State NAEP Scores:  
The Relationship Between Expenditures and Academic Outcomes  
David W. Grissmer

Introduction

Over the past decade there has been a real and continuing crisis in consumer confidence in the public school system and with that have come all sorts of reforms and reproaches. We’ve increased resources and increased teacher and student accountability at all levels and attempted to measure success with a growing number of student assessments. It is time to ask what the nation as a whole has learned from reform efforts by various states and districts about how to improve the performance of our public school system—or indeed whether we can expect such improvement at all.

Two views of public education have predominated during this period of reform initiatives and implementation and will affect the next steps in education reform. The first view is that public education is basically unreformable. This outlook is supported by various research evidence showing that increasing the amount of money spent on education doesn’t translate into higher student achievement, but it is also influenced by one’s confidence in government programs in general. Public education is in fact public, and some people view it as one of many government bureaucracies with many interests to protect and

David W. Grissmer is Senior Management Scientist at RAND in Arlington, Virginia.
few incentives to change. These people support alternatives that create and sustain schools outside of the public system and stimulate competition within the public school system, such as school choice programs and charter schools. The widespread enactment of these kinds of programs indicates that a large number of stakeholders have doubts about the reformability of the public system.

The alternative view of public education is that it is reformable. This view holds that accountability can provide incentives, but that additional resources are needed—resources specifically targeted toward effective programs. A new report from The RAND Corporation suggests that when we invest money wisely in public education, we can in fact improve performance for some populations. This presentation will explain the research methods that led to these findings and examine how these findings could shape future reform efforts.

**In Support of the Relationship Between Expenditures and Outcomes**

Five years ago, the dominant view, at least in nonexperimental historical research, was that money just does not matter. Research studies could show no clear, consistent relationship between instances where we put money into schools and we got better outcomes. Three new reviews support different views of the expenditure-outcome relationship. Two of these reviews used the same original set of studies and came to different conclusions, and the third used a wider sample than the original studies. The new evidence finally supports the hypothesis that increased spending leads to improved performance in the schools. Specifically, the reviews show that spending more
money matters the most, and makes the biggest difference, for students in minority or economically disadvantaged groups.

**Teacher-Student Ratio**

One of the first reforms to show that specific ways of spending money may have a positive impact on student achievement was the Tennessee Experiment, a watershed project in education designed to determine the effect of class size on student performance. The state randomly divided about 6,000 kindergartners into two groups, placing one group of students in classes of 16 and the other group of students in classes of 24. The class sizes were maintained for the students over the next three years, and then all students went back to larger classes. The results showed significant achievement gains for the students from the small classes—and not only at the third-grade level. This same group of students, with just three years of being in a smaller class, showed higher achievement than their peers when they reached the eighth grade, higher high school graduation rates, and higher numbers of college-bound graduates.

We learned from this experiment that early intervention in class size could have a permanent effect on students. But we also learned about other factors in this equation. First, the gains from class-size reduction were only evident among students who had been in the smaller classes for three or four years. Students who spent only one or two years in a class of 16 did not show the same long-lasting positive results. Moreover, the greatest increases in performance were recorded among students in the smaller classes who were from minority or other disadvantaged groups. So not only did we have data that showed specific gains as a result of specific reforms, we also saw that these reforms had effects on very specific populations.
Tracking How Money Is Spent

Perhaps the single most persuasive piece of data that people from the “money doesn’t matter” school of thought are able to produce is this: If you look at per-pupil expenditures adjusted for inflation over the 25-year period from 1967 to 1992, they show an astounding 100 percent increase. Certainly no one is going to argue that student performance has increased significantly over this same period, so this has always been a very difficult statistic for believers in education reform to explain. But the new research suggests that this may be an inaccurate interpretation of the data. The report proposes that the calculation of the percent increase in per-pupil spending was not based on the right cost of living adjustment because most money in education is spent on labor and the cost of labor has not increased as quickly as the cost of capital. If you use an adjustment rate based on cost of labor increases from 1967 to 1992 instead, per-pupil expenditures show only a 60 percent increase.

Still, a 60 percent increase in per-pupil expenditures is a lot, and it is very easy to look at this statistic and wonder just how much that extra 60 percent has really mattered. The problem with trying to make these kinds of direct comparisons is that the very term “per-pupil expenditure” can be somewhat misleading. A lot of the money that schools get is not spent on the kinds of things that would be expected to produce higher test scores or other such quantifiable gains in each and every pupil. For example, a significant portion of a school’s budget these days is earmarked for special education programs. These programs typically affect only a small percentage of students, many of whom may not be required to take the standardized achievement tests on which student achievement statistics are
based. If you adjust the per-pupil expenditure figures to account for the expenditures that were not intended to produce academic gains, the increase becomes only 30 percent. This is an increase in funding that can be easily explained. In fact, most of this increase can be attributed to 1) increases in teacher hiring designed to reduce class sizes and 2) increased spending on compensatory programs for minority and disadvantaged students.

Who Gains the Most from Increased Expenditures?

National Assessment of Educational Progress (NAEP) test scores show achievement gains for all groups of students and in all subjects tested from the 1970s through the 1990s. A closer examination of these results shows that the achievement gains varied among students in different demographic groups. Gains for White students were very small overall, but minority students, including Hispanic students and economically disadvantaged White students, showed tremendous improvement.

This evidence, together with what we have learned from the Tennessee Experiment and the re-evaluation of expenditure statistics, leads to a new interpretation of the current state of the public education system and the value of past and present education reforms. Instead of thinking that we have been throwing vast amounts of money into education and have gotten lower Scholastic Assessment Tests (SAT®) scores and only slightly higher NAEP scores as a result, we now recognize that we have really only invested modest sums into education. Moreover, much of the additional money went toward programs that were designed to elevate the performance of select groups of students, including low-income and minority students, and that is precisely where we are now seeing the greatest results.
Variables in the Expenditures-Outcome Equation

The evidence that public education is reformable, and that certain individual reforms are achieving success, has been accumulating. RAND set out to test this hypothesis further by conducting a comprehensive study of NAEP results. We focused our research on three specific questions:

- Do states that spend more money on public education have higher student achievement?
- Does how and where the money is spent affect student achievement?
- Is there evidence that the state reform efforts that have been going on since the mid-1980s are having an impact on student achievement?

The main variable in and basis for comparing education reforms in our study is the state, not individual schools or districts. This is because the states drive educational policy by furnishing about half of all resources provided to public education. The states have the financial leverage to reform education, and because most legal challenges to educational policy are litigated in the state—not federal—courts, the states have judicial leverage as well. Differences among schools across the nation are therefore likely to originate at the state level.

Variables in Public Education at the State Level

Fortunately for researchers studying education reform, there is a lot of variation in education systems at the state level. Different states have very different education reforms, funding levels, and ways of spending their money. We found that there are four main ways that states may spend money in education, each of which will affect the per-pupil expenditure figures. First, states
can use their resources to hire more teachers and thus lower the pupil-teacher ratio. Second, states set teacher salaries at different levels, even when the regional cost of living is factored in. Third, states may provide their teachers with more resources for teaching, meaning anything from larger classrooms to the numbers of computers or books available. And fourth, public prekindergarten programs add to the student and teacher populations of school systems but may be funded at different levels across states. These four variables account for 95 percent of the disparity in the way money is spent on education across states and therefore comprised the focus of our analysis.

Variation Among States

The degree of variation among states in terms of these factors can be surprising. Overall, per-pupil expenditures, adjusted for regional cost of living differences, range from $4,000 to $9,000. That is a huge difference, which should allow us to track relatively easily its effects as a factor in student achievement. The range in pupil-teacher ratio is just as startling: some states average 14 students per teacher, while in others the number is as high as 26.

Adequacy of teacher resources is another major variable and is something that teachers themselves are able to report on through the NAEP. The percentage of teachers who give the lowest of three possible ratings to describe the adequacy of resources available to them ranges from 17 percent to 57 percent across states. So in some states this is a minor concern, while in others more than half of the teachers surveyed felt their resources were very inadequate. And in terms of teacher salary, the average ranges from $28,000 to $44,000 across states, adjusted for cost of living differences.
These discrepancies are also evident in the last category, prekindergarten expenditures. We studied statistics from the 1980s through the mid-1990s and found that in some states 31 percent of preschool-age children were enrolled in public prekindergarten programs. In most states, however, such programs did not exist or included a negligible number of children. The real and significant differences we discovered in all these spending categories reveal what might be called a nationwide natural laboratory in education reform, the experiments and subjects of which are ripe for study and analysis.

**Characteristics and Results of the NAEP State Program**

RAND decided to use the results from the new NAEP state program as the primary source of data for its study because it is the only national standardized assessment that provides a valid basis for comparison between and among states. Unlike the national and long-term trend NAEP tests, which have sampled student skills at the fourth-, eighth-, and twelfth-grade levels in about a dozen subject areas since 1969, the NAEP state tests have been administered biennially only since 1990 and only to fourth and eighth graders in three fields: mathematics, reading, and science. Participation in all NAEP assessments is voluntary, with between 35 and 45 states and U.S. jurisdictions typically choosing to administer each test and representative samples of about 2,500 students from each participating state selected to take the test. The results of the state mathematics and reading assessments administered between 1990 and 1998, a total of six tests, formed the core of our study. What makes this an especially useful sampling is that while the tests and the number of student examinees were standardized across states during this time period, the criteria for excluding students from testing were also fairly uniform.
Geographical Variation

Not surprisingly, the results of the NAEP tests show a number of correlations, or trends. One notable trend is that states that score high on any one of the tests—say, for example, fourth-grade reading—are also more likely to score high in other assessment areas. But perhaps the most striking trend, at least when states are ranked and listed in order of their raw scores, is purely geographical. Generally speaking, we see northern rural states at the top of the list, some northern urban states in the middle, and southern states at the bottom of the list, signifying the poorest results (Figure 1).

This is not, however, a particularly telling observation, at least in isolation. Are we meant to infer that residents of colder states have nothing better to do in the winter months than homework and studying? Or that perhaps, as New York’s Senator David Patrick Moynihan has jokingly suggested, the pattern is attributable to some as-yet-undefined Canadian influence? In order to discover the true basis of this trend, we need to investigate other factors in student performance, and that is where demographic variation comes in.
Demographic Variation

If we take the same list of states ranked by student performance on the NAEP state tests and use a visual aid such as color to indicate demographic makeup, the resulting pattern is just as
evident as the geographical distribution. We could use almost any social indicator as a criterion—family income, parents’ educational levels, age of mother at child’s birth, racial or ethnic background—since they are all related. We chose to use the percentage of minority students in each state’s student population. In Figure 2, then, states with a low minority percentage are represented by darker bars and those with a high minority percentage are represented by lighter bars. And what do we find? The top of the list, where there are the most northern rural states, is also where we see the most dark bars, whereas most of the lighter bars are clustered at the bottom of the list where we find the low-achieving southern states.

This tells us that demographics are very important variables when it comes to explaining student achievement. In states such as Iowa, Montana, and Maine, which all scored high on the NAEP tests, performance may be a factor of prevailing family traits rather than other educational variables. The same is true of the lower achievers: what we see when we look at their raw scores may be family characteristics, not score characteristics.

Using only NAEP statistics, however, makes it difficult for this trend to come to the surface. The demographic data that NAEP collects come solely from student responses, and we discovered that these can be somewhat unreliable. For example, in a study of fourth graders, 50 percent reported that their parents had graduated from college, when in fact only about 20 percent had. For this reason, RAND undertook its own research to obtain reliable and accurate data about the educational and ethnic backgrounds of students’ families in each state.
The Importance of Factoring in Family Characteristics

It is important for researchers to remember that raw scores on standardized tests are still the most important numbers for individual students and their families. These numbers may well determine what classes a student is enrolled in, what colleges the
student can go to, or what jobs the student is able to get. In this way, such scores can have a large and lasting impact on the lives of individuals of all backgrounds and geographical areas.

But as a measure of the quality of schools, comparisons of raw scores at the state, district, or school level are not good enough. The only valid conclusions about how well schools are performing must be based on comparisons between similar student populations in different schools. Demographic variations can be huge, and states have no control over them. In some states minority students make up less than 1 percent of the school population; in others, the number may be as high as 50 percent. The percentage of parents who have a college degree ranges from 17 percent to 40 percent. Just as we need to control for regional standard of living differences when we compare teacher salaries in different states or for inflation when we compare per-pupil expenditures from different years, we need to control for family characteristics when we compare students from different schools. Only then will we really know what the NAEP student assessment data tell us about the effects of increased expenditures and other education reforms on children in the public schools.

Leveling the Playing Field

RAND used census data and extracted information for fourth and eighth graders to determine the demographic characteristics of the student populations in each state for which NAEP results were available, essentially allowing us to level the playing field for these students. In the original NAEP data, the disparity between the highest- and lowest-scoring states was about 25 to 30 percentile points. When we take family characteristics into consideration, however, the gap closes to about 12 percentile points. That is the real difference between schools, the difference
we want to focus on to determine what works and what does not in education reform. Among students with similar families, Texas, Wisconsin, Montana, Iowa, Maine, Indiana, and New Jersey score the highest, so there is still a slight geographical trend evident, but there are now many more exceptions. What we find when we investigate the reason for the score differences between states at the top of this new list and states like California at the bottom is that the scores are closely linked to schools’ resources.

**What the Raw Data Show**

One relatively easy way to compare groups of similar students is to take the NAEP raw scores and break them down by socio-racial status. A comparison of scores on all NAEP state tests from 1990 through 1998 for non-Hispanic White students from Texas and California (Figure 3) shows that the Texas students outperformed their California counterparts on every one of the tests, typically by 5 or 10 percentile points. On the same set of tests Black students in Texas scored higher than Black students in California (Figure 4), and the trend was even more striking among Hispanic students, with Hispanic students in Texas outscoring Hispanic students in California by a 12 or 14 percentile point margin (Figure 5). In fact, Texas students in all three major socio-racial groups scored as well as or higher than their counterparts in the northern state of New York.
Figure 3
White Students in Texas Have Higher NAEP Scores than White Students in California on Every Test

Figure 4
Black Students in Texas Have Higher NAEP Scores than Black Students in California on Every Test
Figure 5
Hispanic Students in Texas Have Higher NAEP Scores than Hispanic Students in California on Every Test

Explaining the Differences in Performance

The obvious question when we look at the disaggregated data from the Texas and California NAEP tests is why students from similar socio-racial backgrounds still show such difference in their performance on the tests. Our research showed that there are three variables that account for almost all of the disparity. The first is class size. Certainly class-size reduction has been a major reform effort in California over the past decade, but at the time of the tests in the study (late 1980s through early 1990s) class sizes in California were the largest in the nation. Texas, on the other hand, had an average of only 18 students per class, which was in fact slightly below the national average. Another difference between the Texas and California public schools at that time was the availability of public prekindergarten programs. Because Texas billionaire H. Ross Perot had started a subsidized public prekindergarten program in Texas in 1985, almost one quarter of prekindergarten-age children were enrolled in the program. Again, the same could not be said of
California. And finally, another factor in the score discrepancy may be traced to the adequacy of educational and educator resources. In the NAEP survey, more teachers in Texas than in California reported having adequate resources. RAND estimates that if California could duplicate the scholastic conditions of Texas in terms of pupil-teacher ratio, prekindergarten access, and resource availability, the difference in student raw scores would decrease by about two-thirds.

A Look Behind the Numbers: What’s Working?

When RAND investigated why there was such variation between scores for students from similar family backgrounds in different states, we discovered a number of educational policy differences, some of which correlated with higher test scores and some of which, perhaps surprisingly, did not. For example, we could not find any relationship between student achievement and the number of teachers with advanced degrees in the state. When we analyzed teacher salary statistics we found a somewhat ambiguous relationship. But this relationship can be explained with some common-sense thinking: prospective teachers typically will not be able to choose which state they will teach in, whereas they are much more likely to choose which district they will work in based on salary level. Therefore, while we would expect teacher salaries to have a noticeable effect on student achievement at the district level, they do not seem to be a factor at the state level.

Effective Spending

The educational policies that do appear to have an effect on students’ test scores are of course not very surprising. Overall, we found that among students from similar family backgrounds
those who come from states with higher per-pupil spending do better on the NAEP. Just as importantly, the higher-scoring states tend to spend their funds in certain ways. Specifically, they tended to put more money into programs designed to lower pupil-teacher ratios in the lower grades, support public prekindergarten, provide teachers with adequate resources, and reduce the need to hire inexperienced teachers (e.g., by reducing teacher turnover).

We also found that the size of the effect of any one of these factors was related to the demographics of the school-age population in the state. For example, a state with a high proportion of disadvantaged students could expect to see a six-percentile-point gain in their raw scores by reducing class sizes from 25 pupils per teacher to 22 pupils per teacher. States with a low proportion of disadvantaged students, on the other hand, could expect to see a gain of only two percentile points after making the same pupil-teacher ratio reduction. We have an equation for this; the gains decrease as class size decreases from 22 to 19 students, from 19 to 16 students, etc., and at some point any achievement gains become negligible. The same principles apply to support for public prekindergarten programs. Focusing resources on disadvantaged students will have the most significant effects.

Cost-Effectiveness

Based on these data we were able to make cost-effectiveness evaluations as well. This allows us to predict not only how much it would cost a state to reduce its pupil-teacher ratio by one or to enroll a certain percentage of children in a public prekindergarten program, but also what kind of achievement gains the state could expect to see as a result of these measures. The equation also works in reverse: How much would a state have to
spend to raise students’ test scores by three percentile points? It
depends on the proportion of disadvantaged students in the
state. A state with a high proportion of disadvantaged students
that puts its money into lowering pupil-teacher ratios in grades 1
through 4 could expect to see the three-percentile-point gain
after spending an additional $150 per pupil across the state. In a
state with an average number of disadvantaged students the
figure would be about $450 per pupil, and a state with a low
proportion of disadvantaged students would have to spend as
much as $1,000 more per pupil to get the same results. The
same is true of prekindergarten education: it might cost a state
anywhere from $120 per pupil to over $1,000 to raise scores by
three percentile points. The only factor that is an exception to
this pattern is teacher resources. Educators across the nation will
be happy to hear that providing more resources to teachers
appears to benefit all students and is therefore cost effective
across the board.

Results of State-Specific Achievement Tests

The first challenge to the results of our study and the numbers
that I’ve presented here came last year when some state-
produced tests showed tremendous student achievement gains
not recorded on the state NAEP tests. For instance, in Texas,
students’ scores on the Texas Assessment of Academic Skills
(TAAS) have gone up much faster than their NAEP scores.
Similar results were also reported in Kentucky, Maryland, and
California. Does this invalidate our NAEP-based findings? I do
not believe so. I think that the reason the students are
performing differently on the state-specific tests is that these
tests are trying to do something different from a national testing
program like the NAEP. Whereas the NAEP tests are based on
an average curriculum for all fourth and eighth graders in the
country, state tests such as the TAAS are testing students on the content specified in the state curriculum and state learning standards. Students are therefore more likely to have covered the material on the state tests in class.

**Positive Indicators**

Another reason for the increases may be that some states are implementing the very reforms we have found to be cost-effective factors and correlative to student achievement. Many of the states that experienced score increases in the state-specific tests were the ones whose educational resources were on the rise, but this alone could not explain the results. Instead, these states may be providing evidence of education reform that is working. States that started their reform initiatives and/or implemented standards-based reform early on, such as Texas, North Carolina, Indiana, and Kentucky, are the states whose students are now showing among the highest achievement gains.

**A Mandate for Reform**

The recent state-specific test results also serve to highlight the need for another type of reform in resource allocation. Our report focuses on what NAEP and demographic data tell us about how states can spend money wisely to promote student achievement, but we must not lose sight of the fact that one of the most serious reasons for inequality in public education across states is not under state control. The problem is national. Individual states may try, through the court system, to equalize funding for all districts in the state, and this effort may in turn produce district-level achievement gains. But in states with low spending totals, no amount of intrastate resource allocation is going to allow the schools to compete with those in states with large school budgets. The federal government, therefore, must
play a role in trying to address the inequality between states. Most of the disadvantaged children, those who could profit the most from education reform, live in states with low funding levels. And until we address this problem, we are going to leave a lot of kids behind.

Conclusions

Public education is reformable. States that have initiated reforms are now scoring higher on state achievement tests and, when results are controlled for family characteristics, on the National Assessment of Educational Progress (NAEP). Our research now shows that additional spending on education and wise resource allocation, both essential characteristics of education reform, can and do make a difference, especially among disadvantaged student populations. Teacher salaries do not have a significant impact; neither does the number of advanced degrees a teacher has earned.

What has the largest, most direct effect on student achievement is the working conditions of a state’s teachers. The best educational policy is one that helps teachers help their students, especially otherwise disadvantaged students, either by reducing class size, providing teachers with adequate resources, or giving children a head start by funding public prekindergarten programs. The amazing thing about the Tennessee Experiment is that in 1986 in Tennessee teachers with no extra professional development came to school and were given classes of either 16 or 24 students, and at the end of the year the students in the smaller classes had higher scores. The observational data in this case indicate that the students were more on task in smaller classes, probably because teachers could afford to spend more time with individual students, which we know will especially
help disadvantaged children. I believe that the same kind of
cause-and-effect relationship drives all the reforms that improve
student performance. Whenever we give teachers better working
conditions, whether it be by reducing class size, providing
resources, or giving them better-prepared, prekindergarten-
educated children, which in turn makes discipline easier, the
teachers know what to do to help their students learn, and this
shows up on the achievement test scores.

We have a lot of work to do, but the research shows that the
states’ efforts are paying off. The cost-effectiveness equation that
we developed can help states estimate the score increases that
would result from different levels of increased funding and may
help states justify spending increases for education, but the real
benefit of this study has been to identify factors that can promote
student achievement. I hope that this evidence will convince the
nation not only that public education is reformable but that
education reform is worth working for.