transfer plan from Newark, NJ to a 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 bussed students were 9.8 answers higher than the control group on a test on which the bussed 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 bussed 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 experimenzal 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 that time, and conducted his 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 kindergarten and an average gain of 2.6 points for those entering 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 for students who entered the program in grades 2 and 3 and negative treatment effects for stud [illegible] 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 [illegible]sed 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 [illegible] years, which presumably would bias the control group upward. If there [illegible] 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 ears 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 first 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 the control group in reading but did score 6 months ahead of the control group in math. In that year, the treatment group was slightly 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 notorious 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 a 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 that 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 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 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 deviations 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 ETS (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 having relatively large sample sizes. They are also not likely to be affected 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 exceptions 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

*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.

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 Danahy (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 trichotomize the sample before comparing posttest means within each group. Elmer Lemke (1979), studying Peoria, Illinois, studied 180 desegregated and 60 segregated black students five years after desegregation began. He used the Metropolitan Readiness Test and 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 the 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

*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.

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 methodolgically, 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.

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)

Arrow, Kenneth J
1951

Social Choice and Individual Values
New York: Wiley

Bowman, O.H.
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. School Dist.
1975

Desegregation Report. Las Vegas, NV: Author

Crain, Robert L. and 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 choices of Negro inner city children." Ph.D. dissertation, University of Minnesota (University Microfilms No. 72-14285).

Frary, R. B., and T.M. Goolsby Jr.
1970

"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.

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 Illinois - 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., J.E.Lang, H.R. Goldberg and L. W. Heinrich
1968
A Report on a Cooperative Program Between a City School District and a Suburban School Dsitrict, Rochester, NY City School District,

Samuels, J. M.
1971
"A comparison of projects representative of compensatory. busing, and non-compensatory programs for inner-city students." Ph.D dissertation, University of Connecticut (University Microfilms No. 72-14252).

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 — "School 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