Institute for Policy Research
Northwestern University
Working Paper Series

WP-05-04

How Large an Effect Can We Expect from School Reforms?

Spyros Konstantopoulos
Faculty Associate, Institute for Policy Research
Assistant Professor of Human Development, Social Policy, and Learning Sciences
Northwestern University

Larry V. Hedges
Faculty Fellow, Institute for Policy Research
Board of Trustees Professor of Statistics and Social Policy
Northwestern University

DRAFT
Please do not quote or distribute without permission.

2040 Sheridan Rd. ◆ Evanston, IL 60208-4100 ◆ Tel: 847-491-3395 ◆ Fax: 847-491-9916
www.northwestern.edu/ipr ◆ ipr@northwestern.edu

Abstract

Judging the success of school reform requires an interpretative context in which to judge whether effects obtained are large enough to be important or so small as to be a disappointment. The logic of school reform suggests two frameworks with which to judge the importance of effects. One is the size of the existing achievement gaps between important groups in society. The other is the size of gaps between mean achievement among schools (adjusted for student characteristics). NAEP data is used to demonstrate that in national data, gaps which appear large by one standard may appear small by the other. We argue that the most appropriate framework for judging reform effects is the national distribution of school effects.

One of the goals of school reform in the United States is to modify schools so that all students will receive high quality instruction based on a challenging curriculum that will result in high levels of academic achievement for all students. The urgency with which this reform goal will be pursued has been increased with the passage of the No Child Left Behind (NCLB) Act, which provides for incentives and penalties for progress (or lack of it) toward these goals. While there are many desirable outcomes of schooling, such as social responsibility, good character, and other attributes of good citizenship, the No Child Left Behind Act focuses specifically on academic achievement. There are many ways of measuring academic achievement, including work samples portfolios, performance assessments, and other authentic assessments, as well as paper and pencil tests. However, the No Child Left Behind Act privileges academic achievement as measured by the National Assessment of Educational Progress (NAEP) or other assessments that can be benchmarked by NAEP.

Thus it appears that the immediate goals of school reform in America will be to make all schools perform well in generating academic achievement as measured by assessments like NAEP. While the philosophy of school reform is often articulated (in NCLB and elsewhere) in terms of standards (a criterion referenced approach), the effects of reforms are often measured via norm referenced tests such as NAEP, which, despite the imposition of achievement levels as aids in reporting performance, was constructed as a norm referenced test.

While there is relatively little disagreement that this goal will drive reform, there is not a consensus on how to achieve it. Determining the effectiveness of reform strategies is a major part of the educational research agenda for the next decade. Effects of education reforms will be evaluated quantitatively, and an important aspect of this work will be judging *how well* reform strategies work. That is, which reforms produce large improvements in achievement, which produce modest improvements in achievement, and how large those improvements are likely to be.

It is important to distinguish between providing a statistical estimate of the effect size associated with a particular reform and *judging* whether that effect is big enough to be important or so small as to be a disappointment. Estimation of effect size can be accomplished by purely technical means. There may be technical problems in arriving at such an estimate including problems of study design or analysis, but the computation of the effect size estimate is a purely technical procedure. In contrast, judging or evaluating whether the effect is large enough to be important is an interpretive act. This judgment requires a context in which to frame the interpretation: large or small *compared to what*?

Judging the Effectiveness of School Reforms

The rhetoric of contemporary school reform suggests two somewhat different solutions to the problem of the interpretive frame. One solution is derived from the idea that the goal of school reform is to reduce, or better,

eliminate, the achievement gaps between Black and White, rich and poor, and males and females. Consequently it is natural to evaluate reform effects by comparing them to the size of the gaps they are intended to ameliorate. For example, if the (average) achievement gap between Black and White students is one standard deviation of the national achievement distribution, and if school reforms are intended to eliminate this gap, then a reform that would only increase achievement by one tenth of a standard deviation might seem too weak to be important, while a reform that could increase achievement by three quarters of a standard deviation might seem quite important.

The second solution to the problem of interpreting the effects of reforms is derived from the idea that school reforms are intended to make all schools perform as well as the best schools. If so, then it is natural to evaluate reform effects by comparing them to the differences (gaps) in the achievement among schools in America. For example, if the reform is intended to make all schools perform as well as the best schools, then we can evaluate the size of a reform effect by comparing it to the gap between below average schools (say a school at the 25th percentile of all American schools) and an excellent school (say a school at the 80th percentile of all American schools). This interpretative context is explicitly normative, comparing reform effects with the normative distribution of school effects. For example, an effect that would move a median (50th percentile) school only to the 55th percentile of all schools might seem too small to be important, while a reform effect that would move a median school to the 90th

percentile of all schools might be taken to be a large effect.

The purpose of this paper is to explore these two alternative frameworks for interpreting the effects of reforms. We focus on these two frameworks because each is natural in some genres of evaluation. For example, small scale intervention research in the experimental or quasi-experimental tradition is likely to focus on interpretation of effects in terms of student variation. Larger scale school effects research in the tradition of mathematical sociology is more likely to focus on comparisons among schools. While it is conceivable and perhaps desirable to combine the two perspectives, our experience is that one perspective often overshadows the other. In any even insight about each is helpful even if the two frameworks are used together in interpretation.

Our purpose is to gain insight about the implications of each for the frameworks for interpreting the effects of school reform. We proceed by examining empirical evidence from NAEP about the implications of these two frameworks for judging the effects of school reforms. We argue that these two frameworks are likely to lead to different judgments about whether the effects of reforms are large enough to be important. We argue that one of the two interpretive frameworks is more appropriate than the other for interpreting the likely magnitude of school effects. Finally, we hope to shed some light on an important scientific and policy question:

How large an effect of educational reforms on school achievement is it reasonable to

expect, given what we already know about the distribution of achievement in America?

It is important to answer this question for two reasons. First, it is necessary to have an idea of what to expect in order to interpret findings of studies of the effects of school reform. This is the major topic of this paper. Second, research design requires some knowledge of the plausible size of the effects that a successful reform might produce. While optimism is a virtue among those interested in promulgating social reform, realism is a virtue in research design. Many areas of social program research have been plagued by evaluation studies that did not have sufficient statistical power to detect modest but meaningful effects even if they were present (see, e.g., Boruch and Gomez, 1977). Failure to correctly forecast the magnitude of effects that might be obtained in an evaluation can lead to a design that is insufficiently sensitive (has low statistical power), and therefore may fail to detect as statistically significant, program effects that are actually occurring. The issue is not as simple as a generic exhortation to conduct more powerful evaluation studies. In general, obtaining high statistical power costs money (for larger sample sizes or more sophisticated designs and analyses). Conducting adequate studies involving a substantial commitment of resources may necessitate tradeoffs between competing goods (such as more program implementation versus more evaluation). Wise and responsible policy formation in such situations requires a stronger justification than a generic exhortation that "more [money for

evaluation] is always better." Defensible numerical values of the magnitude of treatment effects likely to be obtained are essential to adequately inform research design.

Which Framework is More Appropriate for Interpreting Effects of School Reform?

It is tempting to judge the success or failure of reform efforts in terms of the problem they are meant to solve: achievement gaps between important societal groups. It is appropriate to set goals on the basis of significant societal problems. However identifying a problem and setting a goal of eliminating it does not mean that attaining the goal is feasible in the short term. For example, consider the noble aims of curing cancer, stopping heart disease, arriving at a population that is free of disease. Very significant amounts of resources have been allocated to these goals for decades and, while there has been progress, they are still far from being attained. Most would argue that it is not appropriate to measure the success of the war on cancer by simply asking if cancer is nearly eliminated as a cause of death in America.

Lofty goals have often been set in education as well, like reforming mathematics education in order to assure that American students were as good at mathematics as Soviet students and being first in the world on international comparisons of educational achievement by the year 2000. These goals have also often proven unattainable.

Goals to solve societal problems have many positive functions. They may energize and inspire research communities thereby directing the focus of their actions on a particular research agenda. They may help mobilize support and therefore garner resources for research or interventions. But to serve these ends, the goals themselves must be for substantial changes. If there are no sanctions attached to them, it is not critical that they be realistically attainable and they may be more functional as inspiration if they are not. For example, the goal of curing cancer looks even less attainable now, after decades of research and billions of dollars of research expenditures, than it once did. Yet it still inspires a huge biomedical research community and a National Institute of Health.

The use of normative criteria to interpret the effects of reform is inherently realistic, in the sense that the criteria are developed from actual examples of what is not only possible in the real world, but has actually occurred. For example, if we know that some nontrivial fraction of schools function in a certain way (e.g., produce achievement gains of a certain size), then we know that it is at least *possible* for schools to function that way. In contrast, goals set in the abstract may not be realistic in the sense that no schools have ever been shown to function in ways that meet the goal.

Moreover, we argue that the distribution of observed school effects is a useful gauge to what is *not* possible, or at least has not been done. If virtually no school produces effects of a certain size, then it may be unrealistic (at least in the short run) to expect reforms to reliably create schools that produce effects that

large. Of course it is always possible to so radically change education that new possibilities are created. We should strive to do so. But to *require* such radical change as a criterion of success probably dooms educational reform to failure.

While there is naturally great optimism among proponents of reform about the magnitude of effects reforms might obtain, past experience in education and other empirical sciences (such as medicine) suggests that even treatments eventually understood to be effective may not produce effects that appear to be large without an appropriate interpretive context.

School Effects Models

School effects models can best be described in terms of a hierarchy with two levels (see Bryk and Raudenbush, 1992; Raudenbush and Bryk, 1987). The first level is a within-school *achievement model* that describes the academic achievement of students within a school as a function of the particular school (the school effect) and individual characteristics of the students, such as socio-economic status (SES), gender, race/ethnicity, etc. Thus the achievement model includes a specific term, a school effect, that describes how the average achievement of students in each particular school differs from that of other schools, controlling for the student characteristics in the achievement model. The achievement model usually includes parameters that describe the relation between individual student characteristics (such as SES, gender or race/ethnicity) and achievement *in that specific school*. The parameters in the

school effects model, the school effect (and the effects of student characteristics) on achievement, may vary across schools.

The second level, *the between-school model,* describes the variation across schools of the school effects in the achievement model. Since school effects describe the difference between each school's average achievement and that of the average school (that is they are centered at 0), the average of all school effects is zero. Thus the distribution of school effects is often described by a numerical estimate of a variation (called the between-school variance component). Sometimes additional factors (such as school resources or context) are used in the between-school model to explain variation in school effects.

In this paper we use two different achievement models. The first model simply treats all variation within the school as random.[1] It is used to describe how much of the national variation in achievement is between (across) schools and how much is within schools. Obviously interventions that impact school mean achievement only affect *between-school* variation and, by definition, do not affect the part of variation that is within schools. The second achievement model we employ includes the student characteristics of family SES, gender, and race/ethnicity. [2] Race/ethnicity was characterized by dividing the population into four groups used by NAEP: White, Black, Hispanic, and Other.

In this paper we use essentially the same between-school model in all analyses, which simply represents the variation of effects across schools as random. In the case of the first achievement model discussed above, we measure

how much average achievement varies across schools by the standard deviation of the school average achievement. In the case of the second within-school achievement model, there are six parameters in the achievement model for each school (an average and effects of SES, gender, achievement gaps between White and Black, White and Hispanic, and White and Other). The standard deviation across schools of these parameters provides measures of how much the each effect varies across schools. The computer program HLM and the NAEP sampling weights was used for all analyses.

The National Assessment of Educational Progress

The National Assessment of Educational Progress (NAEP), the Nation's Report Card, is the most important source of information about the academic achievement of our nation's children (Mullis, 1990). Since its inception it has served two important functions (Beaton and Zwick, 1992). First, it has made it possible to compare the academic achievement of population groups (such as regional, racial, or ethnic groups) at any one point in time. Second, it has made it possible to compare the achievement of the nation and population groups over time via its trend sample program. Although other cross sectional surveys have sporadically provided data on representative samples of our nation's children, no other survey has collected achievement data of the same high quality as NAEP and none has done so in a consistent fashion over time in a manner that permits the trend comparisons (with tests that are equated over time) that are possible in

NAEP (Johnson, 1992). Moreover, few other surveys have collected achievement data on pre high school students, making NAEP virtually the only source of information on the achievement of elementary and middle-school or junior high school students.

NAEP has collected achievement data on nationally representative samples of 9, 13, and 17 year-olds in reading since 1971 and mathematics since 1978 as part of its long term trend program. They have kept the instrumentation and the sampling and data collection procedures the same throughout the life of the long term trend program and the scales on which tests are reported have been equated. NAEP also collects data on students' family background, gender, and race/ethnicity. The family background data includes the education level of the parents, family structure, and things found in the home that are indicators of socio-economic status (at least 25 books, newspapers of magazines, and encyclopedia, a computer, etc.)[4]. The NAEP design permits direct estimation of the structure of relationships among background variables and student achievement that are not compromised by the relatively small amount of information obtained from each student assessed. [5] We used the reading and mathematics achievement data from the NAEP long term trend program to estimate school effects reported in this paper.

Findings from Analyses of NAEP

In separate sections below, we consider three issues using our analyses of

the NAEP data. The first issue we consider is how much of the variation in achievement is within schools and how large the between-schools variation is in comparison. The second issue we consider is how the variation in achievement between-schools changes when the effects of student SES, gender, and race/ethnicity are taken into account. In both cases, we examine the trend over time in the distribution of achievement and school effects. The third issue we examine is the implications of the national findings for the likely effects of school reform interventions. Specifically, we show how the distribution of naturally occurring school effects can provide a normative context for school effects that may arise as a consequence of school reform efforts.

How Large is Between-School Variation in Achievement

Table 1 provides information on NAEP reading achievement for the twenty-five years from 1971 to 1996. The table is organized into three panels vertically, with information for age 17 at the top, information for age 13 in the middle, and information for age 9 at the bottom. Within each panel, the top row shows the overall national standard deviation in reading achievement. The second row of each panel gives the estimate of the standard deviation of *school mean* achievement for the same years and the standard deviation of school mean achievement as a percentage of the total standard deviation of national student achievement distribution. The total variation is the sum of between-school and within-school components. Therefore if the between-school variation is less than half of the total variation, most of the variation is within schools. This analysis

reveals one important fact immediately:

> *most of the achievement variation in America is within schools, not between schools*.

The between-school standard deviation ranges from 22% to 48% as large as the national standard deviation of student achievement. This means that even *relatively* large between-school differences may be small in comparison to within school differences in achievement.

The dispersion of reading achievement at age 17 seems to have decreased slightly over the 25 years considered here (from a standard deviation of 45.8 in 1971 to 42.3 in 1996), but the standard deviation of school mean reading achievement has increased over that time (from a standard deviation of 14.9 in 1971 to 16.9 in 1996). As a result of these two trends, between-school variation in reading achievement at age 17 has increased as a fraction of total variation (from 32.6% in 1971 to 40.0% in 1996). The same general trend of between school variation increasing relative to the total also appears to be occurring at ages 9 and 13. Thus schools have become more diverse, more segregated by reading achievement over this time period.

Insert Table 1 About Here

Table 2 provides information on NAEP mathematics achievement for four years between 1978 and 1996, organized in the same way as in Table 1. In all but one case, the between schools achievement variation in mathematics is less than half of the overall national standard deviation. The dispersion of mathematics

achievement seems to have decreased over the 18 years considered here at every age level. For example, at age 17 it decreased from a SD of 34.9 in 1978 to 30.2 in 1996. The standard deviation of school mean mathematics achievement has increased over that time (from a SD of 9.8 in 1978 to 13.4 in 1996 at age 17). As a result of these two trends, between-school variation has increased as a fraction of total variation (from 28% in 1978 to 44% in 1996 at age 17). Thus schools have become more diverse, more segregated by mathematics achievement, over this time period.

Insert Table 2 About Here

Table 3 provides information on NAEP science achievement for four years between 1977 and 1996, organized in the same way as Tables 1 and 2. As in mathematics, in all but one case, the between schools achievement variation in science is less than half of the overall national standard deviation. The dispersion of science achievement seems to have decreased over the 19 years considered here for 8th and 4th graders. For example, at age 13 it decreased approximately 12% from a SD of 43.5 in 1977 to 38.3 in 1996. The standard deviation of school mean science achievement has increased over that time (from a SD of 13.4 in 1977 to 19.8 in 1996 at age 17). As a result, the between-school variation has increased as a fraction of total variation (from 29.7% in 1978 to 43.9% in 1996 at age 17). Thus, in congruence with trends in reading and mathematics, schools have become more diverse, more segregated by science achievement, over this time period.

Insert Table 3 About Here

How Large is Between-School Variation in Achievement Net of Student Background?

The third row of each panel of Table 1 shows the estimate of the standard deviation of school mean reading achievement controlling for SES, gender, and race/ethnicity and this standard deviation as a percentage of the standard deviation of the total national student reading achievement distribution. The standard deviation between schools is only about half as large as the unadjusted between-school standard deviation after as the student background factors of SES, gender, and race/ethnicity are included in the achievement model. This analysis shows that much of the variation between schools in America is explained by student background factors. After controlling for student background, the school mean variation in NAEP reading achievement is only 20-25% as large as the total national standard deviation in 1996.

The third row of each panel of Table 2 shows the estimate of the standard deviation of school mean mathematics achievement controlling for SES, gender, and race/ethnicity and this standard deviation as a percentage of the standard deviation of total national mathematics achievement. As in the case of reading, much of the variation between schools in America is explained by the student background factors of SES, gender, and race/ethnicity. Only a little more than half of the variation between schools remains after these student background

factors are included in the achievement model. After controlling for student background, the school mean variation in NAEP mathematics achievement is only 25% as large as the total national standard deviation in 1996.

The third row of each panel of Table 3 shows the estimate of the standard deviation of school mean science achievement controlling for SES, gender, and race/ethnicity and this standard deviation as a percentage of the standard deviation of total national mathematics achievement. As in the case of reading and mathematics, much of the variation between schools in America is explained by the student background factors of SES, gender, and race/ethnicity. Only a little less than half of the variation between schools remains after these student background factors are included in the achievement model. After controlling for student background, the school mean variation in NAEP science achievement is only about 20% as large as the total national standard deviation in 1996.

How Much Have Within-School Achievement Gaps Changed Over Time?

School reforms might target not just average achievement, but also achievement gaps between groups within schools. Although the details of analyses of within-school achievement gaps are beyond the scope of this paper and are not presented here, we will briefly summarize those results. The average within-school achievement gaps between Blacks and Whites and between Hispanics and Whites (controlling for gender and SES) have decreased considerably over time in reading, mathematics, and science. The gender gap

(controlling for SES and race/ethnicity) has increased slightly in reading and decreased slightly in mathematics and science. The SES gap, measured by the coefficient representing the effect of the change in achievement associated with one unit in our composite SES score, is essentially unchanged in reading, mathematics, and science over the time period of this study.

How Much Do Within-School Achievement Gaps Vary Across Schools?

Although the details are not presented here, we will briefly summarize those results. The variation across schools in the gender gap (measured by the standard deviation of the school-specific gender effects) seems to have increased over time in reading, mathematics, and science. The variation across schools in the Black-White achievement gap (measured by the standard deviation of the school-specific Black-White effects) seems to have increased over time in reading and science but not in mathematics. The variation across schools in the Hispanic-White achievement gap seems to have increased in mathematics and science, but not in reading. Perhaps most interesting, the variation in the SES effects across schools has increased dramatically over the time period studied. In 1996, the standard deviation across schools of the SES effects at age 17 was three times as large in reading as in 1971, over twice as large in mathematics as in 1978, and over twice as large in science as in 1977. This seems to suggest that schools are not just getting more diverse in their average achievement, but more diverse in how well they meet the needs of students at different SES levels.

How Large An Effect Should We Expect from School Reform Programs?

In this section the results of the school effects analyses are used to provide a normative framework for interpreting achievement differences between schools. The premise is that naturally occurring differences between schools yield a population of more effective and less effective schools. Reforms are intended to make less effective schools into more effective ones. Thus the achievement differences resulting from reforms should be similar in magnitude to the achievement differences between naturally occurring less effective and more effective schools. In particular,

> *a school reform is unlikely to create a school that is more effective than any current schools (some of which have reforms in place or are the models on which reforms are based).*

Our school effects analyses demonstrate that a substantial proportion of the variation in school effects is due to differences in student background. Since school reforms are not intended to change student background (that is they do not generally attempt to obtain gains in achievement by eliminating poor children or ethnic minorities from the school), the relevant variation in school effects is the variation left after controlling for student background. That is, an effective school is one that has relatively high mean achievement after controlling for the effects of student background.

Consider first reforms that are targeted at the median school, designed to make it better. Tables 4, 5 , and 6 give the magnitude of the change in school

mean achievement required to move a median school to various percentiles in the school mean achievement distribution (controlling for student background) in reading (Table 4), mathematics (Table 5), and science (Table 6) for 12th, 8th, and 4th graders in 1996. To aid in interpretation of these differences, we have also compared the difference in school mean achievement to three normatively well known national achievement gaps: the gender (Male-Female), race (Black-White), and family background (parental education) achievement gaps which are measured by NAEP (see, e.g., Hedges and Nowell, 1995; Hedges and Nowell, 1999). We measured the parental education gap slightly differently in reading, mathematics, and science because the data available from NAEP are slightly different in the two subject matters. The parental education gap is the mean difference in achievement between students whose parents had not graduated from high school and students whose parents had at least some college (in NAEP reading) or graduated from college (in NAEP mathematics and science).

Insert Tables 4, 5, and 6 About Here

One could argue that a feasible goal that would have real policy significance might be to move a school from median (50th percentile) to the 70th percentile among schools nationally. This would move a school past 20% of the schools in nation (assuming the others stood still). Most principles or superintendents would declare such a change to be a real success. Assuming that school effects are normally distributed (and our analyses strongly support this assumption), such an effect requires a change of about one half of a standard

deviation in the distribution of (student background adjusted) school means and would correspond to an increase of 4.3 NAEP scale score points in reading, 3.9 NAEP scale score points in mathematics, and 4.6 NAEP scale score points in science for 12th graders.

However if we use the size of achievement gaps to judge the importance of this reform, we might arrive at a different conclusion about its importance. The impact on the average student would be only about a tenth of a national standard deviation of student achievement in reading and science, and about an eighth of a standard deviation in mathematics. For students aged 13 and 17, this school effect is nearly 15% of the Black-White achievement gap or the achievement gap associated with parental education (in reading and mathematics), which are usually considered "large" effects. In science the school effect is about 10% of the Black-White achievement gap or the achievement gap associated with parental education. For students aged 13 and 17, the reform effect is a much larger fraction of the gender gap, about 30% of the modest achievement gap favoring females in reading, over 80% of the much smaller achievement gap favoring males in mathematics, and over 40% of the smaller achievement gap in science. The reform effect for students at age 9 is somewhat larger compared to the gender, race, and parental education effects for all achievement scores.

One might consider a very powerful reform to be one that moved the average school to the 90th percentile. Such reforms may be possible on a large

scale, but it seems unrealistic to hold every reform to such a high standard. Tables 4 to 6 show that such a reform would increase average achievement by 10.6 NAEP scale score points in reading, 9.5 NAEP scale score points in mathematics, and 11.3 NAEP scale score points in science for 12th graders, or roughly a quarter of a national student standard deviation in reading, a third of a national student standard deviation in mathematics, and a quarter of a national student standard deviation in science. For students aged 13 and 17, this corresponds to just over a third of the Black-White achievement gap and about three quarters of the gender gap in reading. The reform effect is less than a third of the Black-White achievement gap in mathematics and reading. Thus the effects of even a very powerful reform are considerably smaller than the Black-White achievement gap, and smaller than the gender gap in reading.

Consider now reforms that are targeted at schools that are failing. One might say for the purposes of argument that a failing school is one that is in the bottom 10 percent of the school effects distribution. Obviously there is more room for improvement in these schools than for average schools. What kind of impact on student achievement might be expected by targeting schools at the 10th percentile? Table 7 shows the impact on achievement of moving a school that is at the 10th percentile in reading achievement to various higher percentiles in the national achievement distribution in 1996. Table 8 shows the corresponding information for schools that are at the 10th percentile in mathematics achievement in 1996. Table 9 shows the corresponding information for schools that are at the

10th percentile in science achievement in 1996. To aid in interpretation of these differences, we have also compared the difference in school mean achievement to the same normatively well known national achievement gaps the gender, Black-White, and parental education achievement gaps which are measured by NAEP.

Insert Tables 7, 8, and 9 About Here

One could argue that a feasible goal that would have real policy significance might be to move a school from the 10th percentile to the 30th percentile among schools nationally. Such an effect would require a change of about three quarters of a school standard deviation and would correspond to 6.3 NAEP scale points in reading, a 5.6 NAEP scale points in mathematics, and a 6.7 NAEP scale score points in science for 12th graders. At ages 13 and 17, this would correspond to 15-20% of the Black-White achievement gap and 15-20% of the achievement gap associated with parental education.

A reform with larger impact might be expected to move a school from the 10th percentile to median (the 50th percentile) among schools nationally. Such an effect would require exactly the same change as that of moving the average school to the 90th percentile discussed above (a change of about 1.28 standard deviations in the distribution of student background adjusted school means, corresponding to an increase of 10.6 NAEP scale score points in reading, 9.6 NAEP scale score points in mathematics, and 11.3 NAEP scale score points in science for 12th graders) and would have the same interpretation. At ages 13 and 17 the change would be only about a third as large as the Black-White or parental

education achievement gaps for reading and mathematics, and about a quarter as large for science.

A *very* large reform effect might be to move a 10th percentile school to the 90th percentile. It is unclear of any reforms that can reliably produce such effects exist. If so, they are instruments of extraordinary importance to education reform because they would permit schools to move over practically the entire distribution of American schools (from the bottom to the top). A reform this powerful has an effect that is 50% larger than the gender gap in reading at ages 13 and 17, but its effect would still be considerably smaller than the Black-White achievement gap in reading, mathematics, or science.

Conclusions

Data from school effects analyses of NAEP show that most of the achievement variation in American schools is within schools not between (among) them. When student background characteristics are taken into account, there is even less variation between schools. Therefore interpreting the magnitude of the effects of school reform in terms of individual variation and the achievement gaps between groups may not only be disappointing, but also misleading. For example, the effect in NAEP score units of a reform that would move a school from the 10th percentile to the 90th percentile of effectiveness (student background adjusted mean achievement) is only half to two thirds of a standard deviation of student scores. While Cohen's (1977) convention may state

that half a standard deviation of the student achievement distribution is a "medium sized" effect in terms of individual studies, we would argue that it should be interpreted as a very large effect in terms of school reform. Indeed this effect *is* a much larger fraction of the standard deviation of school effects.

Tables 4 to 9 illustrate how one description of school reforms, a change in percentile rank of the school within the national distribution of schools, can be related to a metric (NAEP scale score points) which can in turn be compared to other achievement differences (such as achievement gaps between policy relevant groups in American society) which have been independently judged to be large or small. Reasonable people could disagree with the feasibility or importance of any particular impact of reform that we have posited here. One might think that moving a school from the 50th percentile to the 70th percentile is either a trivial or a monumental achievement. One might regard gender difference in reading to be a large disadvantage for boys and therefore be reluctant to use it as an index of a modest effect. Regardless, the method suggested here provides a way to gain insight into the plausibility of school effects of various sizes.

One might question whether other sources of data would yield similar results. For example perhaps the 1996 NAEP long-term trend data has some special feature that understates school effects. We do not believe that any feature of the NAEP sampling design would cause an underestimate of between-school variation. Moreover, the fact that between-school variation has increased over

time in NAEP would suggest that the same calculations performed on earlier years of the NAEP data would lead to a distribution of school effects that was less dispersed than that in 1996. That is, school effects that are large in an absolute sense would be even less frequent in earlier years of NAEP data. However our research team has replicated these analyses using other datasets that also have national probability samples (the National Longitudinal Survey of the High School Class of 1972 (NLS-72), High School and Beyond (HS&B), and the National Longitudinal Study of the Eighth Grade Class of 1988 (NELS:88). We reach qualitatively similar conclusions from all of them.

The results of this study and similar investigations can be used to provide a way to obtain plausible treatment (reform) effects for designing studies of school reform. Reasonable values of expected effects are essential to design evaluation studies that have sufficient power to detect the effects of school reform interventions. The results of this study can also provide a context in which to evaluate the results of studies of the effectiveness of school reform in terms of national data. It is essential if we are to have reasonable expectations for school reforms and fairly judge whether they have met reasonable expectations. Such a context helps us to answer the question "Do the results of this study of reform indicate a big effect or a disappointment?"

This study also illustrates one way in which survey data can contribute to evidence-based policy formation. The analysis of between-school achievement distribution to estimate the distribution of naturally occurring school effects

provides a basis for estimating plausible effect magnitudes for planning intervention studies. These effect magnitudes can be used for estimating statistical power of either primary analyses (see, e.g., Cohen 1977) or syntheses of many intervention research studies (see Hedges and Pigott, 2001) and thus should assist in planning and interpretation of both. The analyses of school effects can also provide a context for interpreting treatment effects within the context of naturally occurring variation. It permits the policy researcher to explain the implications of treatment effects within the backdrop of natural variation within which any intervention will operate.

The intent of this paper is not to suggest that achievement gaps are unimportant, nor that research need not address them. Inequality in American education is precisely what is driving school reform and the existing degree of inequality is a major national problem. The danger is that real reform that improves the quality of education must not be judged by standards that preordain its evaluation as a failure.

References

Beaton, A. E. & Zwick, R. (1992). An overview of the National Assessment of Educational Progress. Journal of Educational Statistics, 17, 95-110.

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

Bryk, A. S. & Raudenbush, S. W. (1992). Hierarchical linear models. Newbury

Park, CA: Sage Publications.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. New York: Academic Press.

Hedges, L. V. & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. Science, 269, 41-45.

Hedges, L. V. & Nowell, A. (1999). Changes in the Black-White gap in achievement test scores: The evidence from nationally representative samples. Sociology of Education, 72, 111-135.

Hedges, L. V. and Pigott, T. D. (2001). The power of statistical tests in meta -analysis. Psychological Methods, 6, 203-217.

Johnson, E. G. (1989). Considerations and techniques for the analysis of NAEP data. Journal of Educational Statistics, 14, 303-334.

Johnson, E. G. (1992). The design of the national Assessment of Educational Progress. Journal of Educational Measurement, 29, 95-110.

Mislevey, R. J. (1988). Randomization-based inferences about latent variables from complex samples. Psychometrika, 56, 177-196.

Mullis, I. V. S. (1990) The NAEP guide. Washington, DC: National Center for Education Statistics.

Raudenbush, S. W. & Byrk, A. S. (1987). A hierarchical model for studying school effects. Sociology of Education, 59, 1-17.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi -experimental designs for generalized causal inference. Boston: Houghton

Mifflin.

Footnotes

1. If Y_{ij} is the achievement test score of the i^{th} student in the j^{th} school, this achievement model can be represented symbolically as

$$Y_{ij} = \_0j + \_ij,$$

where $\_0j$ is a school-specific intercept and $\_ij$, is a student-within-school specific residual.

2. The second achievement model we employ includes the student characteristics family SES, gender (the effect of being female), and race/ethnicity modeled via dummy codes for Black, Hispanic, and Other, so that each effect represents the difference in achievement between the named group and Whites, controlling for SES and gender. Thus the achievement model for the i^{th} student in the j^{th} school becomes

$$Y_{ij} = \_0j + \_1jSES_{ij} + \_2jFEMALE_{ij} + \_3jBLACK_{ij} + \_4jHISPANIC_{ij} + \_5jOTHER_{ij} + \_ij,$$

where SES_{ij} is a composite index of socio-economic status of the family, $FEMALE_{ij}$ is a dummy variable for gender, $BLACK_{ij,}$ $HISPANIC_{ij,}$ and $OTHER_{ij}$ are indicator variables for Black, Hispanic, or Other group membership, and $\_ij$ is a student-specific residual.

3. In the achievement model with no level 1 covariates, there is only one _ in the school-specific achievement model ($\_0j$) and thus the between-school model corresponds to

$$\_0j = \_00 + \_0j,$$

where $\_00$ is the average achievement across all schools and $\_0j$ is a school-effect

(the difference between the average achievement in the j^{th} school and that of the average school nationally). The standard deviation of the β_{0j}'s is a measure of how much average achievement varies across schools. In the achievement model with five level 1 covariates, there are six β's (β_{0j}, β_{1j}, β_{2j}, β_{3j}, β_{4j}, β_{5j}), and the specific level two model for the m^{th} coefficient in the j^{th} school β_{mj} is therefore

$$\beta_{mj} = \beta_{0m} + \beta_{mj},$$

where β_{0m} is the average effect across all schools and β_{mj} is a school-effect (the difference between the effect in the j^{th} school and that of the average effect across schools nationally). For the m^{th} coefficient, the standard deviation of the β_{mj}'s is a measure of how much the m^{th} effect varies across schools.

4. We have compared the results of analyses using this specification of SES with others, including those involving parental education and income in High School and Beyond and NELS:88 and found that they yield very similar results.

5. In conventional designs, test scores are estimated for each individual and then analyzed to estimate structural relations. In these analyses unreliability of test scores leads to bias in estimation of structural relations (including variation). The NAEP design does not estimate test scores for individual students, but uses student information in the form of "plausible values" to estimate structural relations. In the NAEP design, the small amount of information obtained from each student increases sampling error of estimates rather than introducing bias (see Mislevy, 1988; Johnson, 1989).

Table 1

NAEP Reading Achievement: Variation Between and Within Schools

	1971		1975		1980		1992		1996	
	SD	%	SD	%	SD	%	SD	%	SD	%
Age 17										
Total	45.8		44.0		41.8		43.0		42.2	
Between-School	14.9	32.6%	13.2	30.1%	10.6	25.4%	16.3	37.9%	16.9	40.0%
Adjusted Between School	6.2	13.5%	5.2	11.9%	3.4	8.2%	7.9	18.3%	8.3	19.6%
Age 13										
Total	35.7		35.8		34.9		39.4		39.1	
Between-School	10.8	30.3%	10.0	30.0%	9.2	26.4%	18.7	47.4%	16.4	42.8%
Adjusted Between School	4.1	11.5%	5.3	14.8%	3.9	11.0%	10.1	25.6%	7.9	20.4%
Age 9										
Total	42.1		38.6		37.9		40.3		39	
Between-School	14.0	33.3%	12.2	31.5%	8.4	22.2%	16.7	41.5%	17.7	43.6%
Adjusted Between School	6.15	14.6%	5.6	14.4%	4.5	11.8%	9.0	22.3%	10.3	25.4%

Table 2

NAEP Mathematics Achievement: Variation Between and Within Schools

	1978		1982		1992		1996	
	SD	%	SD	%	SD	%	SD	%
Age 17								
Total	34.9		32.4		30.1		30.2	
Between-School	9.8	28.1%	9.2	28.3%	12.5	41.4%	13.4	44.3%
Adjusted Between School	5.7	16.4%	5.3	16.5%	6.0	19.9%	7.4	24.5%
Age 13								
Total	39.0		33.4		30.9		31.6	
Between-School	13.7	35.2%	10.5	31.4%	14.8	47.8%	16.5	52.1%
Adjusted Between School	6.6	17.0%	5.7	17.0%	8.1	26.1%	7.5	23.8%
Age 9								
Total	36		34.8		33.1		33.8	
Between-School	10.3	28.6%	10.4	29.9%	13.7	41.4%	14.5	42.9%
Adjusted Between School	6.4	17.9%	5.5	15.8%	7.7	23.2%	8.4	24.8%

Table 3

NAEP Science Achievement: Variation Between and Within Schools

	1977		1982		1992		1996	
	SD	%	SD	%	SD	%	SD	%
Age 17								
Total	45.0		46.7		44.7		45.1	
Between-School	13.4	29.7%	13.3	28.6%	20.3	45.4%	19.8	44.0%
Adjusted Between School	3.7	8.3%	4.2	8.9%	9.1	20.3%	8.8	19.6%
Age 13								
Total	43.5		38.6		36.9		38.4	
Between-School	12.2	28.0%	13.1	33.9%	18.0	48.8%	19.6	50.9%
Adjusted Between School	5.6	12.9%	5.4	13.9%	7.5	20.4%	7.3	18.9%
Age 9								
Total	44.9		40.9		39.9		42.1	
Between-School	14.0	31.3%	15.1	37.0%	17.3	43.3%	18.8	44.8%
Adjusted Between School	5.7	12.8%	5.8	14.2%	8.5	21.2%	9.0	21.3%

Table 4

Effect of Moving a Median School to a Given Percentile as a Percentage of Various Achievement Gaps Estimated from 1996 NAEP Reading Data

		School Effect in Various Metrics			
Target Percentile	NAEP Scores	School SD Units	% Black-White Gap	% Parental Education Gap	% Gender Gap
Age 17					
60	2.1	0.25	7.2	6.9	14.5
70	4.3	0.52	15.0	14.3	29.9
80	7.0	0.84	24.0	22.9	48.1
90	10.6	1.28	36.6	34.9	73.2
95	13.6	1.64	47.0	44.8	93.9
99	19.3	2.33	66.4	63.4	132.9
Age 13					
60	2.0	0.25	6.2	6.8	15.1
70	4.1	0.52	12.9	14.0	31.2
80	6.6	0.84	20.7	22.5	50.1
90	10.1	1.28	31.5	34.2	76.2
95	12.9	1.64	40.5	43.9	97.8
99	18.3	2.33	57.3	62.1	138.4
Age 9					
60	2.6	0.25	9.0	11.8	24.3
70	5.4	0.52	18.7	24.4	50.3
80	8.6	0.84	30.0	39.1	80.8
90	13.2	1.28	45.7	59.6	123.0
95	16.9	1.64	58.7	76.5	157.9
99	23.9	2.33	83.0	108.1	223.4

Table 5

Effect of Moving a Median School to a Given Percentile as a Percentage of Various Achievement Gaps Estimated from 1996 NAEP Mathematics Data

	School Effect in Various Metrics				
Target Percentile	NAEP Scores	School SD Units	% Black-White Gap	% Parental Education Gap	% Gender Gap
Age 17					
60	1.9	0.25	6.9	5.2	40.7
70	3.9	0.52	14.4	10.7	84.2
80	6.2	0.84	23.0	17.2	135.2
90	9.5	1.28	35.1	26.2	205.9
95	12.2	1.64	45.0	33.7	264.2
99	17.2	2.33	63.7	47.6	373.7
Age 13					
60	1.9	0.25	6.5	6.5	48.8
70	3.9	0.52	13.5	13.5	100.9
80	6.3	0.84	21.7	21.6	162.0
90	9.6	1.28	33.1	32.9	246.6
95	12.3	1.64	42.4	42.3	316.5
99	17.5	2.33	60.0	59.8	447.7
Age 9					
60	2.1	0.25	8.4	10.7	54.5
70	4.4	0.52	17.4	22.1	112.9
80	7.1	0.84	27.9	35.5	181.1
90	10.8	1.28	42.5	54.1	275.8
95	13.8	1.64	54.6	69.4	354.0
99	19.5	2.33	77.2	98.1	500.7

Table 6

Effect of Moving a Median School to a Given Percentile as a Percentage of Various Achievement Gaps Estimated from 1996 NAEP Science Data

	School Effect in Various Metrics				
Target Percentile	NAEP Scores	School SD Units	% Black-White Gap	% Parental Education Gap	% Gender Gap
Age 17					
60	2.2	0.25	4.8	4.6	28.3
70	4.6	0.52	10.0	9.5	58.6
80	7.4	0.84	16.0	15.3	94.1
90	11.3	1.28	24.3	23.3	143.2
95	14.5	1.64	31.2	29.9	183.8
99	20.5	2.33	44.2	42.3	260.0
Age 13					
60	1.8	0.25	4.6	5.0	20.9
70	3.8	0.52	9.5	10.4	43.3
80	6.1	0.84	15.2	16.7	69.5
90	9.3	1.28	23.2	25.4	105.8
95	11.9	1.64	29.7	32.6	135.8
99	16.9	2.33	42.0	46.2	192.0
Age 9					
60	2.3	0.25	6.1	7.6	66.7
70	4.7	0.52	12.7	15.8	138.1
80	7.5	0.84	20.3	25.3	221.6
90	11.5	1.28	30.9	38.5	337.4
95	14.7	1.64	39.7	49.4	433.0
99	20.8	2.33	56.1	69.9	612.4

Table 7

Effect of Moving a 10th percentile School to a Given Percentile as a Percentage of Various Achievement Gaps Estimated from 1996 NAEP Reading Data

	School Effect in Various Metrics				
Target Percentile	NAEP Scores	School SD Units	% Black-White Gap	% Parental Education Gap	% Gender Gap
Age 17					
20	3.6	0.44	12.5	11.9	25.0
30	6.3	0.76	21.6	20.6	43.2
40	8.5	1.03	29.3	28.0	58.6
50	10.6	1.28	36.6	34.9	73.1
60	12.7	1.53	43.8	41.8	87.6
70	14.9	1.80	51.5	49.2	103.0
80	17.6	2.12	60.6	57.8	121.2
90	21.2	2.56	73.1	69.8	146.3
95	24.2	2.92	83.5	79.7	167.0
99	29.9	3.61	103.0	98.2	206.0
Age 13					
20	3.4	0.44	10.8	11.7	26.1
30	5.9	0.76	18.6	20.2	44.9
40	8.1	1.03	25.3	27.4	61.1
50	10.0	1.28	31.5	34.2	76.1
60	12.0	1.53	37.7	40.9	91.2
70	14.2	1.80	44.4	48.2	107.3
80	16.7	2.12	52.2	56.7	126.2
90	20.1	2.56	63.0	68.4	152.4
95	23.0	2.92	72.0	78.1	174.0
99	28.3	3.61	88.8	96.3	214.5
Age 9					
20	4.5	0.44	15.6	20.4	42.1
30	7.8	0.76	27.0	35.1	72.5
40	10.5	1.03	36.6	47.7	98.6
50	13.1	1.28	45.7	59.5	122.9
60	15.8	1.53	54.7	71.3	147.2
70	18.5	1.80	64.4	83.9	173.2
80	21.8	2.12	75.7	98.6	203.7
90	26.3	2.56	91.4	119.1	245.9
95	30.0	2.92	104.3	136.0	280.8
99	37.0	3.61	128.6	167.6	346.2

Table 8

Effect of Moving a 10th Percentile School to a Given Percentile as a Percentage of Various Achievement Gaps Estimated from 1996 NAEP Mathematics Data

	School Effect in Various Metrics				
Target Percentile	NAEP Scores	School SD Units	% Black-White Gap	% Parental Education Gap	% Gender Gap
Age 17					
20	3.2	0.44	12.0	9.0	70.4
30	5.6	0.76	20.7	15.5	121.4
40	7.6	1.03	28.1	21.0	164.9
50	9.5	1.28	35.0	26.2	205.6
60	11.3	1.53	42.0	31.4	246.3
70	13.3	1.80	49.4	36.9	289.9
80	15.7	2.12	58.1	43.4	340.8
90	18.9	2.56	70.1	52.4	411.5
95	21.6	2.92	80.1	59.9	469.9
99	26.7	3.61	98.7	73.8	579.4
Age 13					
20	3.3	0.44	11.3	11.3	84.4
30	5.7	0.76	19.5	19.4	145.4
40	7.7	1.03	26.5	26.4	197.6
50	9.6	1.28	33.0	32.9	246.3
60	11.5	1.53	39.5	39.4	295.1
70	13.5	1.80	46.5	46.4	347.2
80	15.9	2.12	54.7	54.5	408.3
90	19.2	2.56	66.1	65.8	492.9
95	22.0	2.92	75.4	75.2	562.8
99	27.1	3.61	93.0	92.7	694.0
Age 9					
20	3.7	0.44	14.5	18.5	94.4
30	6.3	0.76	25.1	31.9	162.6
40	8.6	1.03	34.1	43.3	221.0
50	10.7	1.28	42.5	54.0	275.5
60	12.9	1.53	50.9	64.7	330.0
70	15.1	1.80	59.9	76.1	388.4
80	17.8	2.12	70.4	89.5	456.6
90	21.5	2.56	85.0	108.0	551.3
95	24.6	2.92	97.0	123.4	629.5
99	30.3	3.61	119.7	152.1	776.2

Table 9

Effect of Moving a 10th percentile School to a Given Percentile as a Percentage of Various Achievement Gaps Estimated from 1996 NAEP Science Data

	School Effect in Various Metrics				
Target Percentile	NAEP Scores	School SD Units	% Black-White Gap	% Parental Education Gap	% Gender Gap
Age 17					
20	3.9	0.44	8.3	8.0	49.0
30	6.7	0.76	14.3	13.8	84.4
40	9.1	1.03	19.5	18.7	114.7
50	11.3	1.28	24.3	23.3	143.1
60	13.5	1.53	29.1	27.9	171.4
70	15.9	1.80	34.3	32.8	201.7
80	18.7	2.12	40.3	38.6	237.1
90	22.6	2.56	48.6	46.6	286.3
95	25.8	2.92	55.5	53.2	326.9
99	31.8	3.61	68.5	65.7	403.0
Age 13					
20	3.2	0.44	7.9	8.7	36.2
30	5.5	0.76	13.7	15.0	62.4
40	7.5	1.03	18.6	20.4	84.7
50	9.3	1.28	23.1	25.4	105.7
60	11.1	1.53	27.7	30.4	126.6
70	13.1	1.80	32.6	35.8	148.9
80	15.4	2.12	38.3	42.1	175.1
90	18.6	2.56	46.3	50.8	211.4
95	21.2	2.92	52.9	58.0	241.4
99	26.2	3.61	65.2	71.6	297.7
Age 9					
20	3.9	0.44	10.6	13.2	115.4
30	6.8	0.76	18.2	22.7	198.9
40	9.2	1.03	24.8	30.8	270.3
50	11.5	1.28	30.9	38.4	337.0
60	13.7	1.53	37.0	46.1	403.7
70	16.2	1.80	43.5	54.2	475.0
80	19.0	2.12	51.2	63.7	558.5
90	22.9	2.56	61.8	76.9	674.4
95	26.2	2.92	70.6	87.9	770.0
99	32.3	3.61	87.0	108.3	949.4