School Accountability and Student Performance

Eric A. Hanushek and Margaret E. Raymond*

hanushek@stanford.edu

October 2005

Paper prepared for the

Challenges to Public Education Financing
Facing Missouri and the Nation

Conference co-hosted by:
FEDERAL RESERVE BANK OF ST. LOUIS
MURRAY WEIDENBAUM CENTER, WASHINGTON UNIVERSITY

Washington University
November 4, 2005

* Hoover Institution and National Bureau of Economic Research; Hoover Institution and CREDO, respectively. This research was supported by funding from the Packard Humanities Institute.
Impacts of State Accountability on Student Performance

By Eric A. Hanushek and Margaret E. Raymond

Since the passage of No Child Left Behind, a regular 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 operation 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 infer many of the overall effects of the federal legislation from the prior state experiences.

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

- Performance on standardized tests of the type central to state accountability systems has powerful 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

While much of the study of how schooling impacts on individual earnings has focused just on attainment, or the quantity of schooling, more recent research has turned to issues of quality. This altered research focus is consistent with the current policy attention to student testing and accountability in the United States, United Kingdom and elsewhere.

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 impact of test performance on earnings (Mulligan (1999), Murnane, Willett, Duhaldeborde, and Tyler (2000); Lazear (2003)). These studies employ different nationally representative data sets that follow students after they leave schooling and enter the labor force. When scores are standardized, they suggest that one standard deviation increase in mathematics performance at the end of high schools translates into 12 percent higher annual earnings.²

Figure 1 graphically portrays the impact of higher quality of schooling: if one compares the median earnings in 2001 of a typical individual in the U.S. with the amount they would earn if their school outcomes had been one standard deviation higher in quality (i.e., their measured achievement changed from the 50th percentile to the 85th), expected earnings shifts upward by some 12 percent each year throughout their earnings.

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

² Murnane, Willett, Duhaldeborde, and Tyler (2000) provide evidence from the High School and Beyond and the National Longitudinal Survey of the High School Class of 1972. Their estimates suggest some variation with males obtaining a 15 percent increase and females a 10 percent increase per standard deviation of test performance. Lazear (2003), relying on a somewhat younger sample from NELS88, 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 AFQT score in the NLSY data. By way of comparison, estimates of the value of an additional year of school attainment are typically 7-10 percent.
career. While the research is less extensive, similar or larger magnitudes of earnings improvement have been found in other countries.

Moreover, while not shown in this figure, there are additional gains that would accrue because individuals with greater skills tend to continue farther in schooling, i.e., to have higher school attainment. Murnane, Willett, Duhaldor, and Tyler (2000) separate the direct returns to measured skill from the indirect returns of more schooling and suggest that perhaps one-third to one-half of the full return to higher achievement comes from further schooling. (Figure 1 is just the direct effects of skills, not including the indirect effects coming through added schooling). Note also that the effect of quality improvements on school attainment incorporates concerns about drop out rates. Specifically, higher student achievement keeps students in school longer, which will lead among other things to 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 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 the productivity growth compound over time, to dramatic effect. Using U.S. economic levels in Figure 2, if the economy grew by one percent a year starting in 2000, GDP per capita would increase by 65 percent by 2050. Were the economy of the U.S. to grow at two percent a 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 one standard deviation increase in student achievement (moving from the 50th to the 84th percentile)
translates into one 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 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 is today. 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 the 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. Achieving higher 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 been able to record as significant impacts as previously described. The evidence is consistent across many
countries – US and international, developed countries and developing ones – throwing money at schools does not result in improved results (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, or the pupil-teacher ratio. We have more experiences – 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.

But the rewards are slim to none. As shown in Figure 4, U.S. performance on the National Assessment of Educational Performance (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 have the highest expenditures, and lower performing countries would spend less. However, as laid out in Figure 5 that ranks countries by TIMSS performance, no such pattern exists. Of note, the U.S. is among the highest spending countries but ranks near the middle in terms of performance.

**Focus on Accountability**³

Over the past decade, a sea change has occurred in the design of education policies in many countries around the globe. Policies have shifted from attending to inputs and processes to a focus on the outcomes realized by students. The change has

³ The explicit modeling of accountability is fully developed in Hanushek and Raymond (2005). This section relies on the results there.
rested on a widening practice of testing students with a common test against a set of standards for what they ought to know at each grade. Thus, standards, testing and accountability go hand in hand.

Where countries have a single education administration, as in Taiwan or the UK, 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 states. Over the past 10 years, states adopted their own policies at different times. This 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 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 National Assessment of Education Progress (NAEP). The program is designed to test a representative sample of students in 4th, 8th and 11th grades in Reading, Mathematics, Science on a 4-year cycle. Starting in 1992, the student samples were selected to be representative both nationally and for states individually. Participation until recently was voluntary, so states could join in only one subject test or restrict the grades that are tested. Still, it is the only available common measure of performance across states.
In recognition of the heterogeneity of student results, the US Congress in 2001 passed sweeping education reform legislation, *No Child Left Behind*. While not completely standardized, *NCLB* pushes towards a common practice on accountability throughout the United States. While the law respects the states’ rights to design both education policies and standards/testing policies, it requires each system to have 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 each state 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 evening of program characteristics by NCLB brings the natural experiment of differences to a close.) Combined with the periodic scores reported on NAEP tests, which were given every four years throughout the period, three research questions can be address:

Does accountability work?

Are there differential effects by subgroup?

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, since state participation in NAEP is voluntary, so even among those states with accountability policies, data was lacking for some grades in some years. States also differed in excluding students with disabilities, language proficiency or duration in a school from taking the test, so that there is some mixing of students across states and over time within states. Finally, accountability was not the only reform program states tried over the study period, yet 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 score results. To the extent that states differ in the way they measure school performance, they may produce different signals to schools leading ultimately to different policy results. Our assessment of existing accountability systems takes the state choices as given, 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. Some states merely make the information public (known as report card states) while others introduce consequences in the form of rewards and/or sanctions.
The current analysis looks at consequence differences to test if the design characteristics of states’ accountability systems matter. We return later to issues of overall design.

**Conditional Consequences**

Accountability programs differ in how they may use the accountability scores and this may influence their effectiveness. 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 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 were programs that included consequences, and 14 employed 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 the discussion now turns.

**Modeling the Effectiveness of Accountability**

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

For both Math and Reading, we can test the progress of two cohorts of students in states participating in NAEP. As noted in Figure 6, we can use math scores in 1996 and 2000 and reading scores in 1998 and 2002 (combined with fourth 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 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 Appendix Table A1).

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 with consequence to school performance, but report card states were found not to have significantly different achievement than 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 have an impact on 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, but in this case provides important insight, since 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 a differential effect 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. While 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 if 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.

---

4 A more complete discussion of these design issues is found in Hanushek, Raymond, and Rivkin (2004).
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 all these effects, not simply the school effect.

Depending on the method of aggregating a school score from student-level scores, the school score also captures these other dynamics 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; the influences of family background and prior learning disappear. 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) is used to examine gains relative to other like-situated schools. We refer to this approach as the relative gain score.

To gauge the design effects of school scores the rankings of schools are computed over the same set of student scores and then compared. The student scores from the Texas TASS 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 .37, 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 affect on the economic wellbeing 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.
Figure 1. Median U.S. Individual Earnings with 1.0 s.d. Reform ($=0.12)
Figure 2: Effect of Economic Growth on U.S. Income

- **GDP per capita**
  - 2% annual growth
  - 1% annual growth
  - No growth

- **Year**

- **Values**
  - $57,480
  - $94,071
Figure 3. Improved GDP with Moderately Strong Knowledge Improvement
Figure 4. National Assessment of Educational Progress (NAEP), age 17
Figure 5. TIMSS performance and spending (countries ranked by TIMSS aggregates)
Figure 6. State Accountability over Time
(with NAEP Testing Dates)

Number of states


math 8 read 8 math 8 read 8

Report Cards
Consequential Systems
Figure 7. Effect of Consequential Accountability on Achievement by Race/ethnicity

- White
- Black
- Hispanic

NAEP gains vs. Race/ethnicity:
- No consequences
- Consequences
Figure 8. Racial/Ethnic Gaps by Consequential Accountability Status
(NAEP gains relative to whites)
Table 1. Simple correlation of alternative school accountability measures, TAAS math test for grades 5 and 6

<table>
<thead>
<tr>
<th></th>
<th>Average score</th>
<th>Average gain</th>
<th>Relative gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average score</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average gain</td>
<td>0.27</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Relative gain</td>
<td>0.67</td>
<td>0.86</td>
<td>1.00</td>
</tr>
</tbody>
</table>

a. Correlations are weighted by number of students in each school. Sample for calculations excludes all people 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.
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

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

* significant at 5%; ** significant at 1%

Notes: All models estimated with state fixed effects. Models include NAEP 4th grade scores for reading and for math (lagged four years) and indicator variables for test and period. Absolute value of robust t statistics (with clustering by state) in parentheses.
References


