How Large an Effect Can We Expect from School Reforms?

Spyros Konstantopoulos
Larry Hedges

October 2006

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

How Large an Effect Can We Expect from School Reforms?

Spyros Konstantopoulos
Northwestern University
and IZA Bonn

Larry Hedges
Northwestern University

Discussion Paper No. 2371
October 2006

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
Email: iza@iza.org

Any opinions expressed here are those of the author(s) and not those of the institute. Research disseminated by IZA may include views on policy, but the institute itself takes no institutional policy positions.

The Institute for the Study of Labor (IZA) in Bonn is a local and virtual international research center and a place of communication between science, politics and business. IZA is an independent nonprofit company supported by Deutsche Post World Net. The center is associated with the University of Bonn and offers a stimulating research environment through its research networks, research support, and visitors and doctoral programs. IZA engages in (i) original and internationally competitive research in all fields of labor economics, (ii) development of policy concepts, and (iii) dissemination of research results and concepts to the interested public.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ABSTRACT

How Large an Effect Can We Expect from School Reforms?[*]

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 in 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.

JEL Classification: I20

Keywords: academic achievement, school effects, multi-level models

Corresponding author:

Spyros Konstantopoulos
School of Education and Social Policy
Northwestern University
2120 Campus Drive
Evanston, IL 60208
USA
E-mail: spyros@northwestern.edu

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 NCLB 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 NCLB 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 evaluated via norm referenced assessments like NAEP.

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 largely quantitatively, and an important aspect of this work will be judging *how well* reform strategies work. Such evaluations will indicate which reforms produce large, modest,

small, or no improvements in achievement, or even which reforms have negative effects on academic achievement, and how large those increases (or decreases) in academic achievement 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 minority groups such as Blacks or Hispanics and Whites, rich and poor, and males and females. Ideally the objective of school reform is to increase achievement for all students *and* simultaneously close the achievement gap between lower and higher achieving students (that is produce larger gains for lower achieving students or have all students perform as well as the higher achieving students). It is natural then, 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 for all students by one tenth of a standard deviation might seem too weak to be important, while a reform that could increase achievement for all students by three quarters of a standard deviation might seem quite important (since it is a substantial proportion of the Black-White gap). Notice that in this context, school reform effects are essentially evaluated by comparing them to the size of the student achievement gaps between lower and higher achievers (e.g., the achievement gap between a below average student and an above average student).

The second solution to the problem of interpreting the effects of reforms is derived from a similar idea, that school reforms are intended to make all schools perform as well as the best schools (or reduce the achievement gap between lower and higher achieving 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 a below average school (e.g., a school at the 25th percentile of all American schools) and an above average school (e.g., a school at the 75th percentile of all American schools). This interpretative context is explicitly normative, comparing reform effects with the normative distribution of school effects. Using the normative distribution of school effects one could argue that a school reform 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 school reform effect that would move a median school to the 90th percentile of all schools might be considered a large or important effect.

The purpose of this paper is to explore these two alternative frameworks for interpreting the effects of school 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 however, is more likely to focus on comparisons among schools (see Lee & Bryk, 1989; Raudenbush & Bryk, 1986). While it is conceivable and perhaps desirable to combine the two perspectives, our experience is that one perspective often overshadows the other. 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 also argue that the normative distribution of school effects framework is more appropriate than the other framework for interpreting the likely magnitude of school reform effects. In addition, we show that interpreting the magnitude of the effects of school reforms in terms of individual variation and the achievement gaps between important student groups may not only be disappointing, but also misleading. 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 & 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.

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. 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.

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 what 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. Such knowledge can help set educational goals that are more reasonable and more likely to be attained. In contrast, goals set in the abstract may not be realistic and difficult to attain in the sense that it is not obvious that schools can function in certain ways to meet intangible goals.

Hence, we argue that the distribution of the observed school effects is a useful gauge to what is *not* possible, or what is realistic. 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 and we should strive to do so. But to *require* such radical change as the main criterion of success probably dooms educational reform to failure. While there is naturally great optimism among proponents of reform about the magnitude of the effects that 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 or important without an appropriate interpretive context that reflects reality. Hence, we contend that the distribution of the observed school effects provides a plausible framework that indicates what reform effects are likely to be produced.

School Effects Research

School effects research emerged about 40 years ago with the path-breaking *Equality of Educational Opportunity Study* (EEOS). This was the first large-scale study that examined rigorously the association between school inputs such as school resources and school outputs such as academic achievement using national probability samples of elementary and secondary students in America (Coleman et al., 1966). One of the key findings of the Coleman Report was that family background factors such as SES had a significant impact on student achievement, while school factors had a relatively small impact on student achievement. The main conclusion of the Coleman Report was assumed by many researchers to be that schools had hardly any effect on student achievement or that schools do not matter.

The Coleman Report generated a series of studies that further assessed the effects of schools on academic achievement the last 30 years. It is noteworthy that, there have been disagreements among educational researchers, practitioners, and policymakers about the relative impact-importance of school factors on students' academic achievement. The findings of numerous studies are rather mixed and inconclusive. Some reviewers have concluded that there is little or no evidence of a relationship between school factors and student achievement (Hanushek, 1986; 1989), while others report that the impact of school factors on test scores may be substantial (Hedges, Laine, & Greenwald, 1994; Greenwald, Hedges, and Laine, 1996).

Achievement Models in School Effects Research

In the 1980s methodological advances facilitated school effects research by permitting investigators to gauge the importance of school factors in predicting student achievement more

accurately. During this period, multi-level statistical models were introduced and allowed the use of student and school factors at the appropriate level of analysis (Raudenbush & Bryk, 1986; Bryk & Raudenbush, 1988). Specifically, the flexibility of multi-level models allowed for the use of student characteristics at the student level and the school factors at the school level. In addition, multi-level models allowed the computation of between-school variation in achievement, which is advocated by some researchers to represent school effects (see Bryk, Lee, & Holland, 1993; Constant & Konstantopoulos, 2003; Raudenbush & Wilms, 1995).

Such multi-level models involve typically two levels: a within-school achievement model and a between-school model where school-specific estimates can vary across schools. At the within-school model academic achievement is typically regressed on student demographic characteristics such as SES, race/ethnicity, gender, and previous achievement (Bryk & Raudenbush, 1988; Lee, 2000; Raudenbush & Bryk, 1986, 2002). The school effects are estimated at the between-school model, and the specification at that level depends on the question of interest. That is, many times specific school factors such as structure, organization, or composition are included as predictors in the second level regression (see Lee, 2000). Other times however, the main objective is to compute the between-school variation in achievement and in such cases the between-school models may not include any school predictors. Nonetheless, when one is estimating school effects it is important to control appropriately for student characteristics at the within-school model.

The Within-School Achievement Model

Previous work on the correlates of academic achievement has lead to a considerable consensus about the types of variables that need to be included in academic achievement and school-effects models. Specifically, there is little disagreement over the existence of a positive association between family background and academic achievement (Jencks et al, 1979). For example, the relationship between test scores and family SES is a widely replicated finding in the social sciences (White, 1982; White, Reynolds, Thomas, & Gitzlaff, 1993). The strength of the relationship between SES and academic achievement varies from study to study partly because researchers operationally define SES in different ways and this affects the magnitude or strength of the association (White, 1982). In addition, family SES is sometimes constructed differently in different studies because of data availability. That is, certain data may not include information about family income, but they may include information about household possessions instead. Traditional measures of SES include parental education, family income, and household possessions (Coleman, 1969; Konstantopoulos, Modi, & Hedges, 2001; White et al., 1993). Hence, SES is typically included in specifications that examine achievement models and estimate school effects.

In addition, previous studies have also demonstrated a substantial gap between minority students (e.g., Blacks, Hispanics) and White students. Work that used nationally probability samples found that the Black-White achievement gap ranges from 3/4 to one standard deviation (SD), while the Hispanic gap is smaller (e.g., Campbell, Hombo, & Mazzeo, 2000; Jencks & Phillips, 1998). It is noteworthy, that these differences remain considerable even after adjusting for differences in SES. For example, Hedges and Nowell (1999) found that adjusting for

differences in SES reduced the Black-White achievement gap by about 1/3, but the gap was still

larger than 1/2 SD. Hence, race/ethnicity effects are also taken into account in specifications that

examine achievement models and estimate school effects.

Finally, gender differences in achievement have also been examined extensively. Studies

using national probability samples of students have found gender differences favoring males in

mathematics and females in reading (see Hedges & Nowell, 1995; Willingham & Cole, 1997).

However, the gender achievement gap was estimated to be about 1/5 SD or smaller. This

indicates that gender effects on academic achievement are smaller than SES and race/ethnic

effects. Nonetheless, it is not uncommon to include gender effects in specifications that examine

achievement models and estimate school effects.

Analysis

School effects models can best be described in terms of a hierarchy with two levels (see

Raudenbush & Bryk, 1986; 2002). 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 SES, gender,

race/ethnicity. Thus, the achievement model includes a specific term, the school effect, that

describes how the average achievement of students in each particular school differs from that of

other schools, controlling for student characteristics. The achievement model usually includes

parameters that describe the relation between individual demographics 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 (e.g., random school intercepts) 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 study we use two different achievement models. The first model simply treats all variation within the school as random. It is used to describe how much of the national variation in achievement is between 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. If Y_{ij} is the achievement test score of the ith student in the jth school, this achievement model can be represented symbolically as

$$Y_{ij} = \beta_{0j} + \varepsilon_{ij},$$

where β_{0j} is a school-specific intercept and ε_{ij}, is a student-within-school specific residual. The second achievement model we employ includes the student characteristics of family SES, gender, and race/ethnicity (used as covariates that adjust the school effects). Thus the achievement model for the ith student in the jth school becomes

$$Y_{ij} = \beta_{0j} + \beta_{1j}SES_{ij} + \beta_{2j}FEMALE_{ij} + \beta_{3j}BLACK_{ij} + \beta_{4j}HISPANIC_{ij} + \beta_{5j}OTHER_{ij} + \varepsilon_{ij},$$

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 ε_{ij} is a student-specific residual. The estimate of each race/ethnicity dummy represents the difference in average achievement between the named group and Whites, controlling for SES and gender. Race/ethnicity was characterized by dividing the population into four groups used by NAEP: White, Black, Hispanic, and Other. In this study, we selected Whites as the comparison group. However, any race/ethnic group can be used as the comparison group and the adjustment due to race/ethnic effects remains the same. Family SES was a composite variable including information about parental education and items in the home (the only SES indexes provided by NAEP data).

The between-school specification for the first model simply represents the variation of effects across schools as random and remained the same in all analyses. That is, since the main objective of this study is the computation of the between-school variation in achievement, school level predictors are not included in the model. In the case of the first achievement model discussed above, we measure how much average school achievement varies across schools by the standard deviation of the school average achievement. That is, in the achievement model with no level 1 predictors there is only one coefficient in the school-specific achievement model (β_{0j}) and thus the between-school model corresponds to

$$\beta_{0j} = \gamma_{00} + \eta_{0j},$$

where γ_{00} is the average achievement across all schools and η_{0j} is a school-effect (the difference between the average achievement in the jth school and that of the average school nationally). The standard deviation of the η_{0j}'s is a measure of how much average achievement varies across schools. The distribution of the η_{0j}'s is the normative distribution of school effects.

In the case of the second within-school achievement model, there are six predictors in the achievement model for each school (an average, the effects of SES, gender, and the achievement gaps between White and Black, White and Hispanic, and White and Other). Thus, in the achievement model with five level 1 covariates, there are six coefficients in the school-specific achievement model (β_{0j}, β_{1j}, β_{2j}, β_{3j}, β_{4j}, β_{5j}), and the specific level two model for the mth coefficient in the jth school β_{mj} is therefore

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

where γ_{0m} is the average effect across all schools and η_{mj} is a school-effect (the difference between the effect in the jth school and that of the average effect across schools nationally). This indicates that all level 1 estimates are treated as random as level 2. For the mth coefficient, the standard deviation of the η_{mj}'s is a measure of how much the mth effect varies across schools. The distribution of η_{0j}'s is the normative distribution of school effects adjusted for the effects of student characteristics. 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, 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.)[1]. 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.[2] We used the reading, mathematics, and science achievement data from the

NAEP long term trend program to estimate school effects reported in this paper. In our samples

there were nearly 20 students per school on average.

Findings from Analyses of NAEP

In separate sections below, we consider three issues using our analyses of the NAEP data:

(1) we consider how much of the variation in achievement is within schools and how large the

between-school variation is in comparison; (2) we consider how the variation in achievement

between-schools changes when the effects of student SES, gender, and race/ethnicity are taken

into account, and in both cases, we examine the trend over time in the distribution of achievement

and school effects; (3) we examine the implications of the national findings for the likely effects

of school reform interventions. Specifically, we show how the distribution of the observed school

effects can provide a normative context for school effects that may have arisen as a consequence

of school reform efforts. In addition, we compare the interpretive framework of school effects

(that aims to close the school achievement gap) to the interpretive framework that aims to reduce

the student achievement gap.

How Large is the 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 the 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 47% as large as the national standard deviation of student achievement. Alternatively, the between-school variance ranges from 5% to 22% as large as the national variance 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 unequal in reading achievement over this time period. These findings are congruent with those reported in a recent study that examined trends in school effects using nationally

probability samples of 12th graders (Konstantopoulos, in press).

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 unequal in 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 unequal by science achievement, over this time period.

Insert Table 3 About Here

How Large is the Between-School Variation in Achievement Adjusted by 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 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 science

achievement.   As in 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. In the second within-school achievement model we included

family SES, gender, and race/ethnicity as predictors and hence, we were able to compute the

overall gender, race/ethnic, and SES achievement gaps across all schools. The results of these

analyses are summarized in Table 4 for high school seniors. The estimates reported in Table 4 indicate group differences in NAEP points on average. To facilitate interpretation we standardized the estimates by the total variance in NAEP reading, mathematics, and science achievement respectively and report the results in standard deviation units. The average within-school gender achievement gap increased slightly (favoring females) over time in reading, decreased by more than 40 percent in mathematics, and nearly 50 percent in science. Still in 1996 female students significantly outperformed male students by 1/3 SD in reading, while males significantly outperformed their females peers by 1/10 SD in mathematics and by 1/6 SD in science. The average within-school Black-White achievement gap decreased considerably over time by nearly 40 percent in reading, but the decrease in mathematics and science was much smaller (nearly 15 percent). Still in 1996 White students significantly outperformed Black students by 1/2 SD in reading, and nearly by 3/4 SD in mathematics and science. The average within-school Hispanic-White achievement gap decreased over time by nearly 25 percent in reading, but the decrease in mathematics and science was smaller (seven and 17 percent respectively). Still in 1996 White students significantly outperformed Hispanic students by about 1/3 SD in reading, 1/4 SD in mathematics and about 1/5 SD in 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, positive, and significant in reading, mathematics, and science over time. This indicates that over time higher levels of achievement are consistently associated with high levels of family SES.

Insert Table 4 About Here

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

One of the advantages of two-level models is that the within-school model coefficients (e.g., gender, race/ethnic, and SES effects) can vary across schools in the between-school model. Hence, we were able to compute the between-school variation of gender, race/ethnic, and SES effects across all schools. The results of these analyses are summarized in Table 5 for high school seniors.. 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 especially in reading and science. This indicates that in the 1990s the schools have become less egalitarian with respect to the gender gap and that the gender effect is much more pronounced in some schools than in other schools. 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 reading, mathematics and science. This indicates that in the 1990s the schools have become less egalitarian with respect to the Black-White and Hispanic-White gap and that these race/ethnic effects are much more pronounced in some schools than in other schools. 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 as in 1971 in

reading, and over twice as large as in 1978 in mathematics and in 1977 in science. This seems to

suggest that schools are not just getting more diverse in average achievement, but also more

diverse in the family SES achievement gap. This indicates that in the 1990s the schools have

become less egalitarian with respect to family SES and that the SES effects are much more

pronounced in some schools than in other schools. Similar findings were reported in a recent

study that examined trends in school effects using NLS, HSB, and NELS data (Konstantopoulos,

in press).

Insert Table 5 About Here

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 the

observed differences in achievement 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 less effective and more effective schools. In particular, *a*

school reform is unlikely to create a school that is more effective than any of the current schools

(some of which have reforms in place or are the models on which reforms are based). Notice

however, that the degree of effectiveness of the current schools is a function not only of the

intention to treat, but also, among other things, of the way the reform efforts are implemented in

schools, and of the students' family background. This makes the adjustment for SES effects

critical, since ideally the distribution of school effects should be adjusted for family background.

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.

Tables 6, 7, and 8 give the magnitude of the change in school mean achievement required to move a school at a specific percentile to various percentiles in the school mean achievement distribution (controlling for student background) in reading (Table 6), mathematics (Table 7), and science (Table 8) 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 or Hispanic-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). We assume that the school mean achievement distribution is normally distributed.

Insert Tables 6, 7, and 8 About Here

First, consider 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. What kind of impact on student achievement might be expected by targeting

schools at the 10th percentile? 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. This change

would correspond to a 15-20% of the Black-White achievement gap, a 17-26% of the Hispanic-

White gap, and a 15-20% of the parental education achievement gap.

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 an average school to the 90th percentile because of symmetry.

This indicates a change of about 1.28 standard deviations in the distribution of adjusted school

means, corresponding to an increase of 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.

Alternatively, this corresponds to an increase of only 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. One might consider this a very powerful

reform, and it may seem unrealistic to hold every reform to such a high standard. However, for

17 year olds this reform effect would be only about a third as large as the Black-White gap in

reading and mathematics, and about a quarter as large in science. This indicates that even

powerful interventions would be considered as not so effective if evaluated under the

achievement gap framework.

Second, consider reforms that are targeted at average schools. 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 17 year

olds, this school reform effect is nearly 15% of the Black-White, Hispanic-White, or the parental

education achievement gap (in reading and mathematics). In science the school effect is about

10% of the Black-White, Hispanic-White, or parental education achievement gap. For 17 year

olds, 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.

Finally, consider a very large reform effect where a low achieving school moves for the

10th percentile to the 90th percentile. It is unclear that reforms which 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). However, even a reform this powerful would still be account for

70% of the Black-White, Hispanic-White, or parental education gap in reading, 50-90% of the

Black-White, Hispanic-White, or parental education gap in mathematics, and nearly 50% of the

Black-White, Hispanic-White, or parental education gap in science.

Conclusion

Data from school effects analyses of NAEP show that most of the achievement variation

in American schools is within schools not among them. This finding is in congruence with one of

the main findings of the Coleman Report which indicated that most of the variation in

achievement is within schools. 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 (nearly three SD).

Tables 6 to 8 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. 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, analyses reported in a recent study that used national probability samples of high school seniors reached to qualitatively similar conclusions (Konstantopoulos, in press).

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 the observed 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 the observed variation. It permits the policy researcher to explain the implications of treatment effects within the backdrop of observed variation within which any intervention will operate.

In addition, by studying the distribution of school effects one can identify schools that may serve as a model for study. For example, the distribution of school effects includes schools that are more egalitarian (e.g., smaller Black-White, Hispanic-White, or parental education achievement gap) and more effective (e.g., higher academic achievement). These schools can be identified and studied thoroughly to determine the factors that contribute to the success of closing the achievement gap and increasing achievement for all students. This could eventually help with reconsidering the nature of (or guiding) school reform efforts and its implementation in schools.

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. (1988). Toward a more appropriate conceptualization of research on school effects. A three-level hierarchical linear model. *American Journal of Education*, *97*, 65-108.

Bryk, A.S., Lee, V.E., & Holland, P.B. (1993). *Catholic schools and the common good*. Cambridge, MA: Harvard University Press.

Campbell, J. R., Hombo, C. M., & Mazzeo, J. (2000). *NAEP trends in academic progress*. Washington, DC: US Department of Education, National center of Education Statistics.

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

Coleman, J. S. (1969). *Equality and achievement in education*. Boulder, CO: Westview Press.

Coleman, J. S., Campbel, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1996). *Equality of Educational Opportunity*. Washington DC: U.S. Government Printing Office.

Constant, A., & Konstantopoulos, S. (2003). School effects and labor market outcomes for young adults in the 1980s and 1990s. *Applied Economics Quarterly, 49*, 5-22.

Greenwald, R., Hedges, L. V., & Lane, R. D. (1996). The effects of school resources on student achievement, *Review of Educational Research, 66*, 361-396.

Hanushek, E. A. (1986). The Economics of Schooling: Production and Efficiency in Public Schools. *Journal of Economic Literature 24*, 1141-77.

Hanushek, E. A. (1989). The impact of differential expenditures on school performance. *Educational Researcher, 18,* 45-51.

Hedges L. V., & Nowel, 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.

Hedges, L. V., Laine, R., & Greenwald, R. (1994). Does Money Matter: A Meta-Analysis of Studies of the Effects of Differential School Inputs on Student Outcomes. *Educational Researcher, 23*, 5-14.

Jencks, C. S., & Phillips, M. (1998). *The Black-White test score gap.* Washington DC: Brookings Institution Press.

Jencks, C. S., Bartlett, S., Corcoran, M., Crouse, J., Eaglesfield, D., Jackson, G., McClelland, K., Mueser, P., Olneck, M., Schwartz, J., Ward, S., & Williams, J. (1979). *Who gets ahead? The determinants of economic success in America*. New York: Basic Books.

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.

Konstantopoulos, S. (in press). Trends of School Effects on Student Achievement: Evidence from NLS:72, HSB: 82, and NELS:92. *Teachers College Record*.

Konstantopoulos, S., Modi, M., & Hedges, L. V. (2001). Who are America's gifted? *American Journal of Education, 109,* 344-382.

Lee, V. E. (2000). Using hierarchical linear modeling to study social contexts: The case of school effects. *Educational Psychologist, 35,* 125-141.

Lee, V. E., & Bryk, A. S. (1989). A multilevel model of the social distribution of high-school achievement. *Sociology of Education, 62,* 172-192.

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. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59*, 1-17.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models.* Newbury Park, CA: Sage Publications.

Raudenbush, S. W., & Wilms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics, 20,* 307-335.

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

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, NJ: Lawrence Erlbaum.

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin, 91*, 461-481.

White, S. W., Reynolds, P. D., Thomas, M. M., & Gitzlaff, N J. (1993). Socioeconomic status and achievement revisited. *Urban Education, 28*, 328-343.

Footnotes

1. 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 and found that they yield very similar results.

2. 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.
Trends of Mean Estimates of the Achievement Gap Over Time: Grade 12

	Reading				
	Year of Survey				
Variable	1971	1975	1980	1992	1996
Female	12.391*	11.828*	8.473*	11.654*	14.016*
Black	-36.018*	-35.987*	-37.001*	-25.153*	-20.820*
Hispanic	-	-22.314*	-15.259*	-14.531*	-16.631*
SES	0.788*	0.759*	0.782*	0.767*	0.770*

	Mathematics				
	Year of Survey				
Variable	1978	1982		1992	1996
Female	-6.550*	-5.626*		-3.771*	-3.662*
Black	-25.084*	-24.212*		-17.139*	-21.197*
Hispanic	-9.574*	-13.328*		-13.907*	-10.252*
SES	0.772*	0.766*		0.787*	0.798*

	Science				
	Year of Survey				
Variable	1977	1982		1992	1996
Female	-14.435*	-16.972*		-9.998*	-7.012*
Black	-40.000*	-50.536*		-33.147*	-33.783*
Hispanic	-14.872*	-33.543*		-20.875*	-17.625*
SES	0.760*	0.724*		0.779*	0.773*

* $p < 0.05$

Table 5.
Trends of Variance Estimates of Gender, Race, and SES Effects
Over Time: Grade 12

Reading				
Year of Survey				
Variable	1971	1975	1980	1992	1996
---	---	---	---	---	---
Female	2.93*	5.09*	3.54	7.07*	7.10*
Black	8.14	13.95*	6.63	13.44	14.73
Hispanic	10.50*	14.24*	8.51*	18.57*	12.71*
SES	0.10*	0.17*	0.09	0.25	0.33

Mathematics				
Year of Survey				
Variable	1978	1982		1992	1996
---	---	---	---	---	---
Female	3.20*	3.32		4.70*	4.12
Black	6.69	4.83*		11.13*	4.57
Hispanic	6.42*	6.07		11.18*	7.88
SES	0.16*	0.19*		0.26	0.34

Science				
Year of Survey				
Variable	1977	1982		1992	1996
---	---	---	---	---	---
Female	2.417*	3.81		8.358*	7.265
Black	3.973	6.505*		14.792*	8.405
Hispanic	6.420*	4.12		17.650*	9.766
SES	0.096*	0.14		0.202*	0.248

* $p < 0.05$

Table 6
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	% Hispanic-White Gap	% Parental Education Gap	% Gender Gap
Age 17						
20	3.6	0.44	12.5	12.2	11.9	25.0
30	6.3	0.76	21.6	21.1	20.6	43.2
40	8.5	1.03	29.3	28.6	28.0	58.6
50	10.6	1.28	36.6	35.7	34.9	73.1
60	12.7	1.53	43.8	42.8	41.8	87.6
70	14.9	1.80	51.5	50.3	49.2	103.0
80	17.6	2.12	60.6	59.2	57.8	121.2
90	21.2	2.56	73.1	71.4	69.8	146.3
95	24.2	2.92	83.5	81.6	79.7	167.0
99	29.9	3.61	103.0	100.6	98.2	206.0
Age 13						
20	3.4	0.44	10.8	12.5	11.7	26.1
30	5.9	0.76	18.6	21.5	20.2	44.9
40	8.1	1.03	25.3	29.2	27.4	61.1
50	10.0	1.28	31.5	36.4	34.2	76.1
60	12.0	1.53	37.7	43.6	40.9	91.2
70	14.2	1.80	44.4	51.3	48.2	107.3
80	16.7	2.12	52.2	60.4	56.7	126.2
90	20.1	2.56	63.0	72.9	68.4	152.4
95	23.0	2.92	72.0	83.2	78.1	174.0
99	28.3	3.61	88.8	102.6	96.3	214.5
Age 9						
20	4.5	0.44	15.6	18.2	20.4	42.1
30	7.8	0.76	27.0	31.3	35.1	72.5
40	10.5	1.03	36.6	42.5	47.7	98.6
50	13.1	1.28	45.7	53.0	59.5	122.9
60	15.8	1.53	54.7	63.5	71.3	147.2
70	18.5	1.80	64.4	74.7	83.9	173.2
80	21.8	2.12	75.7	87.9	98.6	203.7
90	26.3	2.56	91.4	106.1	119.1	245.9
95	30.0	2.92	104.3	121.2	136.0	280.8
99	37.0	3.61	128.6	149.4	167.6	346.2

Table 7
Effect of Moving a 10th Percentile School to a Given Percentile
 as a Percentage of Various Achievement Gaps
Estimated from 1996 NAEP Mathematics Data

Target Percentile	NAEP Scores	School SD Units	% Black-White Gap	% Hispanic-White Gap	% Parental Education Gap	% Gender Gap
			School Effect in Various Metrics			
Age 17						
20	3.2	0.44	12.0	15.1	9.0	70.4
30	5.6	0.76	20.7	26.1	15.5	121.4
40	7.6	1.03	28.1	35.5	21.0	164.9
50	9.5	1.28	35.0	44.2	26.2	205.6
60	11.3	1.53	42.0	53.0	31.4	246.3
70	13.3	1.80	49.4	62.3	36.9	289.9
80	15.7	2.12	58.1	73.3	43.4	340.8
90	18.9	2.56	70.1	88.5	52.4	411.5
95	21.6	2.92	80.1	101.0	59.9	469.9
99	26.7	3.61	98.7	124.5	73.8	579.4
Age 13						
20	3.3	0.44	11.3	12.9	11.3	84.4
30	5.7	0.76	19.5	22.2	19.4	145.4
40	7.7	1.03	26.5	30.2	26.4	197.6
50	9.6	1.28	33.0	37.7	32.9	246.3
60	11.5	1.53	39.5	45.1	39.4	295.1
70	13.5	1.80	46.5	53.1	46.4	347.2
80	15.9	2.12	54.7	62.4	54.5	408.3
90	19.2	2.56	66.1	75.4	65.8	492.9
95	22.0	2.92	75.4	86.1	75.2	562.8
99	27.1	3.61	93.0	106.1	92.7	694.0
Age 9						
20	3.7	0.44	14.5	16.6	18.5	94.4
30	6.3	0.76	25.1	28.6	31.9	162.6
40	8.6	1.03	34.1	38.8	43.3	221.0
50	10.7	1.28	42.5	48.4	54.0	275.5
60	12.9	1.53	50.9	58.0	64.7	330.0
70	15.1	1.80	59.9	68.2	76.1	388.4
80	17.8	2.12	70.4	80.2	89.5	456.6
90	21.5	2.56	85.0	96.9	108.0	551.3
95	24.6	2.92	97.0	110.6	123.4	629.5
99	30.3	3.61	119.7	136.4	152.1	776.2

Table 8
Effect of Moving a 10th percentile School to a Given Percentile
as a Percentage of Various Achievement Gaps
Estimated from 1996 NAEP Science Data

Target Percentile	School Effect in Various Metrics					
	NAEP Scores	School SD Units	% Black-White Gap	% Hispanic-White Gap	% Parental Education Gap	% Gender Gap
Age 17						
20	3.9	0.44	8.3	10.3	8.0	49.0
30	6.7	0.76	14.3	17.8	13.8	84.4
40	9.1	1.03	19.5	24.2	18.7	114.7
50	11.3	1.28	24.3	30.1	23.3	143.1
60	13.5	1.53	29.1	36.1	27.9	171.4
70	15.9	1.80	34.3	42.5	32.8	201.7
80	18.7	2.12	40.3	50.0	38.6	237.1
90	22.6	2.56	48.6	60.3	46.6	286.3
95	25.8	2.92	55.5	68.9	53.2	326.9
99	31.8	3.61	68.5	84.9	65.7	403.0
Age 13						
20	3.2	0.44	7.9	9.4	8.7	36.2
30	5.5	0.76	13.7	16.3	15.0	62.4
40	7.5	1.03	18.6	22.1	20.4	84.7
50	9.3	1.28	23.1	27.6	25.4	105.7
60	11.1	1.53	27.7	33.1	30.4	126.6
70	13.1	1.80	32.6	38.9	35.8	148.9
80	15.4	2.12	38.3	45.7	42.1	175.1
90	18.6	2.56	46.3	55.2	50.8	211.4
95	21.2	2.92	52.9	63.0	58.0	241.4
99	26.2	3.61	65.2	77.7	71.6	297.7
Age 9						
20	3.9	0.44	10.6	12.3	13.2	115.4
30	6.8	0.76	18.2	21.2	22.7	198.9
40	9.2	1.03	24.8	28.8	30.8	270.3
50	11.5	1.28	30.9	35.9	38.4	337.0
60	13.7	1.53	37.0	43.0	46.1	403.7
70	16.2	1.80	43.5	50.6	54.2	475.0
80	19.0	2.12	51.2	59.5	63.7	558.5
90	22.9	2.56	61.8	71.9	76.9	674.4
95	26.2	2.92	70.6	82.1	87.9	770.0
99	32.3	3.61	87.0	101.2	108.3	949.4