The Promise and Pitfalls of Using Imprecise School Accountability Measures

Thomas J. Kane and Douglas O. Staiger

Forthcoming in *The Journal of Economic Perspectives*, Fall 2002.

Thomas J. Kane is Professor of Policy Studies and Economics, University of California at Los Angeles, Los Angeles, California. Douglas O. Staiger is Associate Professor of Economics, Dartmouth College, Hanover, New Hampshire. Kane is a Faculty Research Fellow and Staiger is a Research Associate of the National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are tomkane@ucla.edu and doug.staiger@dartmouth.edu.

Abstract

In recent years, most states have constructed elaborate accountability systems using school-level test scores. However, because the median elementary school contains only 69 children per grade level, such measures are quite imprecise. We evaluate the implications for school accountability systems. For instance, rewards or sanctions for schools with scores at either extreme primarily affect small schools and provide weak incentives to large ones. Nevertheless, we conclude that accountability systems may be worthwhile. Even in states with aggressive financial incentives, the marginal reward to schools for raising student performance is a small fraction of the potential labor market value for students.

Over the last decade, states have constructed elaborate incentive systems for schools using school-level test scores. By the spring of 2002, all fifty states and the District of Columbia had implemented some form of accountability for public schools using test scores. For instance, the state of California spent nearly $700 million on financial incentives in 2001, with bonuses of up to $25,000 for teachers in schools with the largest test score improvements. The federal No Child Left Behind Act of 2001 mandates even broader accountability, requiring all states to test children in grades three through eight within three years and to intervene in schools failing to achieve specific performance targets.

Economists have paid scant attention to the properties of school accountability systems.[1] The nature of the incentives presented schools ultimately depends upon the strengths and weaknesses of the school-level mean test score measures upon which most systems are based. Accordingly, in this article, we describe the statistical properties of school test score measures, which are less reliable than is commonly recognized, and explore the implications for school incentives. Many accountability systems that appear reasonable at first glance perform in perverse ways when test score measures are imprecise.

Elements of a School Accountability System

School accountability systems typically include three elements: testing students, public reporting of school performance, and rewards or sanctions based on some measure of school performance or improvement.[2] As seen in Figure 1, by the 2001-2002 school year, nearly all states issued report cards on their schools. It was less common for states to include explicit financial rewards (18 states) or sanctions (20 states).[3] However, the 27 states that provided either rewards or sanctions in 2001-2002 included nine of the ten largest states, and contained 75 percent of the U.S. population. We briefly describe some of the common design elements currently being used by states in each of these three areas.

Student Testing

Although some states use “off-the-shelf” tests such as the Stanford Achievement Test or Iowa Test of Basic Skills in their testing programs, many states have developed their own tests. Nearly every state and the District of Columbia currently administer some type of English and math test for at least one grade level in elementary schools, in middle schools and in high schools (testing, for instance, in grades 4, 8 and 10). Only one-third of the states currently administer such tests in grades 3-8, as mandated by the No Child Left Behind Act of 2001. In addition, roughly two-thirds of the states administer tests in other subject areas – primarily science, history or social studies -- in at least one grade.

States differ in their requirements determining which students must take the test. For example, students with learning disabilities, limited English proficiency, or who are absent on the day of the test are often exempted from taking the test. Such rules give school personnel considerable opportunity to manipulate which students take the test, and thus affect average performance. Many states have attempted to limit such behavior by penalizing schools with a low proportion of students taking the test. In Massachusetts and Colorado, for example, absent students are counted as failing the test, while in Florida and Michigan any school with a large proportion of students not taking the test is ineligible for the state’s highest rankings.

Public Reporting

Most states provide the public with a report card on each school, containing information on student test score performance.[4] In 30 states, this information is used to form an overall performance index for each school, often in the form of a letter grade (A-F) or equivalent ranking. Half of these states rely solely on student test scores to construct the rankings, while the remainder use student test scores combined with other information such as attendance and dropout rates. In addition to being publicly reported, these rankings often serve as the basis for determining which schools are eligible for rewards and sanctions.

Although states differ in how they use test scores to gauge school performance, all states use some combination of three measures. The most commonly used measure is average test score *levels* among students in a given grade. Test score levels are often reported in terms of the percentage of students at a school scoring in various ranges, such as the proportion failing, proficient, or advanced. Second, many of the largest states – including California, Florida and New York – rely in part on *changes* in average test scores in a given school between one year and the next. California, for example, ranks schools according to whether they have exceeded an annual growth target, which is somewhat higher for schools with poor baseline performance. The third approach, used by a handful of states including Arizona, North Carolina, and Tennessee, rates schools using the average *gain* in test performance between the end of one grade and the end of the next grade. Thus, where the change approach would measure the performance of this year's 4th grade students relative to last year's 4th grade students, the gain approach would measure the performance of this year's 4th grade students relative to their own performance in 3rd grade. The latter methodology requires states to invest in data systems enabling them to link individual students' scores across years. Test score changes and gains, which focus on improvement in performance, are generally seen as less biased methods of comparing schools serving different student populations.

Finally, to encourage schools to raise the performance of all youth, 17 states report separate measures of the performance for specified subgroups, including minority, low-income, and limited-English-proficient students. The accountability system in eleven states, including Texas and California, explicitly penalize schools with poor subgroup performance.

Sanctions and Rewards

The majority of state accountability systems now use some combination of explicit sanctions for low performing schools and monetary rewards for high performing schools. Sanctions are used sparingly, with fewer than 5 percent of schools typically eligible for some form of sanction in a given year. Typically, the lowest performing schools in any single year are required to submit an improvement plan, and are eligible for various forms of assistance from the state. Schools that continue to be low performing over multiple years are subject to increasingly stringent sanctions, such as permitting their students to transfer to other schools (eleven states) and eventual closure or reconstitution (17 states). However, such sanctions have been rarely used, even in states with long-standing accountability systems such as Texas. The federal No Child Left Behind Act will require much tougher sanctions in the coming years including providing vouchers to parents for out-of-school programs, and eventually replacing school staff or converting failing schools to charter schools.

Financial awards are typically made to a school for schoolwide use, although a handful of states including California, North Carolina and Pennsylvania allow for bonuses to teachers. The magnitude of the financial awards vary. For example, award-winning schools in Texas received less than $5,000 per school, while the average award in California was over $50,000 per school in 2001. Financial incentives are often spread across a large number of schools with performance thresholds near the state average, but sometimes large rewards are targeted on a few schools with exceptional performance. For instance, in California, roughly half of all schools received an award in 2001, but the teachers in less than 1 percent of schools with the largest test score increases received awards of $5,000 to $25,000 per teacher.

Signal and Noise in School-Level Test Scores

In the rush to implement accountability systems, little attention has been paid to the imprecision of the test score measures on which they are based. Yet two well-known facts suggest that school performance is difficult to measure with one year's worth of test data. First, we have known at least since the analysis by Coleman et al. (1966), that the between-school variance in student test scores represents only 10 to 15 percent of the variance in student test scores. In other words, the difference in mean test performance between the best- and worst-performing school is not nearly as large as the difference in performance between the best and worst student in the typical school. Second, the median elementary school in the United States has only 69 students per grade level.[5] With a sample that small, even a few particularly talented or rambunctious youth can have a large impact on the grade-level score for a school from year to year. When volatility in test scores is combined with a relatively narrow distribution of school performance, it implies that the 95 percent confidence interval for the average 4th grade reading or math score in a school with 69 students per grade level would extend from roughly the 25th to the 75th percentile among schools of that size (Kane and Staiger, 2002b).

The importance of sampling variation in school-level mean test scores is immediately apparent in Figure 2, which plots various measures of 4th grade math performance against the number of 4th grade students taking the test for 1163 elementary schools in North Carolina in 1998. These plots illustrates two facts which are important in the discussion of volatility. First, there is high variability in test scores among small schools. The upper left-hand panel shows the mean level of math performance in fourth grade; the lower left-hand panel reports the mean gain in math performance of individual students tracked from the end of third grade to the end of fourth grade. For both test score levels and gains, virtually all of the schools with the highest and lowest scores were small schools. Second, this variability in test scores among small schools is not solely due to heterogeneity among small schools. The right-hand panels of Figure 2 report the one-year change in each school's mean test score and mean gain score between 1998 and 1999. The small schools are also much more likely to report large *changes* in mean scores and mean gains from one year to the next.

There are many potential sources of short-term fluctuations in student performance. Sampling variation is surely one factor. But there are also likely to be other sources of variation at the classroom level, generated by classroom chemistry between a teacher and class, or the presence of particularly disruptive students in a class. Indeed, certain sources of variation might generate temporary fluctuations in test performance for a whole school, such as a dog barking in the parking lot on the day of the test or inclement weather.

One way of testing for the importance of sampling variation and other one-time shocks to performance is to determine whether improvements (or declines) in a school's test score performance tend to be reversed in the following year. If a change in test scores is due to a one-time phenomenon, then the school is likely to revert back to its prior performance level in the next year. In the extreme, if school test scores were pure noise and independent from one year to the next, we would expect a correlation between the change this year and the change next year to be -0.5.[6] On the other hand, when test score changes reflect permanent improvements, they serve as the basis for subsequent improvements or declines, and we would not expect such mean reversion. Indeed, we would expect a correlation of 0 in the change from one year to the next. If some changes are permanent and some changes are purely transitory, one would expect a negative correlation between 0 and -.5.

Using data from North Carolina, we estimate that changes in 4th grade math scores have a correlation of -0.37 with changes in the next year. This estimate suggests that 74 percent of the variance in the change in math scores is transitory (-0.37 is 74 percent of -0.5).[7] Similarly, changes in 4th grade math gains have a correlation of -0.45 with changes in the next year, suggesting that 90 percent of the variance in the change in this measure is transitory. Estimates suggest even less persistence in reading scores (Kane and Staiger, 2001). In other words, if one were to look for signs of improvement by closely tracking changes in school-level scores from one year to the next, most of what one observed would be temporary -- either due to sampling variation or some other non-persistent cause.

But the variation in school-level test score measures is not entirely transitory. Figure 3 plots the correlation of school test scores one to five years apart, using 4th grade math scores (levels and gains) from North Carolina between 1994 and 1999. The correlation in school performance measures one year apart is 0.77 for math levels and 0.41 for math gains. The correlation falls off gradually as one looks at school test scores more than one year apart, declining by about 10 percent per additional year for test score levels and by about 20 percent per additional year for gains. In other words, schools with high test scores in a given year tend to have high test scores in future years, but the correlation drops dramatically with a one year lag and then fades out gradually over time.

There is a very simple interpretation of Figure 3. Suppose that observed test scores are composed of two parts. One part represents persistent factors influencing test scores, such as the quality of teachers at the school or the curriculum being used, which are highly correlated from year to year. As time passes, these persistent factors influencing test scores may gradually evolve, due to such factors as teacher turnover or the adoption of new curriculum, causing a steady decline in the correlation as one looks at school test scores further apart in time. But a second part of observed test scores represents purely non-persistent factors influencing test scores, such as sampling variation or a particularly disruptive student in a class, which are equally uncorrelated at a one-year or five-year lag. These non-persistent factors, even if they have real effects on student learning (as is the case with a disruptive student), are the types of things that schools have little control over and cannot replicate on a consistent basis. Thus, the low correlation between scores one year apart reflects the presence of non-persistent factors, while the more gradual decline in correlation in school test scores more than one year apart reflects the slow change in persistent factors.

This interpretation of the data suggests a simple way of decomposing the variance in test scores into persistent and non-persistent components. The gradual decline seen in Figure 3 beyond the first year implies that the correlation in the persistent component declines by about 10 percent per year for test score levels, and by about 20 percent per year for test score gains. In fact, this annual decline in the correlation of the persistent component is broadly consistent with the annual teacher turnover hazard of between 10 and 15 percent in North Carolina (Kane and Staiger, 2001). Assuming that the annual decline in the correlation is constant, this would imply that school test scores one year apart should be correlated 0.9 in levels, and 0.8 in gains. Any *additional* decline in the correlation at the first year is the result of the non-persistent component of test scores. This logic suggests that about 13 percent of the variation in test score levels is non-persistent (the difference between the 0.9 correlation expected from the persistent component, and the 0.77 correlation in observed test scores), while about 39 percent of the variation in test score gains is non-persistent (the difference between the 0.8 correlation expected and the 0.41 correlation observed).

Figure 4 uses this approach to decompose the variation in school test scores into persistent and non-persistent sources. In addition, we estimate the proportion of the variance in school test scores which is the result of sampling variation alone. It is straightforward to calculate the contribution of sampling variation: each school's test score is a sample average over n students, so has a sampling variance of σ^2/n where σ^2 is the within school variance in the test score.[8] We do the calculation for three commonly used measures of performance: test score levels, test score gains, and changes in test score levels. To illustrate the increased importance of noise in measures for small schools, we do the calculation separately for the smallest, middle and largest quintile of schools (based on average number of students taking the test between

1994 and 1999).

Several results are worth highlighting in Figure 4. First, much of the difference in the test score *levels* is persistent. Even among the smallest quintile of schools, non-persistent factors account for only 27 percent of the variance between schools. Among the largest quintile of schools, such factors account for only 13 percent of the variance. However, since we are not adjusting for initial performance levels or for the demographic characteristics of the students, much of that reliability may be due to the unchanging characteristics of the populations feeding those schools and not necessarily due to persistent educational practices in the schools.

Second, gain scores and changes in average level of performance are measured unreliably. More than half (56 percent) of the variance among the smallest quintile of schools in mean gain scores is due to sampling variation and other non-persistent factors. Even among the largest quintile of schools, non-persistent factors are estimated to account for 34 percent of the variance in gain scores. Changes in mean test scores from one year to the next are measured even more unreliably. More than 80 percent of the variance in the annual change in mean test scores among the smallest quintile of schools is due to one-time, non-persistent factors.

Third, the reliability of using changes in average level of performance cannot be much improved by combining information from more than one grade. For instance, even though the largest quintile of schools were roughly four times as large as the smallest quintile, the proportion of the variance in annual changes due to non-persistent factors was still over 60 percent.

Overall, these results suggest that annual test scores are quite unreliable measures of performance differences across schools and over time, particularly in smaller schools. An alternative approach that reduces the impact of year-to-year variation is to pool test score data for individual schools over time. The No Child Left Behind Act of 2001 provides states with the option of calculating three-year weighted averages of school performance on test scores. In Kane and Staiger (2001), we show that the optimal weights for averaging performance over time (which maximize the reliability of the resulting performance measure) incorporate information on the signal variance, sampling variation and the degree of persistence in the signal over time: One places more weight on past performance when a single year's test score is unreliable and when the persistent component of performance is strongly correlated over time. The resulting estimate is a sophisticated moving average of each school's prior test scores, and is equivalent to a Bayesian estimate of the posterior mean of the persistent component of each school's test score. Our work suggests that this approach can greatly improve the accuracy of school ratings. For example, we found that the optimally weighted average of past test scores was much more successful in picking schools that were likely to perform well one or two years in the future, more than doubling the forecast R-squared compared to using a single year of data. But, in most cases, simple averages of past test scores achieved more than half of this gain in forecast performance. Thus, even simple averaging over several years leads to markedly improved performance measures.

Three Cautionary Tales

How much should we be concerned about the imprecision of the test score measures being used in school accountability systems? After all, no performance measure is perfect. Even noisy performance measures may provide useful information that can be incorporated into a carefully designed incentive contract. The problem resides not with the measures themselves, but with the way that these measures are often used. In this section, we provide three cautionary tales of common practices that appear to be reasonable, yet perform in perverse ways because test score measures are unreliable.

Focusing on Schools with the Very Best and Worst Test Scores

Many state accountability systems focus on outliers, targeting the highest and lowest performing schools for rewards or sanctions. Yet, in light of the properties described above, such apparently reasonable schemes will primarily affect small schools, and will leave large schools largely unaffected.

For example, North Carolina has programs that target schools at both ends of the distribution. At one extreme, they have provided special recognition each year since 1997 to the 25 schools in the state with the largest mean "growth composite index," essentially a measure of the mean gain in performance for a school's students since the previous spring. Teachers in these schools are given a banner to celebrate their status and a banquet is held in their honor. At the other extreme, the state assigned assistance teams to intervene in schools with the poorest performance on state tests. Table 1 displays the proportion of elementary schools that were either in the "top 25" or assigned an assistance team between 1997 and 2000, in total and separately by school size deciles. Small schools were much more likely to receive awards and more likely to receive sanctions. More than a quarter (27.7 percent) of the smallest decile of elementary schools were among the top 25 schools at some point over the 4 years the awards have been given. In fact, the smallest schools were 23 times more likely to win a "Top

25" award than the largest schools!   Similarly, all but one of the elementary schools assigned an assistance team were in the bottom four deciles by school size. The smallest decile of schools would have received an even larger share of the assistance teams, except for a rule requiring that the proportion of students scoring below grade level to be statistically significantly below 50 percent.

Small schools are over-represented in these two programs largely because test score gains are much noisier in small schools. Table 1 also presents information on the mean gain scores in math for 4th and 5th grade, and the between-school variance in school mean gain scores by size decile. There is no significant difference in average performance across the deciles, but the variance in mean gain scores among schools in the smallest size decile was more than four times the variance among the largest decile of schools. Thus, by simply splitting themselves into smaller units, large schools could increase their chances of being among the top 25 – but would also increase the risk of being assigned an assistance team. Without additional noise in their test score measures, large schools have very little chance of ever appearing in the extremes of the distribution.

This pattern of test scores by school size poses some challenges for designing a system of rewards and sanctions. If the state requires schools to exceed a high threshold in order to receive a reward, there will be a much higher chance that the small school will win the award than a large school – even if the small and a large school have the same “true” performance level. Indeed, large schools are likely to be unmotivated by school accountability programs that emphasize the extremes, since they have very little chance of ever appearing in the tails of the distribution. While they may not be fully cognizant of the statistical issues involved, those working at schools of varying sizes would be expected learn over time that their efforts have a greater or lesser impact on their likelihood of winning.

One possible remedy for this problem would be to establish different thresholds for schools of different sizes. For example, grouping schools according to size (as is often done in high school sports) and giving awards to the top 5 percent in each size class would tend to even out the incentives (and disparities) between large and small schools.[9] Another alternative would be to establish thresholds closer to the middle of the test score distribution, where both large and small schools have a chance to win so that the difference in incentives is less extreme. This is what many states in fact do – give smaller rewards to a larger group of schools. In North Carolina, for example, the focus was not solely on the outlier schools. Roughly two-thirds of the schools in 1999 were identified as having achieved "exemplary" growth and these schools received the lion's share of the award money.

Drawing Inferences Based on One-Year Changes

Volatility in test scores can wreak havoc if policymakers draw inferences from short-term fluctuations. For example, when test scores were released in Massachusetts in November of 1999, Provincetown showed the greatest improvement over the previous year. The *Boston Globe* published an extensive story describing the various ways in which Provincetown had changed educational strategies, and only in the detailed table of school results did one learn that only 22 students took the test in 1999 (Tarcy, 1999). Such dramatic swings in test scores are not uncommon for a small school, and are very unreliable indicators of improving performance. Yet, there is a large literature in psychology that suggests that people are often over-confident in predicting future performance on the basis of current performance (Kahneman and Tversky, 1971; Plous,1993). That same over-confidence may lead local school administrators to draw unwarranted conclusions on the effectiveness or ineffectiveness of policies based upon such short-term fluctuations in performance.

Similarly, some accountability systems require schools to improve their student's performance continuously. For example, the initial versions of the No Child Left Behind Act passed by the House of Representatives and the Senate in 2001 included provisions that required schools to show improvement in math and reading scores each year or be subject to a variety of sanctions. But even when a school is on the right track, the path to improved student performance is rarely a straight path. The natural fluctuation in noisy test score measures means that two steps forward are often followed by one step back. The No Child Left Behind Act had to be rewritten in conference committee after an initial analysis– carried out after passage by both houses -- suggested that virtually every school in the country would have failed to improve at least once in a five-year period because of the natural volatility in test score measures (Kane, Staiger and Geppert, 2002; Broder, 2001).

For example, in North Carolina between 1994 and 1999, the proportion of students in grades three through five scoring at the "proficient" level or higher in mathematics rose from 55 percent to 70 percent -- roughly 3 percentage points per year. The proportion of third- through fifth-grade students scoring at the proficient level in reading grew from 61 to 70 percent, or nearly 2 percentage points per year. Progress of that magnitude has made North Carolina the envy of many other states. Nevertheless, only 11 percent of North Carolina schools witnessed an increase in math proficiency for 5 straight years, only 6 percent witnessed an increase in reading proficiency for 5 straight years, and less than 2 percent of schools witnessed an increase in both subjects for 5 straight years. Thus, under the initial standard proposed in the No Child Left Behind Act, over 98 percent of North Carolina elementary schools would have "failed" and been subject to sanctions!

When looking for signs of improvement at the school level, the clear lesson is that one must look to trends over several years rather than the change in any single year.

Racial and Ethnic Subgroup Rules

Despite some closing of the gap in academic performance between whites and blacks between the mid-1970s and the late 1980s, the gaps in standardized test scores by race and ethnicity remain quite large (Jencks and Phillips, 1998). Among non-Hispanic youth, the gap in mean test scores between whites and blacks is often between three-quarters and one standard deviation. The gap between non-Hispanic white and Latino youth is slightly smaller. Because of concerns about these gaps in test performance, many accountability systems (including the No Child Left Behind Act of 2001) require schools to achieve minimum standards not just for the school as a whole, but for each racial and ethnic subgroup within the school as well. The goal is worthwhile, but the design of accountability rules may well be counterproductive – targeting fewer resources to diverse schools simply because of their diversity and providing strong incentives to schools to segregate by race or ethnicity.

The California system provides an interesting example. In order to win a "Governor's Performance Award" in California, a school must achieve a minimum improvement in performance at the school level, but also for each "numerically significant" racial or ethnic subgroup within the school.   To be numerically significant, a group must represent at least 15 percent of the student body and contain more than 30 students, or represent more than 100 students regardless of their percentage. There are eight different groups which could qualify as "numerically significant," depending upon the number of students in each group in a school: African American, American Indian (or Alaska Native), Asian, Filipino, Hispanic, Pacific Islander, White non-Hispanic or "socioeconomically disadvantaged" students (defined as a student of any race neither of whose parents completed a high school degree or who participates in the school's free or reduced price lunch program).

Of course, measured changes in test scores for small subgroups are often quite volatile. Moreover, to the extent the higher variance is due to sampling variation, their fluctuation is likely to be nearly independent. As a result, California's rules are analogous to a system that makes every school flip a coin once for each minority subgroup, and then gives awards only to schools that get "heads" on *every* flip. Schools with more minority subgroups must flip the coin more times and, therefore, are put at a purely statistical disadvantage relative to racially homogeneous schools. Moreover, because improvement at most schools exceeded their target, the additional volatility of scores for small sized subgroups – and small schools in general – further increased the chances of falling below the target (unlike the case of the "Top 25" award program in North Carolina where volatility helped small schools to achieve a high target).

This statistical disadvantage is clearly seen in Table 2, which reports the proportion of California elementary school's winning their Governor's Performance Award by school size quintile and number of numerically significant subgroups in each school. Among the smallest quintile of elementary schools, racially heterogeneous schools were almost half as likely to win a Governor's Performance award as racially homogeneous schools. This is particularly ironic given that the more integrated schools had slightly larger overall growth in performance between 1999 and 2000 (36.0 points versus 33.4 points). The statistical bias against racially heterogeneous schools is also apparent among larger schools, but somewhat less pronounced because subgroups in these schools are larger in size and, as a result, their scores are less volatile.[10]

Table 2 has several important implications. First, under rules that focus on performance by subgroups, a district would have a strong incentive to segregate by race/ethnicity. Consider a district with four small schools, each being evenly divided between African American, Latino, Asian American and white, non-Hispanic. According to the results in Table 2, the district could nearly double each school's chance of winning an award by segregating each group and creating four racially homogeneous schools. Second, because minority youth are more likely to attend heterogeneous schools than white non-Hispanic youth, the average school attended by minority youth actually receive less award money as a result of the minority subgroup rule. For example, the average African American student attended a school with 2.8 subgroups, compared to 2.2 subgroups in the schools attended by the average white non-Hispanic student. As a result, African Americans in California were nearly 5 percent less likely to be in an award winning school solely because of the statistical bias against schools with subgroups (Kane and Staiger, 2002c). A similar calculation suggests that Latinos were 2.5 percent less likely to be in an award winning school because of the subgroup bias.

Other Limitations of School Test Score Measures

School test score measures are imperfect measures of schools' output for at least three other reasons. First, test score measures may reflect factors outside of a school's control, such as family background, which grant schools in more wealthy districts an advantage, particularly when schools are rated on the basis of their level of performance. One partial solution to this problem is to focus on so-called "value-added" measures of achievement, such as the gain scores we have analyzed above. However, such a solution is only partial, since students differ not only in their baseline performance, but also in their subsequent trajectory.[11]

Test score improvements and value-added measures may be less biased methods of rating school performance than test score levels. But the reduction in bias comes at the cost of greater unreliability. Relative to test score levels, gains and changes in test scores are attempting to identify much smaller differences across schools with measures that are at least as noisy. Test score measures are being used for two purposes-- to provide incentives and to identify best practice. Although the incentive problems created by greater imprecision can be accommodated by attaching less weight to the noisy measures, the search for best practice requires less noisy measures. As a result, when using value-added measures or changes in scores over time, it is particularly important to implement some method for pooling data over time.

A second problem with test score measures is that they are incomplete measures of school output. Most test-based accountability systems are based upon reading and math scores alone.[12] Other academic subjects (such as science) and dimensions of learning that contribute to the public good (such as civics) are typically assigned zero value. Similarly, standardized tests are often designed to determine basic proficiency, and may assign little value to advanced knowledge of a subject. Rewarding schools for tested subjects will implicitly punish schools that excel in other valued, but difficult-to-measure domains, and distort instruction toward the subset of subject areas and concepts that are tested. For example, in the Kentucky accountability system in the early Nineties, science was tested in fourth grade and math was tested in fifth grade. Stecher and Barron (1999) found that teachers had reallocated their time so that they spent much more time on science in fourth grade when students took the science test and on math in fifth grade when students took the math test. Jacob (2001) found that scores on science and social studies leveled off or declined in Chicago after the introduction of an accountability system focused on math and reading performance.

A third source of error in test score measures is occasionally introduced by the test publishers themselves. The most egregious errors (such as mistaken scoring sheets or test booklets with missing pages) are usually found. But more subtle errors are introduced when companies attempt to "equate" tests from year to year, when new test items are introduced. Such errors can be difficult to find, as New York City learned in 1999, when 9000 students were mistakenly sent to summer school and the superintendent resigned on the basis of test scores that were subsequently revised upward by the test publisher due to an equating error[13] (Steinberg and Henriques, 2001a, 2001b).

The Impacts of Accountability Systems on Student Achievement

Although there is some evidence that school-level and teacher-level incentives were associated with improved student test performance in Israel (Lavy 2002a, 2002b), the evidence on the impact of accountability systems on student achievement in the U.S. has been limited and ambiguous. The experience in North Carolina and Texas-- two states with high visibility accountability systems-- has been cited as evidence of the value of accountability systems. Figure 5 portrays the improvement in state math scores for public school students on the National Assessment of Educational Progress (NAEP) in 4th grade and 8th grade between 1992 and 2000.   At the national level, 4th grade math scores rose by slightly under a quarter of a standard deviation over the decade, and 8th grade math scores grew by one-fifth of a student-level standard deviation.   However, mean math scores in 4th grade and 8th grade grew by .6 standard deviations in North Carolina and by roughly .5 and .3 standard deviations in 4th and 8th grade, respectively, in Texas. To provide some appreciation for the magnitude of this increase, remember that the gap in test scores between blacks and non-Hispanic whites is typically between three-quarters and one standard deviation.

Some analysts have drawn a connection between the test score increases in the states of North Carolina and Texas and those states' emphasis on school accountability (Grissmer and Flanagan, 1998). However, there are three reasons to be cautious in drawing such an inference. First, North Carolina and Texas were not the only states pursuing aggressive school accountability policies during the 1990s. Indeed, by the late 1990s, only a handful of states were <u>not</u> doing so. For instance, Kentucky was an early innovator in the move toward school-level accountability and their improvements on the NAEP during the 1990s was close to the national average. Ohio achieved impressive gains in both 4th and 8th grade math, yet it has no policy of attaching financial incentives to school test scores.   The successes in North Carolina and Texas may well be because of their accountability policies, but the success may have been due to other policy differences as well. For instance, North Carolina also invested heavily in encouraging teachers to prepare for certification by the National Board on Professional Teaching Standards, while Texas reduced class size over this period.

Second, at least part of the increase in North Carolina and Texas may have been due to an increase in the proportion of sampled youth excluded from the NAEP sample. The NAEP test has traditionally excluded the test scores of students who are granted testing accommodations, such as extra time to take the test. Most states granted accommodations to a larger share of students following the passage of the Individuals with Disabilities Act in 1996.[14] Between 1992 and 2000, the average state increased its exclusions by 3.5 percentage points (from 5 percent of sampled youth to 8.5 percent of sampled youth) while Texas and North Carolina, increased their exclusion rates by 5 and 10 percentage points respectively. In fact, the increase in exclusion rates in North Carolina was larger than in any other state.

However, it is unlikely that the change in exclusion rates accounts for all of the rise in test scores. To provide an upper bound of the effect of the exclusion rates, suppose that students are randomly excluded from the NAEP in 1992, but North Carolina alone systematically excluded 13 percent of students from the bottom tail of the distribution in 2000. This is an extreme scenario since not all of the students excluded in 2000 would have scored in the bottom tail. If the distribution of test scores is normal at the student level, then truncating scores at the 13th percentile would have raised test scores by .25 standard deviations. In contrast, 8th grade math scores in North Carolina grew .42 student-level standard deviations faster than the rest of the country between 1992 and 2000. Thus, even under this extreme scenario, the exclusions cannot explain all of the excess rise in test scores. (Grissmer and Flanagan (forthcoming) reach a similar conclusion.)

Third, there is no evidence that the rate of improvement in North Carolina accelerated in 1997, when the state first began offering financial incentives to teachers and schools.[15] Between 1993 and 1999, 4th grade and 8th grade math scores grew by .58 and .50 standard deviations respectively (.10 and .09 standard deviations per year), with similar improvements at the 90th and 10th percentile. However, the rate of increase between 1996 and 1999 -- after the financial incentives were added -- was no higher than the rate of increase between 1993 and 1996.

More generally, performance on state high stakes tests must be interpreted cautiously. Unlike North Carolina, the improvements in most states on the test used for accountability purposes are far larger than improvements on the NAEP scores (Linn and Dunbar, 1990). For example, Koretz and Barron (1998) found that improvements on the fourth grade math test Kentucky used for its accountability system were four times as large as that state's improvement in the NAEP between 1992 and 1996. Klein et. al. (2000) also report that the increase in scores on the Texas Assessment of Academic Skills was much larger than the improvement on the NAEP.

There are several reasons we might see larger improvements on a state's high stakes tests than on other tests. For example, the tests may not measure the same skills. The Kentucky test focused more on general problem-solving skills, while the NAEP tends to focus more on specific mathematical knowledge. The North Carolina test was probably more similar to the NAEP than the Kentucky test – so it may not be surprising that the increases on NAEP and the state test were similar.

But the differentials in improvement may also reflect more perverse incentives that are unavoidable with a high-stakes test. First, for any concept to be tested, the questions must eventually take written form, and coaching on that form can often have an impact on student performance. Koretz (forthcoming), who provides a particularly illuminating discussion of these issues, cites an arresting example reported by Shepard (1988), drawing on data from the New Jersey Department of Education (1976). Students in New Jersey were asked to add and subtract decimals. When students were asked to add decimals in the familiar vertical format, the passing rate was 86 percent; when the decimals were provided in horizontal format, the passing rate was 46 percent; when students were asked to subtract rather than add, the passing rates in the vertical and horizontal formats were 78 and 30 percent respectively. When Koretz et. al. (1996) asked teachers in Kentucky to report the importance of several different factors to account for the improvements in student test scores, more than half of the teachers said "increased familiarity" with Kentucky's accountability test and "work with practice tests and preparation materials" had been important, while only 16 percent reported that "broad improvements in knowledge and skills" accounted for the improvement.

Second, when the stakes are sufficiently high, accountability systems can generate cheating by teachers and school administrators. There have been a number of anecdotal examples of educators falsely raising student achievement: investigators in New York City charged that dozens of educators had cheated, telling students to change incorrect answers or giving them practice tests containing questions from the actual test (Goodnough, 1999). Michigan launched a widespread investigation of teacher cheating on the 2001 state assessment (Wilgoren, 2001). However, evidence on the actual extent of cheating has been understandably scarce. Jacob and Levitt (2002) take a novel approach, identifying cheating by focusing on schools with unusual patterns of common student responses combined with unexpected rises and subsequent declines in student performance.   Their estimates suggest that 4-5 percent of classroom test scores in Chicago elementary schools were affected by such cheating.

Thus, even if school accountability programs bring increases in test scores, such increases should be interpreted conservatively until they can be connected to other evidence of gains in learning.[16]

The Promise of School Accountability

The success of a school accountability system will depend upon the quality of the performance measures available and the way in which rewards and sanctions are structured. If performance measures are noisy or if they have the potential to distort behavior in undesirable ways, then the literature on optimal incentives suggests that the marginal payoff attached to such measures be downweighted (Baker, 1992, 2000). That does not mean that test-score measures are of no value. It simply means that accountability systems must be carefully designed to reflect the noise and distortion inherent in test-score measures, just as the design of incentive contracts in other organizations reflects the quality of the performance measures available (Gibbons, 1998).

How *should* a test-based school accountability system be designed? Kane and Staiger (2002a) argue that, under plausible assumptions, the optimal incentive contract should have three features that are currently lacking from most accountability systems. First, it would sort schools into separate size classes to account for the fact that smaller schools have more variable performance measures. Second, it would use average school performance over many years to increase the reliability of the performance estimate and reward those schools that have persistently high test scores. Finally, to preserve schools' incentives in the short-term, the optimal contract would reward schools for improving upon their expected performance, with smaller schools receiving smaller incentive payments in line with their less reliable test-score measures.

But critics of school accountability worry that current systems already place too great a weight on imperfect measures of academic achievement and, on net, may do more harm than good. To evaluate these concerns, one must have a sense of the potential value that we should place on an increase in student achievement. Some calculations reveal that the monetary value of even a small improvement in academic achievement can have very large payoffs.

Two recent papers provide estimates of the impact of test performance on the hourly wages of young workers. Murnane, Willett and Levy (1995) estimate that a one-standard deviation difference in math performance is associated with an 8 percent hourly wage increase for men and an 12.6 percent increase in for women.[17] These estimates probably understate the value of test performance, since the authors also control for years of schooling completed.[18] Neal and Johnson (1995), who do not condition on educational attainment, estimate that an improvement of one standard deviation in test performance is associated with a 18.7 and 25.6 percent increase in hourly wages for men and women, respectively.[19] Using a discount rate of 3 percent, the present value at age 18 of an increase of one standard deviation in test performance is worth roughly $110,000 per student using the Murnane, Willett and Levy estimates and $256,000 per student using the higher estimates from Neal and Johnson.[20] Discounting these values back to age 9 – that is, 4th grade -- would reduce the estimates to $90,000 and $215,000 per student. One might argue that a 3 percent discount rate is a bit low, since investment in human capital is not risk free. But even with a discount rate of 6 percent, these estimates are only reduced by about one third.

Such estimates are quite large relative to the rewards offered to schools for increasing student test performance. For example, California paid elementary schools and their teachers an average award of $122 per student if their school improved student performance by an average of at least 0.03 student-level standard deviations.[21] Based on the calculations in the preceding paragraph, the present value of such an increase in test scores to students in elementary school is in the range of $2700 to $6400 per student (0.03 times $90,000 or $215,000). In other words, the labor market value of the test score increase would have been worth roughly 20 to 50 times the value of the incentive provided in 2001 by California – the state with the most aggressive incentive strategy.

This calculation suggests that even the most aggressive state is paying schools much less than the marginal payoff – at least if we thought the test score improvements reflected true achievement that would be rewarded in the labor market. In other words, critics' concerns about inaccuracies in performance measures may already be reflected in the relatively weak financial incentives offered in most states. In fact, the strength of incentives for schools in California is remarkably similar to what Hall and Liebman (1998) found for CEOs: $1 in compensation for every $40 increase in firm valuation.

Conclusion

Can school accountability systems, redesigned to reflect the noise and distortion inherent in test-score measures, generate enough real improvement in academic achievement to justify their expense? One would have to say that the jury is still out. But the lack of statistical power for discerning the very small increases in achievement needed to justify these systems means that the jury may never come back in! For example, California's current annual spending of $114 per student on financial incentives is worth about $1800 in present value by the time a student is age 18. A true increase in test performance of only .01 to .02 standard deviations is enough to justify this expense, but it may be very difficult to discern empirically. Therefore, one's interpretation of the empirical evidence will depend even more than usual on the null hypothesis. If one starts from the presumption that incentives matter, and that an increased emphasis on student achievement is likely to encourage schools to improve on the margin, then the burden of proof on those arguing that the we are placing inordinate weight on school test scores is quite high.

While financial incentives may encourage individual schools to improve, they may not be sufficient by themselves to generate the rapid growth many reformers seek. Firm-level evidence from other industries suggests that an important channel through which market reforms affect productivity growth is by shifting production to more productive firms and closing down less productive firms, rather than by gradual productivity improvements in every firm (Pavcnik, 2002; Tybout, forthcoming). Thus far, school accountability systems have led to little reallocation of students across schools. In practice, the impact of strategies to reallocate students (such as school choice or school closure) is limited by the lack of reliable performance measures, and may also create unintended consequences. Nonetheless, we expect that effective school accountability systems, pooling information on school performance over time to more accurately identify the most and least effective schools, will include both financial incentives and some mechanism for re-allocating students between schools.

Acknowledgements

This paper was written with support from Dartmouth College and the William and Flora Hewlett Foundation. We thank Amitabh Chandra, Nina Pavcnik, and the editors of this journal for helpful comments.

References

Baker, George. "Incentive Contracts and Performance Measurement," *Journal of Political Economy*, June 1992.

Baker, George. "Distortion and Risk in Optimal Incentive Contracts," working paper, June 2, 2000.

Black, Sandra "Do Better Schools Matter? Parental Valuation of Elementary Education" Quarterly Journal of Economics (1999) Vo. 114, No. 2, pp. 577-599.

Broder, David S. "Long Road to Reform: Negotiators Forge Education Legislation" Washington Post, December 17, 2001, p. A01.

Clotfelter, Charles and Helen F. Ladd "Recognizing and Rewarding Success in Public Schools" in Helen Ladd (ed.) Holding Schools Accountable (Washington, DC: Brookings Institution, 1996).

Coleman, James S., E.Q. Campbell, C.J. Hopson, J. McPartland, A.M. Mood, F.D. Weinfeld, and R.L. York Equality of Educational Opportunity (Washington, DC: U.S. Department of Health, Education and Welfare, 1966).

Gibbons, Robert. "Incentives and Careers in Organizations," *Journal of Economic Perspectives*, Fall 1998.

Goodnough, A. "Answers Allegedly Supplied in Effort to Raise Test Scores" New York Times December 8, 1999.

Grissmer, David and Ann Flanagan, "Exploring Rapid Achievement Gains in North Carolina and Texas" Paper written for the National Education Goals Panel, (November, 1998).

Grissmer, David and Ann Flanagan, "Tracking the Improvement in State Achievement Using NAEP Data" in Michael Ross and George Bohrnstedt (eds), Instructional and Performance Consequences of High Poverty Schooling (forthcoming).

Grogger, Jeff and Eric Eide, "Changes in College Skills and the Rise in the College Wage Premium" Journal of Human Resources (1995) Vol. 30, No. 2, pp. 280-310.

Hall, Brian J. and Jeffrey B.Leibman, "Are CEO's Really Paid Like Bureaucrats?" Quarterly Journal of Economics (1998) Vol. 113, No. 3, pp. 653-691.

Hanushek, Eric and Margaret Raymond "Lessons and Limits of School Accountability

Systems" Paper presented at "Taking Account of Accountability" conference, Kennedy School of Government, Harvard University, June 9-10, 2002.

Jacob, Brian A. "The Impact of High-Stakes Testing on Student Achievement: Evidence from Chicago" Unpublished paper, Kennedy School of Government, Harvard University, June 2001.

Jacob, Brian A. and Steven D. Levitt, "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating" Unpublished paper, Kennedy School of Government, Harvard University, April 2002.

Jencks, Christopher and Meredith Phillips (eds.), The Black-White Test Score Gap (Washington, DC: Brookings Institution, 1998).

Jencks, Christopher and Meredith Phillips, "Aptitude or Achievement: Why Do Test Scores Predict Educational Attainment and Earnings?" in Susan Mayer and Paul Peterson (eds.) Earning and Learning: How Schools Matter (Washington DC: Brookings Institution and Russell Sage Foundation Press, 1999), pp. 15-48.

Kahneman, Daniel and Amos Tversky (1971). "Belief in the law of small numbers." Psychological Bulletin 76(2): 105-110.

Kane, Thomas J. and Douglas O. Staiger, "Improving School Accountability Measures" *National Bureau of Economic Research Working Paper* No. 8156, March 2001.

Kane, Thomas J. and Douglas O. Staiger, "Improving School Accountability Systems," manuscript, May 2002a.

Kane, Thomas J. and Douglas O. Staiger "Volatility in School Test Scores: Implications for Test-Based Accountability Systems" in Diane Ravitch (ed.) Brookings Papers on Education Policy, 2002 (Washington DC: Brookings Institution, 2002b).

Kane, Thomas J. and Douglas O. Staiger "Racial Subgroup Rules in School Accountability Systems", paper presented at the Taking Account of Accountability Conference, Kennedy School of Government, June 2002c.

Kane, Thomas J., Douglas O. Staiger and Jeffrey Geppert. "Assessing the Definition of Adequate Yearly Progress in the House and Senate Education Bills," working paper, July 15, 2001 (revised version published in *Education Next*, 2002).

Keller, Bess "Controversy Surrounds Release of Maryland Test Results" Education Week , February 6, 2002.

Klein, Stephen P., Laura Hamilton, Daniel McCaffrey and Brian Stecher "What Do Test Scores in Texas Tell Us?" Education Policy Analysis Archives (2000) Vol. 8, No. 49, pp. 1-22.

Koretz, Daniel. "Limitations in the Use of Achievement Tests as Measures of Educators' Productivity" Unpublished paper, Rand Corporation, May 2000. Paper initially presented at "Devising Incentives to Promote Human Capital", National Academy of Sciences Conference, December 1999. (Forthcoming in the *Journal of Human Resources.*)

Koretz, Daniel M. and S.I. Barron, "The Validity of Gains on the Kentucky Instructional Results Information System" (Santa Monica: RAND Corporation, 1998).

Koretz, Daniel, S. Barron, K. Mitchell and B. Stecher "The Perceived Effects of the Kentucky Instructional Results Information System" (Santa Monica: RAND Corporation, 1996), Report Number MR-792-PCT/FF.

Krueger, Alan B. "Economic Considerations and Class Size" *NBER Working Paper 8875*, April, 2002.

Lavy, Victor. "Paying for Performance: The Effect of Teacher's Financial Incentives on Students' Scholastic Outcomes" Working Paper, Hebrew University of Jerusalem, July 2002a.

Lavy, Victor. "Evaluating the Effect of Teacher Performance Incentives on Student Achievement" *Journal of Political Economy* (forthcoming, 2002b).

Linn, Robert and S.B. Dunbar "The Nation's Report Card Goes Home: Good News and Bad About Trends in Student Achievement" Phi Delta Kappan (1990) Vol. 72, No. 2, pp. 127-133.

Manzo, Kathleen Kennedy "Testing Glitch Prompts N. Carolina to Order System Audit" Education Week, May 30, 2001.

Murnane, Richard J., John B. Willett and Frank Levy, "The Growing Importance of Cognitive Skills in Wage Determination" Review of Economics and Statistics Vol. 77, No. 2 (May 1995): 251-266.

Neal, Derek and William Johnson "The Role of Premarket Factors in Black-White Wage Differentials" Journal of Political Economy (1996) Vol. 104, pp. 869-895.

New Jersey Department of Education "Educational Assessment Program: State Report 1975-76" 1976.

Pavcnik, Nina, "Trade Liberalization, Exit, and Productivity Improvements: Evidence From Chilean Plants, " Review of Economic Studies (2002) Vol. 69, pp. 245-76.

Plous, Scott The Psychology of Judgement and Decision-Making (New York: McGraw Hill, 1993).

Sandham, Jessica "California Score Glitch Throws Wrench into Bonus Plan" Education Week, October 10, 2001.

Shepard, Lori "The Harm of Measurement-Driven Instruction" Paper presented at the annual meeting of the American Educational Research Association, Washington, DC April, 1988.

Stecher, Brian and S. Barron "Quadrennial Milepost Accountability Testing in Kentucky" CSE Technical Report No. 505, (Los Angeles: Center for the Study of Evaluation, Standards and Testing, 1999).

Steinberg, Jacques and Diana Henriques "When a Test Fails the Schools, Careers and Reputations Suffer" New York Times May 21, 2001

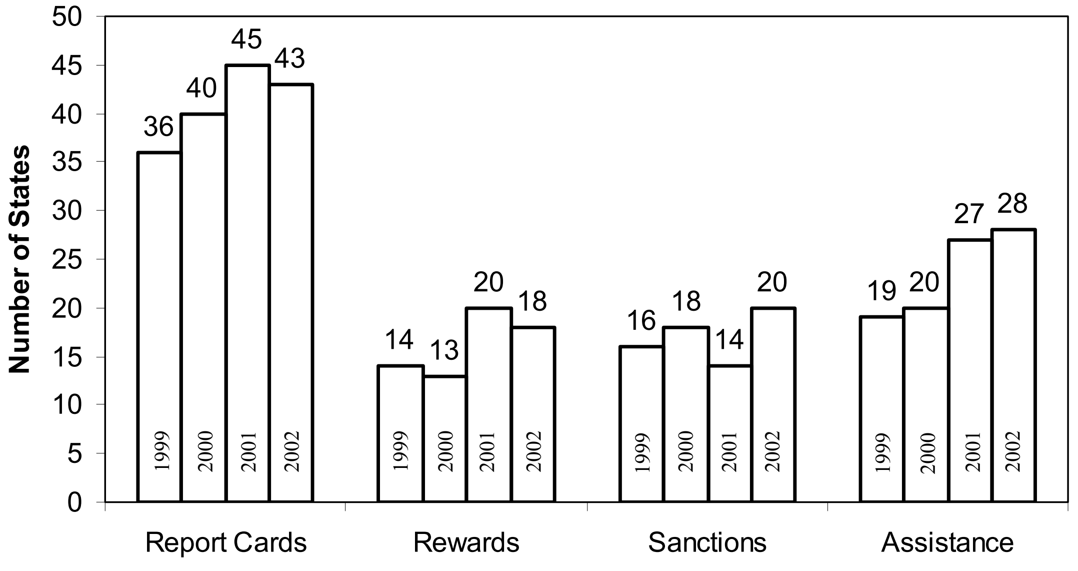
Steinberg, Jacques and Diana Henriques "Right Answer, Wrong Score: Test Flaws Take Toll" New York Times May 20, 2001

Tarcy, Brian "Town's Scores the Most Improved" *Boston Globe*, December 8, 1999, p. C2.

Tybout, James, "Plant- and Firm-level Evidence on 'New' Trade Theories," in James Harrigan (ed.) Handbook of International Trade (Basil Blackwell), forthcoming.

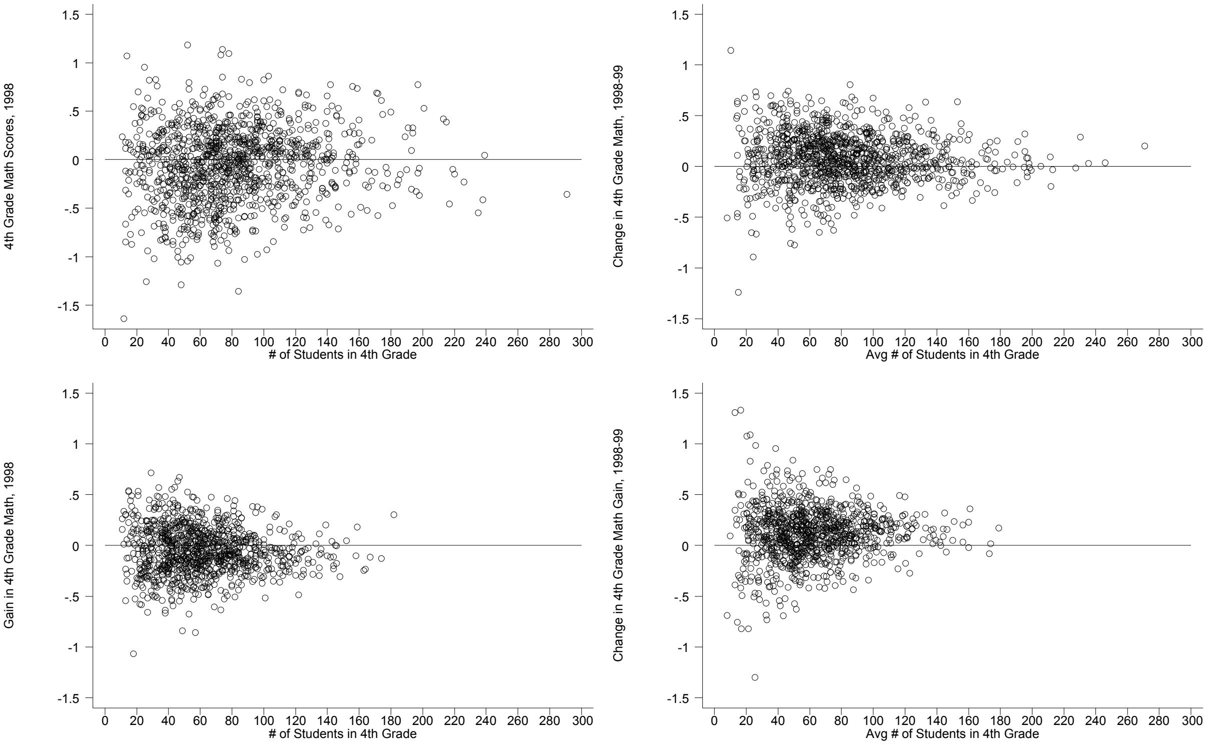
Wilgoren, J. "Possible Cheating Scandal is Investigated in Michigan" The New York Times, June 9, 2001.

Figure1. Trends in State Accountability Systems



Source: Education Week, 2002.

Figure 2



4th Grade Math Scores and Value-Added by School Size

Figure 3

Correlation in School Test Scores

4th Grade Math Levels and Gains

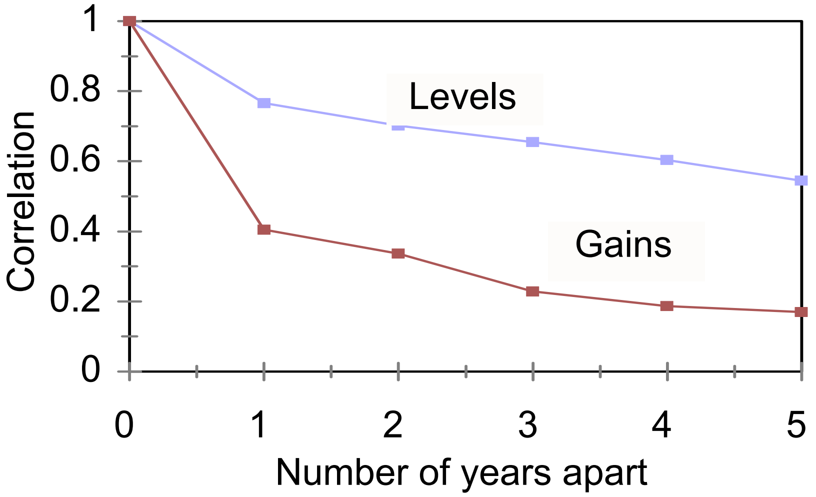
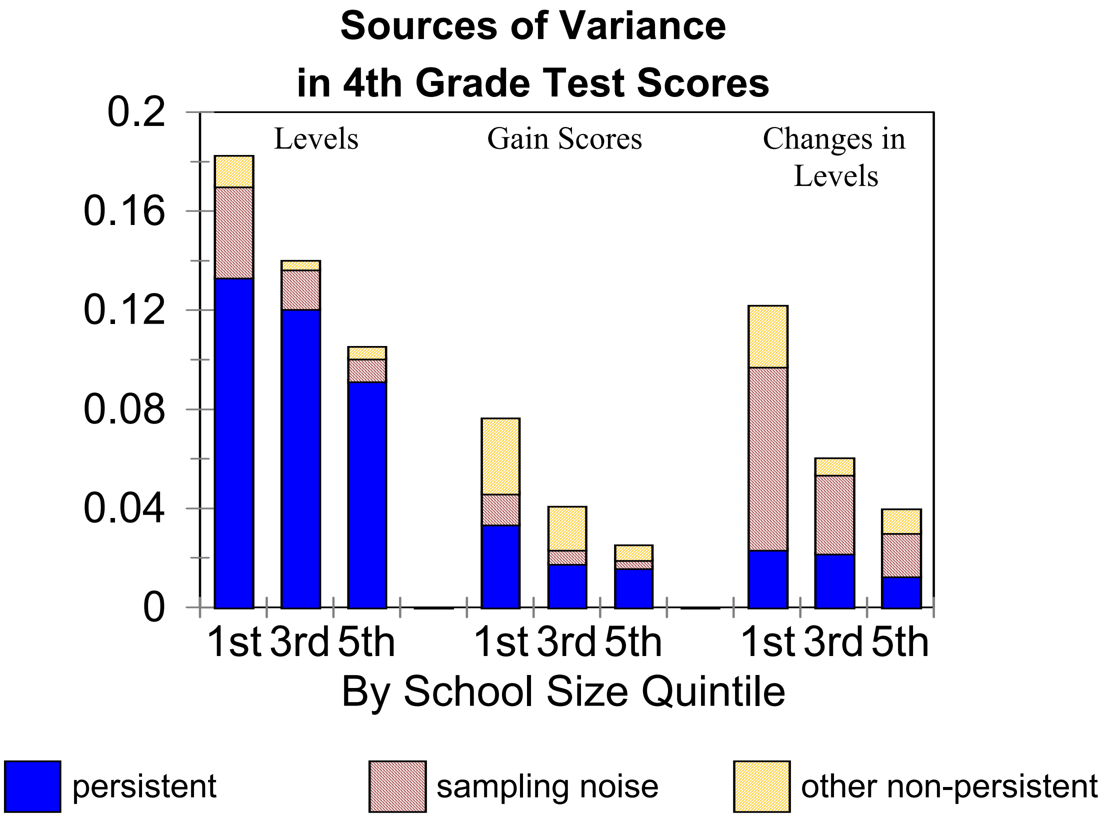
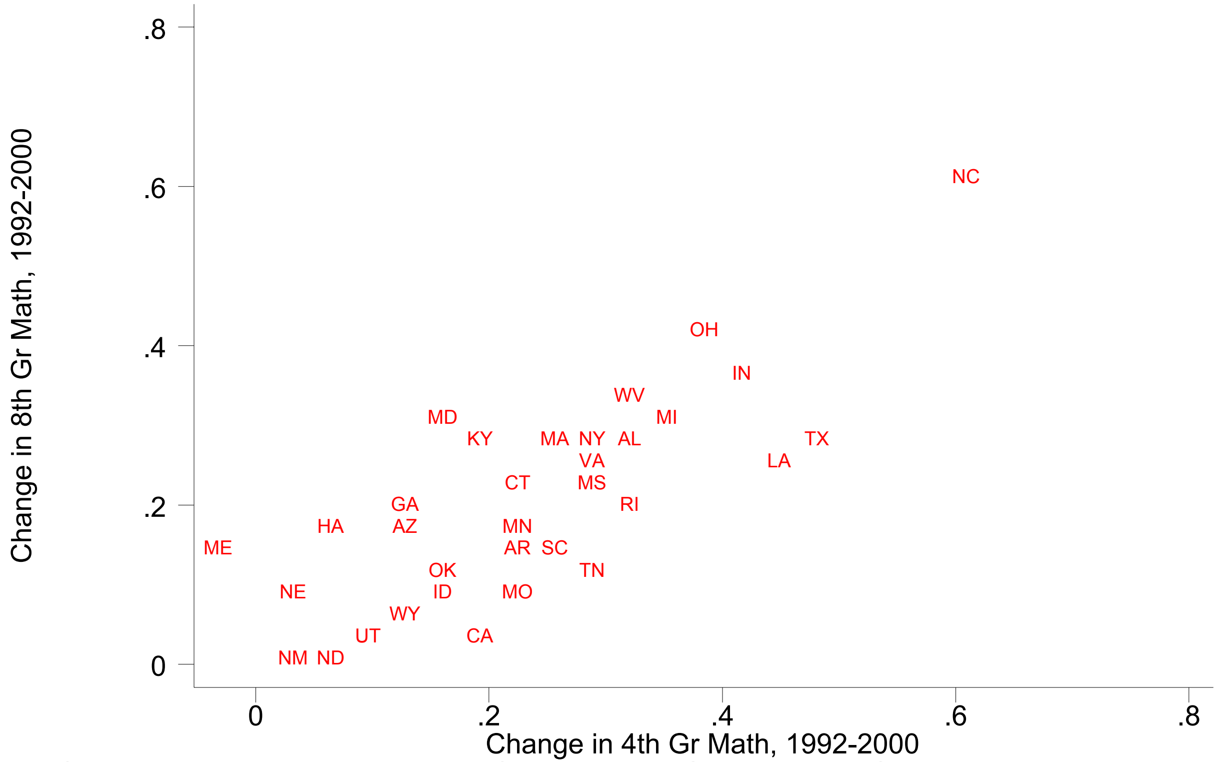


Figure 4



Sources of Variance in 4th Grade Test Scores

Figure 5



Changes in NAEP Math Scores by State and Grade, 1992-2000

Table 1.
Awards and Sanctions Among Elementary Schools in North Carolina

School Size	Percent Ever "Top 25" 1997-2000	Percent Ever Assigned Assist. Team 1997-2000	Mean Gain in Math	Between-School Variance in Mean Gain in Math
Smallest Decile	27.7%	1.2%	.020	.048
2nd	11.8	4.7	-.007	.030
3rd	8.2	7.1	.008	.028
4th	3.6	1.2	.009	.026
5th	2.4	0	-.002	.024
6th	3.6	0	.019	.018
7th	4.8	0	.007	.016
8th	7.1	0	.006	.016
9th	0	1.2	-.007	.015
Largest Decile	1.2	0	-.011	.011
Total	7.0	1.5	.004	.023

Note: The above refers to the 840 regular public elementary schools for whom we had data from 1994 through 2000. Charter schools are not included.

Table 2.
Proportion of California Elementary Schools Winning Governor's Performance Awards by School Size and Number of Numerically Significant Subgroups

Proportion Winning
(Average Growth in API 1999-2000)
[# of Schools in Category]

	# of Numerically Significant Subgroups				Total:
	1	2	3	4+	
Smallest	.824 *(33.4)* [204]	.729 *(45.6)* [343]	.587 *(42.2)* [349]	.471 *(36.0)* [51]	.683 *(41.2)* [947]
2nd	.886 *(29.9)* [158]	.769 *(42.6)* [337]	.690 *(42.2)* [358]	.670 *(43.9)* [94]	.749 *(40.5)* [947]
3rd	.853 *(26.8)* [156]	.795 *(36.3)* [308]	.708 *(38.9)* [390]	.667 *(44.6)* [93]	.756 *(36.6)* [947]
4th	.903 *(28.0)* [144]	.823 *(41.8)* [328]	.776 *(39.5)* [379]	.656 *(40.8)* [96]	.799 *(38.7)* [947]
Largest	.876 *(29.5)* [89]	.776 *(37.9)* [370]	.726 *(36.9)* [387]	.686 *(40.5)* [102]	.755 *(37.0)* [948]
Total:	.864 *(29.8)* [751]	.778 *(40.9)* [1686]	.699 *(39.9)* [1863]	.647 *(41.7)* [436]	.749 *(38.8)* [4736]

Note: Reflecting the rules of the Governor's Performance Award program, the above was limited to elementary schools with more than 100 students.

Endnotes

1.One early exception was a paper by Clotfelter and Ladd (1996), which evaluated the impact of accountability systems in Dallas and South Carolina.

2.The information on school accountability systems presented in this section is derived from Education Week's *Quality Counts* reports from various years, in combination with our own compilation of information from state web sites.

3.School performance differences are at least partially capitalized in local housing values, so the publication of school test scores may have an impact on teacher incentives through parental pressure – albeit incompletely, since each parent would be tempted to free-ride on other parents' efforts. For example, Black (1999) evaluated the effect of school test scores on housing values, using differences in housing values near boundaries of school assignment zones. Her results suggest that a one student-level standard deviation in school test scores is associated with a $1,941 difference in housing value.

4.Much of the data on school level test scores used in this paper can be downloaded from the California Department of Education (http://www.cde.ca.gov/psaa/api/), the Texas Education Agency (http://www.tea.state.tx.us/perfreport/account/) and from the North Carolina Department of Public Instruction (http://www.ncpublicschools.org/abcs/). To calculate student-level gain scores, we obtained student-level data from the North Carolina Department of Public Instruction.

5.In the 1999 Common Core of Data, among schools with a 4th grade classroom, the median school contained 69 students in the 4th grade and the mean number of students was 74.

6.If test scores (y_t) are identically and independently distributed, then $\text{corr}(y_t\text{-}y_{t-1},y_{t-1}\text{-}y_{t-2}) = \text{cov}(y_t\text{-}y_{t-1},y_{t-1}\text{-}y_{t-2})/\text{var}(y_t\text{-}y_{t-1}) = -\sigma_y^2/(2\sigma_y^2)= -0.5$.

7.Formally, suppose that test scores (y_t) are the sum of a fixed component (α), a persistent component (ν_t) which follows a random walk ($\nu_t = \nu_{t-1} + u_t$), and an independent and identically distributed transitory component (ε_t). Then it is straightforward to show that (1) $\text{Var}(\Delta y_t) = \sigma_u^2 + 2\sigma_\varepsilon^2$, and (2) $-2\text{corr}(\Delta y_t,\Delta y_{t-1}) = 2\sigma_\varepsilon^2/(\sigma_u^2 + 2\sigma_\varepsilon^2)$. Thus, -2 times the correlation provides an estimate of how much of the total variation in test score changes is accounted for by the transitory component. This model implies that changes more than one year apart are uncorrelated. In the North Carolina data, we estimate these correlations to be below .05 in absolute value. Note that we can also derive an estimate of how much of test scores (not changes) is accounted for by the transitory component, using $\sigma_\varepsilon^2 = -\text{corr}(\Delta y_t,\Delta y_{t-1})*\text{Var}(\Delta y_t)$. We use a slightly different approach in the results reported below in Figure 4, but the estimates using either method are similar.

8.We estimate the three components as follows, with the calculation done separately for each school size quintile. Let σ^2_y be the total variance across schools in the test score measure. Sampling variance for each school was calculated by dividing the average within-school variance by the school's sample size. The total variance due to sampling variance (σ^2_{samp}) is the sample average of this across schools. The variance due to persistent factors was calculated as $\sigma^2_{pers} = (\sigma^2_y)(\rho_{-1}/\rho_{pers})$, were ρ_{-1} is the correlation of the test measure with a 1-year lag, and ρ_{pers} is an estimate of the first order serial correlation in the persistent component. We estimate ρ_{pers} with the average of (ρ_{-k-1}/ρ_{-k}), for k=1 to 4, i.e. with the average proportional decline in correlation between lags 1 and 5. The variance of the other non-persistent component was estimated as the residual $(\sigma^2_y-\sigma^2_{pers}-\sigma^2_{samp})$. For *changes* in test score levels, sampling variance and other non-persistent variance were estimated by doubling the corresponding estimates from the levels, and the persistent variance was estimated as the residual. See Kane and Staiger (2001) for a more efficient estimator and standard errors.

9.Indeed, the underlying reasons for separate sports leagues for high schools of different size may also be statistics of sampling distributions. A high school with 3000 students is much more likely to find 5 students over 6' 5" in height to form a basketball team than a school with 300 students.

10.Kane and Staiger (2002c) present additional evidence suggesting that racial subgroup rules have had little impact on the performance of minority students in Texas. The improvement in performance for minority students in schools where they were sufficiently numerous to count as a separate subgroup was similar to the improvement for minority students in schools where there were not sufficient numbers of such students to qualify for separate subgroup status.

11.For instance, in our analysis of the North Carolina data, students with more educated parents not only had higher baseline scores, but they gained more from year to year.

12.The No Child Left Behind Act of 2001 requires states to test reading and math skills in all grades 3-8 by the 2005-6 school year. Science tests would not be added until 2007-08, and there are no plans to require states to test other skills.

13.For more examples of test publisher errors, please see Sandham (2001), Keller (2002) and Manzo (2001).

14.Hanushek and Raymond (2002) report that the states establishing accountability systems were also more likely to witness an increase in their special education populations.

15. A somewhat similar issue of the timing of school accountability programs and test results arises in the Clotfelter and Ladd (1996) study the impact of an accountability system in Dallas in 1991. Test scores did rise more quickly in Dallas than in other Texas cities, but the timing of the increase pre-dated the accountability system by one year.

16.In Kane and Staiger (2001), we report that the elementary schools in North Carolina with the largest improvments in mean gain scores did not improve on other measures of student engagement, such as absenteeism, television viewing and time spent on homework.

17.Using similar data, but conditioning on high school grades as well as educational attainment, Grogger and Eide (1995) find that a standard deviation in math scores was associated with a 5 percentage point wage increase for men and 7.5 percentage point increase for women in 1986.

18.The Murnane, Willett and Levy estimates may also differ because they include only the math test score measure and not the composite measure of reading and math skills.

19. The correlation between test scores and earnings is not simply reflecting the payoff to innate abilities, since improvements in test scores are also associated with higher earning prospects. Jencks and Phillips (1999) find that a one standard deviation improvement in math scores between 10th and 12th grade was associated with a 26 percent increase in earnings 10 years after high school graduation.

20.We used the following calculation, incorporating productivity growth as suggested by Krueger (2002):

$$\text{PV at Age 18} = \sum_{\text{i=1}}^{46} \beta \text{w}_{\text{i}} \left(\frac{(1+\gamma)}{(1+r)} \right)^{i-1}$$

where: β is the proportional rise in wages associated with a given test score increase; w_i represent wages from age 18 through 64 estimated using full-time, year-round workers in the 2000 CPS; γ represents the general level of productivity growth, assumed to equal 0.01; and r is the discount rate, assumed to equal 0.03.

21.The School Site Employee Bonus program provided $591 per full-time equivalent teacher to both the school and teacher, or $59 per student based on an average of 20 students per teacher. The Governor's Performance Award (GPA) program provided an additional $63 per student. The growth target for the average elementary school was 9 points on the state's Academic Performance Index (API). Because the state did not publish a student-level standard deviation in the API scores, we had to infer it. A school's API score was a weighted average of the proportion of students in each quintile of the national distribution on the reading, math, language and spelling sections of the Stanford 9 test. For elementary schools, the average proportion of students across the four tests in each quintile (from lowest to highest) was .257, .204, .166, .179 and .194 and the weights given to each quintile were 200, 500, 700, 875 and 1000. Under the assumption that students scored in same quintile on all four tests, we could calculate the student-level variance as $.257(200\text{-}620)^2+.204(500\text{-}620)^2+.166(700\text{-}620)^2+.179(875\text{-}620)^2+.194(1000\text{-}620)^2=89034$, implying a standard deviation of 298. This is nearly 5 times the school-level variance, which is roughly consistent with expectations.