School Accountability and Student Performance

Eric A. Hanushek and Margaret E. Raymond

The introduction of student accountability systems across the United States has been controversial both because of its focus on standardized achievement tests and because of questions about its effectiveness. Past evidence, however, shows that performance on standardized tests of the type central to state accountability systems has powerful economic effects. Additionally, analysis of performance across states indicates that accountability policies in general lead to higher levels of achievement, though the magnitudes of the effects are influenced by the design of the policy. Finally, however, despite positive effects overall, recent work shows that these policy instruments are not effective at closing the black-white achievement gap.

Federal Reserve Bank of St. Louis *Regional Economic Development*, 2006, *2*(1), pp. 51-61.

Since the passage of the No Child Left Behind Act (NCLB), a common question has been: Is it working? Of course, analyzing the overall impacts of NCLB is difficult if not impossible. The policies are very recent. But, more than that, there is no obvious comparison group because all states fall under the purview of NCLB. Nonetheless, because many states had previously introduced their own accountability systems—systems that became the heart of most states' responses to NCLB—it is possible to examine these states' experiences and infer many of the overall effects of the federal legislation.

This paper presents a nontechnical overview of the findings of analyses of state accountability. It summarizes three central results:

- Performance on typical "state accountability" standardized tests is tied directly to economic effects;
- accountability policies in general lead to higher levels of achievement, though the magnitudes of the effects are influenced by the design of the policy; and,
- despite positive effects overall, recent work shows that these policy instruments are not effective in repairing existing disparities in performance by race.

THE IMPORTANCE OF SCHOOL QUALITY

Much research on how schooling affects individual earnings has focused merely on attainment, or the *quantity* of schooling, but more-recent research has turned to issues of *quality*. This alternative focus is consistent with the current attention policymakers are paying to student testing and accountability in the United States, United Kingdom, and elsewhere.[1]

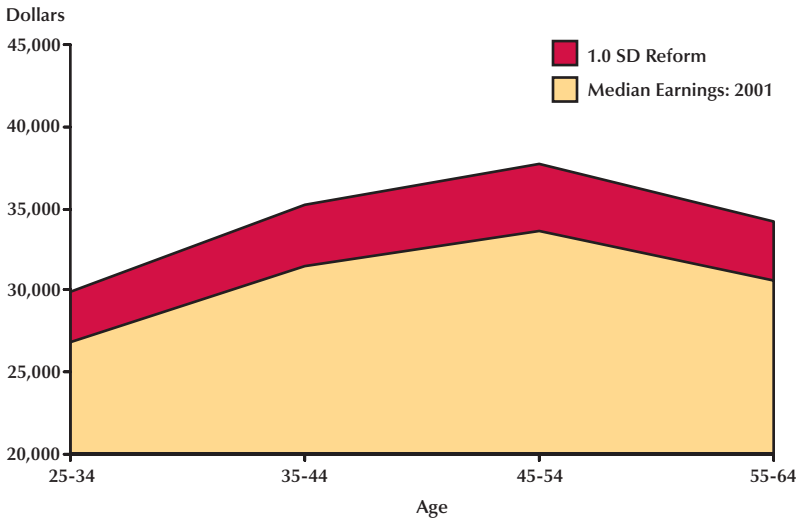
Recent research in the United States shows that the quality of schooling relates to real differences in earnings and attainment. Three recent studies provide direct and quite consistent estimates of the

[1] A more complete discussion of the issues in this section can be found in Hanushek (2004): www.hanushek.net.

Eric A. Hanushek is the Paul and Jean Hanna Senior Fellow, and Margaret E. Raymond is the director of CREDO (Center for Research on Education Outcomes) at the Hoover Institution, Stanford University.

Figure 1

Median U.S. Individual Earnings with 1.0 SD Reform



NOTE: SD is standard deviation.

impact of test performance on earnings (Mulligan, 1999; Murnane et al., 2000; and Lazear, 2003). These studies use different nationally representative data sets that follow students after they leave school and enter the labor force. When scores are standardized, they suggest that a 1-standard-deviation increase in mathematics performance at the end of high school translates into 12 percent higher annual earnings.[2]

Figure 1 graphically portrays the impact of higher quality of schooling: Comparing the median earnings in 2001 of a typical individual in the United States with the amount they would earn if the quality of their schooling had been 1 standard deviation higher (i.e., if their measured achievement changed from the 50th percentile to the 84th) shows that expected earnings shifts upward by some 12 percent each year throughout their working career. Although the research is less extensive, similar or larger magnitudes of earnings improvement have been found in other countries.

Moreover, although not shown in this figure, there are additional gains that would accrue because individuals with greater skills tend to continue farther in schooling—that is, to have higher school attainment. Murnane et al. (2000) separate the direct returns to measured skill from the indirect returns to more schooling and suggest that perhaps one-third to one-half of the full return to higher achievement comes from further schooling. (Figure 1 shows just the direct effects of skills, not including the indirect effects from added schooling.) Note also that the other side of increases in school attainment from quality improvements is a decrease in school drop-out rates. Specifically, higher student achievement keeps students in school longer, which will lead to, among other things, higher graduation rates at all levels of schooling.

Another place to look for the economic impact of school quality is the effect on the growth in

[2] Murnane et al. (2000) provide evidence from the "High School and Beyond" survey and the national longitudinal survey of the high school class of 1972. Their estimates suggest some variation: male students obtain a 15 percent increase and female students a 10 percent increase in earnings per standard deviation of test performance. Lazear (2003), relying on a somewhat younger sample from the national education longitudinal study of 1988, provides a single estimate of 12 percent. These estimates are also very close to those in Mulligan (1999), who finds 11 percent for the normalized Armed Forces Qualification Test in the national longitudinal survey of youth data. By way of comparison, estimates of the value of an additional year of school attainment are typically 7 to 10 percent.

Figure 2

Effect of Economic Growth on U.S. Income

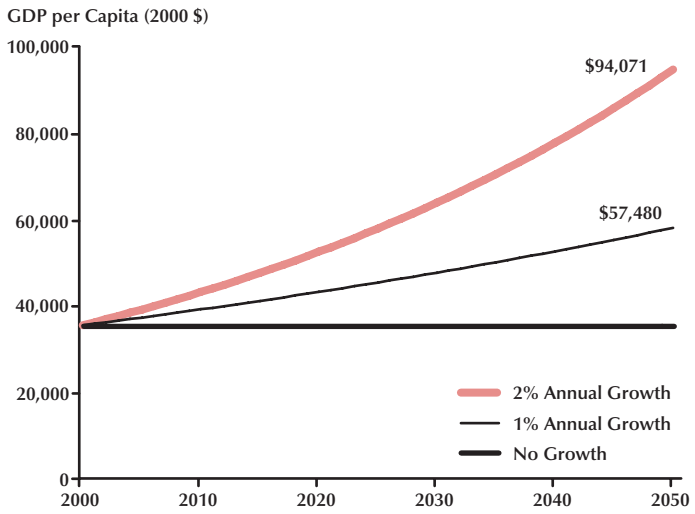
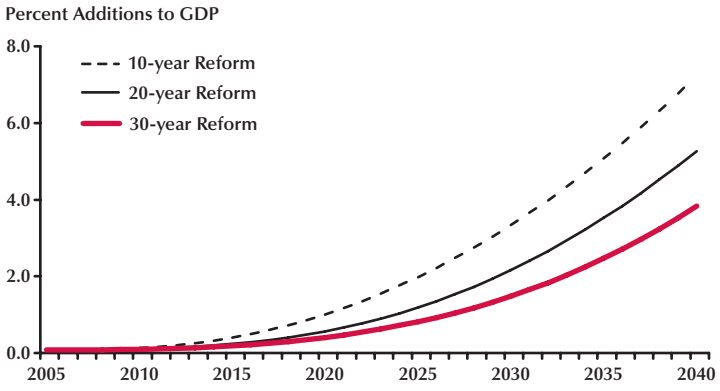


Figure 3

Improved GDP with Moderately Strong Knowledge Improvement

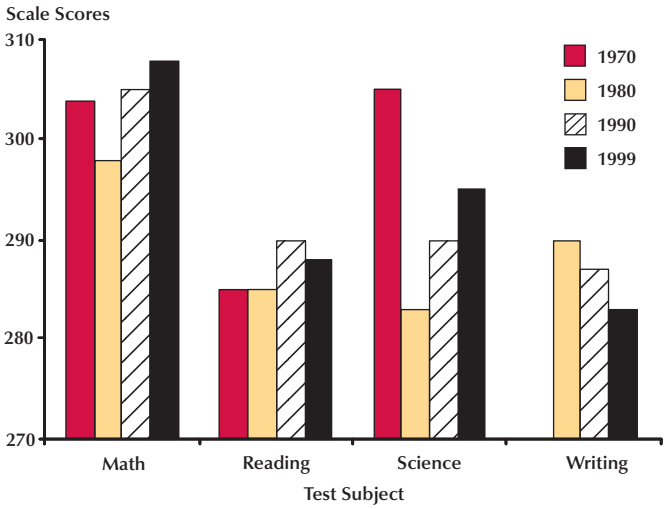


national income. Economists have demonstrated that productivity gains that are directly related to human capital fuel increases in the gross domestic product (GDP) of a nation. GDP growth, in turn, is what improves the standard of living for its citizens. Furthermore, the benefits of productivity growth compound over time, to dramatic effect. With U.S. economic levels as shown in Figure 2, if the economy grew by 1 percent per year starting in 2000, GDP per capita would increase by 65 percent by 2050. Were the economy of the United States to grow at 2 percent per year, the GDP per capita would go from roughly $35,000 to over $94,000.

Research on how school quality affects growth shows that a 1-standard-deviation increase in student achievement (moving from the 50th to the 84th

Figure 4

National Assessment of Educational Progress (NAEP), Age 17



percentile) translates into 1 percent faster growth (Hanushek and Kimko, 2000). That is, after allowing for any other factors that might affect growth, improvements in student outcomes have a very powerful impact on growth, leading to the kind of gains found in Figure 2.

The pattern of economic effects depends on two factors: the size of achievement gains and the speed with which they are obtained. The faster the United States introduces quality-oriented education reforms, the faster it will be able to realize the benefits offered by such an approach. Consider the effects of achieving a moderately strong gain in knowledge, as measured by moving from the median to the 69th percentile (i.e., 0.5 standard deviations) over differing time horizons. Figure 3 illustrates the impact on the level of GDP arising from moderately strong gains in knowledge over 10-, 20- and 30-year time frames. If it takes 30 years to achieve that level of improvement, the GDP in 2040 will be approximately 4 percent higher than it otherwise would be. This gain in GDP would essentially pay for all primary and secondary expenditures. In other words, the growth dividend *from true reforms that led to real student achievement gains* would make schooling free. If those quality gains can be realized in 20 years, then the compounding is more pronounced and 2040 GDP will be greater by 5 percent. With a 10-year horizon for improvement, the GDP gain in 2040 will be nearly 7 percent. Reaching higher achievement levels in a shorter period of time is clearly more difficult, but it yields compensating gains.

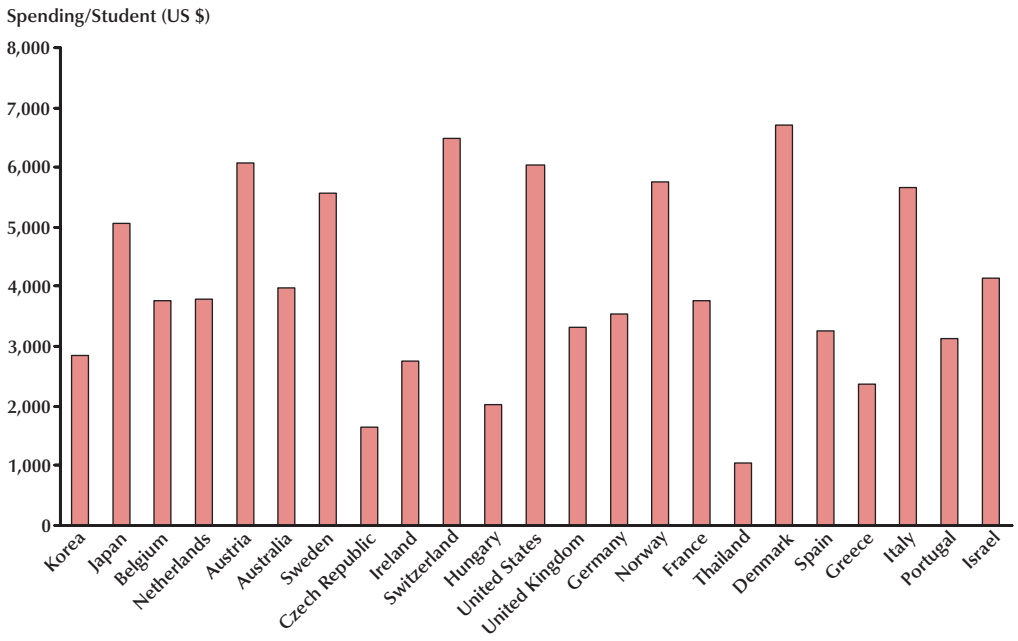
RESOURCE RICH, RESULTS POOR

Despite decades of effort, no resource-oriented policies have achieved results that are as significant as those described here. The evidence is consistent across many countries—in the United States and foreign nations, in developed countries and developing ones—that "throwing money at schools" does not result in improvements (see Hanushek, 2003).

We have learned this lesson with difficulty in the United States. We have witnessed large growth in teacher investments, as measured by the share of teachers with Master's degrees and decreased pupil-teacher ratios. We have more experienced—and thus more highly paid—teachers than in past decades. And perhaps most dramatically, we have tripled our average real per-pupil spending since 1960.

Figure 5

TIMSS Performance and Spending (Countries Ranked by TIMSS Aggregates)



But the rewards are slim to none. As shown in Figure 4, U.S. performance on the National Assessment of Educational Progress (NAEP) has gained only slightly in reading and math and has actually declined in science and writing. This is hardly a sterling endorsement for increasing spending further.

The international picture is similarly unsupportive. If resources were significantly and positively related to performance, one would expect to see that countries who scored the highest in the Trends in International Math and Science Study (TIMSS) would spend the most and that lower-performing countries would spend less. However, as laid out in Figure 5, which ranks countries by TIMSS performance, no such pattern exists. Of note, the United States is among the countries with the highest expenditures but ranks near the middle in terms of performance.

FOCUS ON ACCOUNTABILITY

Over the past decade, a sea change has occurred in the design of education policies in many countries around the globe.[3] Policies have shifted from attending to inputs and processes to a focus on the outcomes realized by students. The change has emerged through the widening practice of testing students against a common set of expectations about learning objectives for each grade. Thus, standards, testing, and accountability go hand in hand.

Where countries have a single education administration, as in Taiwan or the United Kingdom, students often face national exams. Countries with federal systems of government in which education is a federal responsibility operate in similar ways. In the United States, the responsibility for education resides in the 50 individual states. Over the past 10 years, states have adopted their own policies at different times, which created a diversity of accountability policies and testing programs as well as different adoption dates. States differed also in the use of rewards and sanctions. Figure 6

[3] The explicit modeling of accountability is fully developed in Hanushek and Raymond (2005). This section relies on the results in that study.

shows the pattern of the adoption of accountability systems by states. It also shows the division between "report card" states (those simply reporting results to the population) and "consequential" states (those attaching varying rewards and sanctions to school performance).

Not surprisingly, the adoption of accountability policies has produced a range of education outcomes as well.

The closest thing to a national examination in the United States is the NAEP. The program is designed to test a representative sample of students in 4th, 8th, and 12th grades in reading, mathematics, science, and other subjects regularly. Starting in 1992, the methods used to select student samples were intended to provide representative results at the state level. Participation, until recently, was voluntary for individual states, so states could test in only one subject or restrict the grades that were tested. Still, it is the only available common measure of performance across states.

In recognition of the heterogeneity of student results, the U.S. Congress in 2001 passed sweeping education reform legislation, the No Child Left Behind Act. Although not completely standardized, NCLB pushes toward a common practice on accountability throughout the United States. Although the law respects the states' rights to design both education policies and standards/testing policies, it requires each system to test students annually, requires all states to report on a limited set of performance metrics, and introduces a common set of consequences for schools that fail to show acceptable results. The policy also requires states to establish their own standards of proficiency using their state standardized test, though the actual thresholds of "below proficient" and "proficient" may differ across states.

We are able to capitalize on the staggered adoption and diversity of accountability programs to study in a general way the effect of this important change on student performance. (Clearly, the NCLB has equalized the program characteristics and, thus, has ended the national accountability experiment of state-level differences.) Combined with the periodic scores reported on NAEP tests, which were given every four years throughout our evaluation period, three research questions can be addressed:

- Does accountability work?
- Are the impacts common for all subgroups?
- Are there policy attributes that affect results?

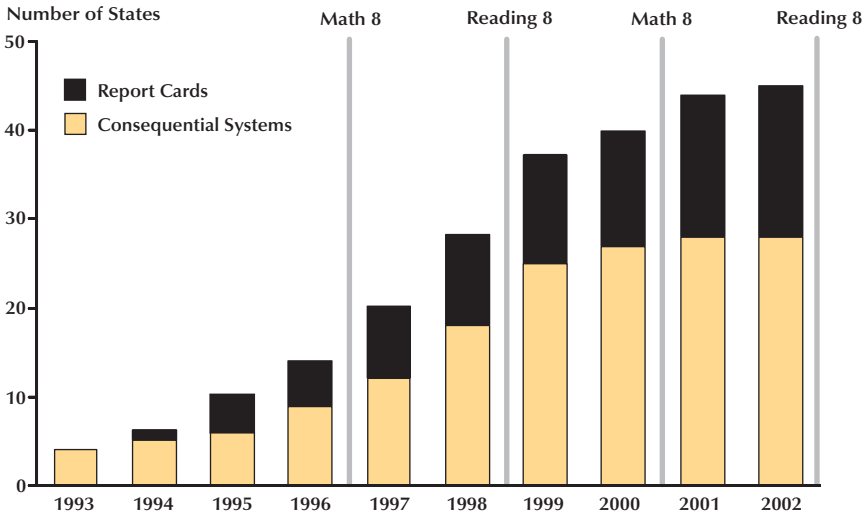
It is important to note that the analysis is limited in several respects. Some states did not adopt an accountability system at all until required by NCLB, thus limiting the observations of accountability effects. Second, state participation in NAEP is voluntary; so, even among those states with accountability policies, data were lacking for some grades in some years. States also differed in their decisions to exclude students on the basis of disability, language proficiency, or time since entry into a school from taking the test; accordingly, there is some mixing of students across states and over time within states. Finally, accountability was not the only reform initiative that states implemented over the study period, but the impacts of these other initiatives are difficult to isolate.

How Well Do Accountability Systems Work?

Accountability policies have two general characteristics: They provide performance information about a school in a consistent way, and they require all schools to face similar treatment based on their results. States create an aggregate score for each school based on individual student test scores. Because states differ in the way they measure school performance, they may send different signals to schools and, ultimately, promote different policy results. Our basic assessment of existing accountability systems does not distinguish among design features of different states, although later we suggest that the designs are very different and are likely to affect performance more or less strongly. The school score is then used to determine the performance of schools against some pre-set criteria. These evaluative ratings are intended to provide feedback and offer objective motivation to spur improvement. Second, what happens to schools once they obtain their scores differs by state. As noted previously, some states merely make the information public (known as report card states), whereas others introduce consequences in the form of rewards and/or sanctions. The current analysis looks at the impact of having consequences to test whether the design

Figure 6

State Accountability Over Time



NOTE: Gray bars indicate NAEP testing dates.

characteristics of state accountability systems matter. We return later to issues of overall design.

Conditional Consequences. Accountability programs differ in how they use accountability scores, and such differences may influence the effectiveness of the program. Earlier research identified two general approaches. The first uses public disclosure to motivate interested parents, school boards, the media, and civic leaders to demand better performance from low-scoring schools. This approach relies on release of scores over the Internet and publication and comment in local papers. The second and more direct approach incorporates into the accountability program's design a set of consequences—typically monetary awards, Blue Ribbon designations, or punitive actions such as probationary status or threat of reconstitution—to prompt schools to improve.

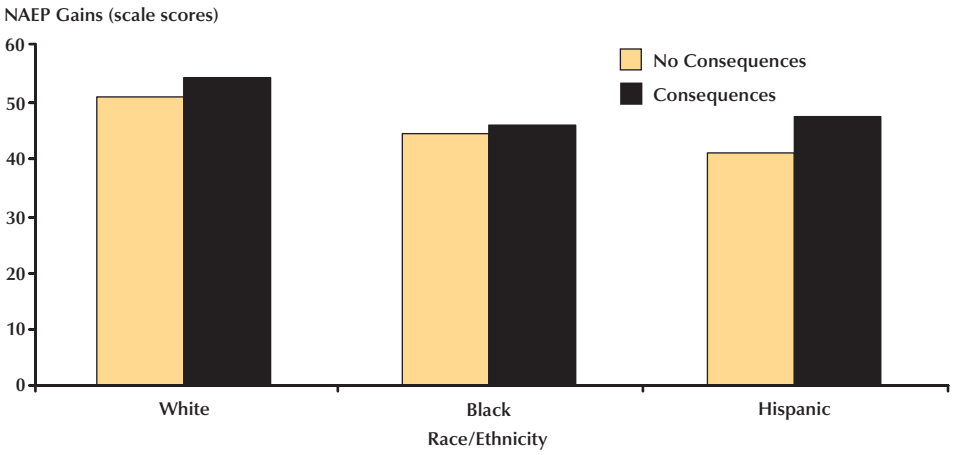
As shown in Figure 6, between 1993 and 2002, 43 states adopted accountability programs. Of these, 29 programs included consequences and 14 used a report card approach. The markedly different mechanisms of influence provide the chance to study whether this design feature is influential in the educational improvement of states. Consideration of the type of accountability system was incorporated into the overall test of the effectiveness of accountability to which our discussion now turns.

Modeling the Effectiveness of Accountability. The availability of NAEP test results on repeated administrations of the test provides a unique opportunity to examine the staggered adoption of accountability policies across the states and to test their impact on the rate of improvement in student academic achievement. Conveniently, NAEP tests students in the 4th and 8th grades in reading and math every four years; so, for states that test students in both grades, over time the same cohort is captured as it moves through school.

For both reading and math, we can test the progress of two cohorts of students in states participating in the NAEP. As noted in Figure 6, we can use 8th grade math scores in 1996 and 2000 and reading scores in 1998 and 2002 (combined with 4th grade scores four years prior). As long as we can control for cohort differences in family background (e.g., parental education, race/ethnicity, poverty), average state education spending, and testing exclusions over the period, the growth in achievement across cohorts can be compared for

Figure 7

Effect of Consequential Accountability on Achievement by Race/Ethnicity



states that had adopted accountability over the period of study against states that did not. We further exploit the disaggregation of NAEP results by race and ethnicity (white, black, and Hispanic). We pool the disaggregated state test data for both reading and math.

We also consider the difference in the system design (consequence vs. report card) and a fixed-state effect to reflect any other policy changes that the state might have adopted to improve student performance. Multivariate econometric modeling was used to discern the impacts of the factors we examined. (The full models estimated are reported in Table A1 of the appendix.)

The overall difference in performance between 4th and 8th grades that comes from accountability is displayed in Figure 7. For each group of students, the expected growth in achievement is higher in states that implement accountability systems than in states that do not.

The improvement was realized by states that attached consequences to schools' performance. However, states with "report card" accountability programs had no significantly different achievement levels from those of states without any accountability program.

Other results are also noteworthy. Testing exclusion rules were negatively significant—the more students excluded, the better the results; nonetheless, exclusion rates vary across states in a way that does not affect the estimated importance of accountability. Differences in per-pupil spending were not significant in explaining the differences in learning gains. This latter finding is consistent with a large body of earlier work; in this case, the finding provides especially important insight because many states face pressure to dramatically increase spending to promote better learning.

At the same time, by comparing the gains for each group, it is clear that accountability has different impacts on the groups. The overall differences are shown in Figure 8, which identifies the black-white and Hispanic-white achievement gaps both with and without accountability. The comparisons (measured in standard deviation units) show that accountability closes the gap for Hispanics but widens it for blacks.

DESIGN ISSUES

Although each adopted its accountability system independently, states copied student testing and school scoring design from each other.[4] Although

[4] A more complete discussion of these design issues is found in Hanushek, Raymond, and Rivkin (2004).

Figure 8

Racial/Ethnic Gaps by Consequential Accountability Status: NAEP Gains Relative to White Students

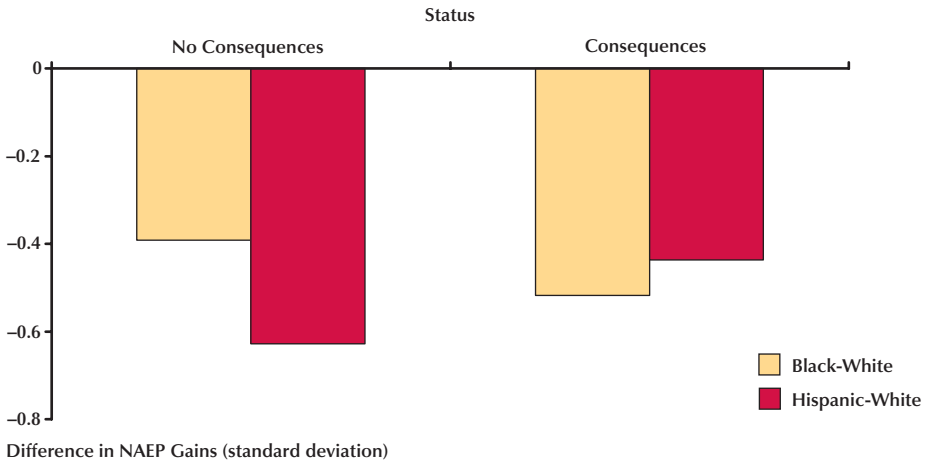


Table 1

Simple Correlation of Alternative School Accountability Measures: TAAS Math for Grades 5 and 6

	Average score	Average gain	Relative gain
Average score	1.00		
Average gain	0.27	1.00	
Relative gain	0.67	0.86	1.00

NOTE: Correlations are weighted by the number of students in each school. These data exclude all students moving into school during the year plus those eligible for special education or bilingual programs. Each measure is calculated for individual grades and then aggregated to the school level.

small distinctions arose, the systems fall into a few groups; the differences provided the chance to examine the design features of these systems and learn whether they influence the effectiveness of accountability as a policy. We found that design does matter: The results that states obtain can be markedly different based only on the approach they use.

School Scores

We begin by looking at the individual student test score. We know that the score a student receives on an achievement test is influenced by multiple factors: earlier learning, family background, test measurement error, and the actual contribution of his schooling in the year tested. But a test score at one point in time captures the effect of all these, not simply that of the school.

Depending on the method of aggregating a school score from student-level scores, the school score also captures these other factors to varying extents. *Simple averages* of annual test scores produce results that can differ over time simply because of changes in the student population, a real problem in schools with high student mobility rates. Purer results are obtained when school scores aggregate the *gain scores* for individual students over time (that is, the improvements in their scores); the influences of family background and prior learning tend

to disappear when these scores are used. Still, the magnitude of gains may depend on the starting point—low-performing students may achieve higher gains than high-performing ones—so comparison across schools may be problematic. For this reason, a third method (not currently in use but valuable for comparison purposes) examines gains relative to other like-situated schools. We refer to this approach as the *relative gain score*.

To gauge the effects that program design has on school scores, we compute then compare the rankings of schools over the same set of student scores. The student scores from the Texas Assessment of Academic Skills (TAAS) test for 5th and 6th graders for over 1000 schools were used. If no difference in the computational methods existed, the correlations of school ranks should be unitary. The correlation results are shown in Table 1. The low correlation of the simple average and gain scores, at 0.27, is particularly troublesome since these are the two methods most widely used in the United States today. Even more troubling is the finding that the different rankings result in many schools moving from the top quartile to the bottom and vice versa, completely reversing the signal about the effectiveness of the school. Better alignment is seen between the other comparisons, which may suggest new options for calculating scores. It is difficult to judge the success of national reform programs if the outcome metrics used in those inquiries are so unrelated.

CONCLUSIONS

Improving educational quality has a dramatic effect on the economic well-being of individuals and nations. The original research described here reinforces the idea that public policies can positively affect the course of education quality. The findings demonstrate that, overall, the adoption of accountability policies produces higher academic gains than having no policy, but that the impacts are not equally distributed across all student groups. We also find that the designs of the systems themselves must receive careful attention so that consistent and accurate information about school performance can be obtained.

REFERENCES

Hanushek, Eric A. "The Failure of Input-Based Schooling Policies." *Economic Journal*, February 2003, *113*(485), pp. F64-98.

Hanushek, Eric A. "Some Simple Analytics of School Quality." NBER Working Paper 10229, National Bureau of Economic Research, January 2004.

Hanushek, Eric A. and Kimko, Dennis D. "Schooling, Labor-Force Quality, and the Growth of Nations." *American Economic Review*, December 2000, *90*(5), pp. 1184-208.

Hanushek, Eric A. and Raymond, Margaret E. "Does School Accountability Lead to Improved Student Performance?" *Journal of Policy Analysis and Management*, Spring 2005, *24*(2), pp. 297-327.

Hanushek, Eric A.; Raymond, Margaret E. and Rivkin, Steven G. "Does It Matter How We Judge School Quality?" Paper presented at the American Education Finance Association annual meetings, Salt Lake City, Utah, March 11-13, 2004.

Lazear, Edward P. "Teacher Incentives." *Swedish Economic Policy Review*, 2003, *10*(2), pp. 179-214.

Mulligan, Casey B. "Galton Versus the Human Capital Approach to Inheritance." *Journal of Political Economy*, December 1999, *107*(6, Part 2), pp. S184-224.

Murnane, Richard J.; Willett, John B.; Duhaldeborde, Yves and Tyler, John H. "How Important Are the Cognitive Skills of Teenagers in Predicting Subsequent Earnings?" *Journal of Policy Analysis and Management*, Fall 2000, *19*(4), pp. 547-68.

APPENDIX

Table A1

Differential Racial and Ethnic Impact of Accountability on State Growth in NAEP Reading and Mathematics Performance (4th to 8th Grade), 1992-2002

	Accountability by ethnicity	Disaggregation of state accountability
Consequential accountability	3.40 (2.8)**	3.54 (3.0)**
Consequential accountability x black	–2.04 (2.0)*	
Consequential accountability x Hispanic	3.10 (2.4)*	
Disaggregated x Hispanic		–2.35 (2.0)*
Disaggregated x black		3.02 (2.0)*
Report card system	0.72 (0.6)	0.72 (0.6)
(%Population age 25+) ≥ high school	0.05 (0.7)	0.06 (0.9)
School spending, $/ADM ($1000)	–1.14 (0.6)	–1.07 (0.6)
Change in exclusion rates	0.50 (3.5)**	0.51 (3.5)**
Black	–6.34 (2.5)*	–6.76 (2.6)**
Hispanic	–10.17 (4.4)**	–9.80 (4.2)**
Minority exposure x black	–8.59 (2.7)**	–8.16 (2.4)*
Minority exposure x Hispanic	–4.90 (1.4)	–4.98 (1.4)
Observations	348	348
Number of states	42	42
R^2	0.956	0.956

NOTE: */** Indicates significance at the 5/1 percent levels. All models are estimated with state fixed effects. Models include NAEP 4th grade scores for reading and math (lagged four years) and indicator variables for test and period. Absolute value of robust *t* statistics (with clustering by state) in parentheses.