

Volatility in School Test Scores: Implications for Test-Based Accountability Systems

by

Thomas J. Kane
School of Public Policy and Social Research
UCLA
tomkane@ucla.edu
(310) 825-9413

Douglas O. Staiger
Department of Economics
Dartmouth College
douglas.o.staiger@dartmouth.edu
(603) 646-2979

August, 2001

“Errors, like straws, upon the surface flow;
He who would search for pearls must dive below.”

-John Dryden (1631-1700), *All for Love*

I. Introduction

By the spring of 2000, forty states had begun using student test scores to rate school performance. Twenty states are going a step further and attaching explicit monetary rewards or sanctions to a school's test performance. For example, California plans to spend \$677 million on teacher incentives this year, providing bonuses of up to \$25,000 to teachers in schools with the largest test score gains. In this paper, we highlight an under-appreciated weakness of school accountability systems-- the volatility of test score measures-- and explore the implications of that volatility for the design of school accountability systems.

The imprecision of test score measures arises from two sources. The first is sampling variation, which is a particularly striking problem in elementary schools. With the average elementary school containing only 68 students per grade level, the amount of variation due to the idiosyncracies of the particular sample of students being tested is often large relative to the total amount of variation observed between schools. A second source of imprecision arises from one-time factors that are not sensitive to the size of the sample: a dog barking in the playground on the day of the test, a severe flu season, one particularly disruptive student in a class or favorable "chemistry" between a group of students and their teacher. Both small samples and other one-time factors can add considerable volatility to test score measures.

Initially, one might be surprised that school mean test scores would be subject to such fluctuations, since one would expect the averaging of student scores to dampen any noise. Although the averaging of students' scores does help dampen volatility, even small fluctuations in a school's score can have a large impact on a school's ranking, simply because schools' test scores do not differ dramatically in the first place. This reflects the longstanding finding from

the Coleman report (*Equality of Educational Opportunity*, issued in 1966), that less than 16 percent of the variance in student test scores is between-schools rather than within-schools. We estimate that the confidence interval for the average 4th grade reading or math score in a school with 68 students per grade level would extend from roughly the 25th to the 75th percentile among schools of that size.

Such volatility can wreak havoc in school accountability systems. First, to the extent that test scores bring rewards or sanctions, school personnel are subjected to substantial risk of being punished or rewarded for results beyond their control. Moreover, to the extent such rankings are used to identify best practice in education, virtually every educational philosophy is likely to be endorsed eventually, simply adding to the confusion over the merits of different strategies of school reform. For example, when the 1998-99 MCAS test scores were released in Massachusetts in November of 1999, the Provincetown district 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 between 1998 and 1999, interviewing the high school principal and several teachers.¹ As it turned out, they had changed a few policies at the school-- decisions that seemed to have been validated by the improvement in performance. One had to dig a bit deeper to note that the Provincetown high school had only 26 students taking the test in 10th grade. Given the wide distribution of test scores among students in Massachusetts, any grouping of 26 students is likely to yield dramatic swings in test scores from year to year, that is large relative to the distribution of between-school

¹Brian Tarcy, "Town's Scores the Most Improved" *Boston Globe*, December 8, 1999, p. C2.

differences. In other words, if the test scores from one year are the indicator of a school's success, the *Boston Globe* and similar newspapers around the country will eventually write similar stories praising virtually every variant of educational practice. It is no wonder that the public and policymakers are only more confused about the way to proceed.

The remainder of the paper is divided into several parts. In Section II, we describe our data sources. Section III presents a non-technical description of the variation in school test scores. The discussion proceeds in four steps: First, we discuss the importance of sampling variation in mean test scores and in mean gain scores. Second, we provide estimates of the total variance in test scores from one year to another. Third, we provide some estimates of the proportion of this variance in single-year changes in test scores that is persistent, rather than transient. Fourth, we put all the pieces together and decompose the variation in test scores into persistent and non-persistent factors. For those interested in a more precise discussion of the technical details, similar results are presented in Kane and Staiger (2001).

The accountability system one would design if test scores were stable measures of a school's performance might look very different from the system one would choose in light of the volatility of test scores. In Section IV, we draw four lessons regarding the design of test-based accountability systems, given the statistical properties of school-level test scores. First, states should not reward or punish those schools with test scores at either extreme. Such rules primarily effect small schools, that are most likely to observe large fluctuations in test scores. For instance, in one component of its accountability system, North Carolina recognizes the 25 schools with the top scores on that state's "growth composite". Our results suggest that the smallest tenth of elementary schools were 23 times more likely than the largest tenth of schools

to win such an award, even though their average performance was no better. Due to larger random fluctuations in small schools, large schools are simply unlikely to ever have test scores among the top 25 or bottom 25 in any year. Because large schools are unlikely to have test scores at the very top or bottom of the distribution, such incentives do not affect them.

Second, the requirement that schools achieve minimum gains for all of its racial or ethnic subgroups before winning an award greatly disadvantages diverse schools. With the support of civil rights groups, a number of states (including California and Texas) require some minimum performance threshold for all racial and ethnic subgroups in a school. The U.S. Congress included similar provisions in the recent re-authorization of the Elementary and Secondary Education Act. However, as we illustrate, such rules can actually harm the intended beneficiaries—disadvantaged minorities. Elementary schools in California with 4 or more racial or ethnic subgroups were much less likely than schools with only one racial subgroup to win a Governor’s Performance Award, even though their improvements in performance were slightly larger. The reason is simple: to the extent that there is any “luck of the draw” involved in achieving an increase in performance from one year to the next, diverse schools must be sufficiently lucky on each of multiple draws to win an award, while schools with only one racial subgroup have only to win on one draw. Because disadvantaged minorities are more likely to attend racially diverse schools, their schools on average are more likely to be arbitrarily left out of the award money. Moreover, such rules provide incentives for schools to segregate by race.

Third, rather than use a single year’s worth of test scores, states should pool test scores over multiple years in order to identify exemplars and minimize the importance of chance fluctuations. In earlier work, we have proposed sophisticated ways of pooling information over

time that take account of the different amounts of noise in test scores in small and large schools. We illustrate the value of such techniques in identifying differences in performance among schools.

Finally, such fluctuations make it difficult to evaluate the impact of policies that are targeted at high- or low-scoring schools. For instance, North Carolina provides assistance teams to schools with the lowest scores in a given year. In the past, many of the schools assigned such assistance teams experienced large increases in performance subsequently. Rather than reflecting the effect of the assistance teams alone, such test score increases may simply reflect volatility. We illustrate how the effect of assistance teams can be overstated unless one carefully accounts for the natural volatility in test scores.

II. Sources of Data

We obtained math and reading test scores for nearly 300,000 students in grades 3 through 5, attending elementary schools in North Carolina each year between the 1992-93 and 1998-99 school years. (The data were obtained from the N.C. Department of Public Instruction.) Although the file we received had been stripped of student identification numbers, we were able to match a student's test score in one year to their test score in the previous year using day of birth, race and gender.² In 1999, 84 percent of the sample had unique combinations of birth date, school and gender. Another 14 percent shared their birth date and gender with at most 1 other

²In addition, the survey contained information on parental educational attainment reported by students. Given changes in student responses over time, we did not use parental education to match students from year to year, although we did use these data to when attempting to control for the socioeconomic background of students.

students in their school and grade and 2 percent shared their birth date with 2 other people. (Less than 1 percent shared their birth date and gender with 3 or more students in the school and no match was attempted for these students.) Students were matched across years only if they reported the same race. If there was more than 1 person with the same school, birth date, race and gender, we looked to see whether there were any unique matches on parental education. If there was more than one person that matched on all traits-- school, birth date, race, gender and parental education-- the matches that minimized the squared changes in student test scores were kept.

However, because of student mobility between schools and student retention, the matching process was not perfect. We were able to calculate test score gains for 65.8 percent of the 4th and 5th grade students in 1999. (The matching rate was very similar in other years.) Table 1 compares the characteristics of the matched and the non-matched sample of fourth and fifth grade students in 1999. The matched sample had slightly higher test scores (roughly .2 student level standard deviations in reading and math), a slightly higher proportion female, a slightly higher proportion black and Hispanic, and a slightly lower average parental education than the sample for which no match could be found.

In most of the discussion, we employ the test scores in reading and math used by the North Carolina Department of Public Instruction. However, a one-unit change in such scores does not have any intuitive reference point. To provide readers with an intuitive sense of magnitude of such scores, we subtracted the mean and divided by the standard deviation in scores in each grade to restate test scores in terms of student-level standard deviations from the mean. However, because we used the overall mean and the overall standard deviation for the whole

period 1994 through 1999 to do the standardization, we allow for changes over time in the distribution of scaled scores. We also calculated student-level gains by taking the differences in these scores (standardized in the above way) from one year to the next. As a result, both test score levels and test score gains are in units of student-level standard deviations in levels. We also experimented with using “quasi-gains”, by first regressing a student’s score on his or her score in the previous year, and taking the residuals as a measure of student improvements. However, because the results were similar, we are only reporting the results using gain scores and test score levels here.

In our previous work (Kane and Staiger (2001)), we also adjusted each individual student’s score for race, gender and parental education. That has the effect of removing between-school differences due to differences in race and parental education. Throughout most of this paper, we use the unadjusted test score data. The exception is analysis presented in Tables 5 and 6, which report the results of our filtering technique.

We also use school and grade level data on California’s Academic Performance Index scores in 1998 through 2000. The Academic Performance Index is based upon school-level scores on the Stanford 9 tests. Schools receive 1000 points for each student in the top quintile, 875 points for students in the next quintile, 700 points for students in the middle quintile, 500 points for students in the 20th to 39th percentiles and 200 points for students in the bottom quintile. A school’s average is based upon an equally weighted average of their scores in the Reading, Spelling, Language and Mathematics portions of the Stanford 9 tests.³ We use the

³For a more detailed description of the California Academic Performance Index, see California Department of Education, Policy and Evaluation Division, “2000 Academic Performance Index Base Report Information Guide”, January 2001.

California data to highlight the generality of the measurement issues we describe and to analyze some of the properties of that state's accountability system.

III. Sources of Volatility in School-Level Test Scores

In this section, we illustrate three characteristics of school-level test score measures that are vital to the design of test-based accountability systems. The discussion is generally non-technical and intuitive and, as a result, takes up more space than technical shorthand would allow. Nevertheless, our purpose is to provide a primer for our subsequent discussion of some of the flaws in the design of test-based accountability systems that do not take account of test volatility.

We emphasize three fundamental characteristics of school-level test measures: First, there is a considerable amount of variation in test scores at the school level due to sampling variation. Each cohort of students that enters first grade is analogous to a random draw from the population of students feeding a school. Even if that population remains stable, performance will vary depending upon the specific group of students reaching the appropriate age in any year. Using standard sampling theory, we can directly estimate the amount of variation we would expect to occur. Given that there are only 68 students per grade level in the typical elementary school, such variation can be substantial.

Second, there are other factors producing non-persistent changes in performance in addition to sampling variation. Possible sources of such variation would be a dog barking in the parking lot on the day of the test, a severe flu season, the chemistry between a particular group of students and a teacher, a few disruptive students in the class or bad weather on test day.

Although we cannot estimate the magnitude of this second source of variation directly without explicitly monitoring each influence on scores, we can do so indirectly, by observing the degree to which any changes in test scores from year to year persist and, thereby, infer the total amount of variation due to non-persistent factors. Any non-persistent variation in test scores that is not due to sampling variation we put into this second category.

The third fact we highlight is that by focusing on mean gains in test scores for students in a given year or changes in mean test score levels from one year to the next, many test-based accountability systems are relying upon quite unreliable measures. Schools differ very little in their rate of change in test scores or in their mean value-added-- certainly much less than they differ in their mean test score levels. Moreover, those differences that do exist are often non-persistent-- either due to sampling variation or other causes. For instance, we estimate that more than 70 percent of the variance in changes in test scores for any given school and grade is transient. For the median-sized school, roughly half of the variation between schools in gain scores (or value-added) for any given grade is also non-persistent.

Below, we describe each of these facts below in more detail, before proceeding to a discussion of their implications for the design of test-based accountability systems.

Sampling Variation

A school's mean test score will vary from year-to-year, simply because the particular sample of students in a given grade differs. But just how much it varies depends upon two things: the variance in test scores in the population of students from which a school is drawing and the number of students in a particular grade. In schools where the students are particularly heterogeneous or in schools with a small number of students in each grade, we would expect test

scores to fluctuate more.

In 1999, there were nearly one thousand schools in North Carolina with students in the fourth grade. Averaging across these schools (and weighting by school size), the variance in math scores among students in a given school was nearly nine-tenths as large (.87) as the student-level variance in scores; the ratio of the average within-school variance in 4th grade reading scores to the total variance in reading scores was .89. In other words, the heterogeneity in student scores within the average school was nearly as large as the heterogeneity in scores overall.

This is not some idiosyncratic characteristic of North Carolina's school system. Rather, it reflects a long-standing finding in educational assessment research. In their classic study of inequality of student achievement published in 1966, James Coleman and his colleagues estimated that only between 12 and 16 percent of the variance in verbal achievement among white 3rd grade students was due to differences across-schools; the remainder was attributable to differences within schools.⁴ In other words, two students drawn at random from within a given school are likely to differ nearly as much as two students drawn at random from the whole population.

Applying the rules from elementary sampling theory, one would simply divide the average within-school variance by the sample size in order to calculate the expected variance in the mean test score for a given school due to sampling variation. According to the National Center for Education Statistics, there were 68 students per grade level on average in schools

⁴Coleman et. al. (1966), p. 326.

serving the elementary grades.⁵ Dividing .87 and .89 respectively by 68, we would expect a variance of .013, due simply to the effect of drawing a new sample of students.

In North Carolina elementary schools near the national average in size (between 65 and 75 students with valid test scores), the variance in mean reading and math scores was .087 and .092 respectively. Dividing the estimated amount of variance due to sampling variation for a school of average size (.013) by the total variance observed for such schools, we would infer that 14 to 15 percent of the variation in 4th grade math and reading test scores was due to sampling variation.

It is sometimes difficult to gain a strong intuitive sense for the magnitude of sampling variation with a proportion of variance calculation. An alternative way to gauge the importance of sampling variation would be to calculate the 95 percent confidence interval for a school's mean test score. One would do so by adding and subtracting 1.96 times the standard error of the estimate for the mean ($\sqrt{.013}$), which is equal to .223 student-level standard deviations.

Among schools with between 65 and 75 students with valid test scores, such a confidence interval would extend from roughly the 25th to the 75th percentile.

Sampling Variation and Mean Gain Scores Across Schools

North Carolina-- like a handful of other states including Arizona, Florida and Tennessee--

⁵In 1996-97, there were 51,306 public schools in the U.S. with a 3rd grade. (*Digest of Education Statistics*, 1998, Table 99, p. 119) This included 4910 schools with grades PK, K or 1st through 3rd or 4th, 20570 schools with PK, K or 1st through grade 5, 15578 schools with PK, K or 1st through grade 6, 4543 schools with PK, K or 1st through grade 8 and 5705 schools with other grade spans. Moreover, in the fall of 1996, there were 3.518 million students in the U.S. enrolled in 3rd grade. (*Digest of Education Statistics*, 1998, Table 43, p. 58)

rates its schools by focusing on the average gain in performance among students attending a particular school.⁶ Advocates tout the value-added methodology as a more “fair” method of ranking schools, by explicitly adjusting for the fact that some students enter school with higher scores than others. However, to the extent that schools differ less in their “value-added” than in their test score levels, such measures can be particularly vulnerable to sampling variation.

The point can be illustrated with a few simple calculations. The variance in the gain in test performance between the end of 3rd grade and the end of 4th grade within the average school in North Carolina was .331 in math and .343 in reading (stated in terms of the student-level standard deviation in 4th grade math and reading test scores). Note that the variance in gains is smaller than the variance in test scores in 4th grade (or 3rd grade), despite the fact that one is taking one imperfect measure of a child’s performance and subtracting another. Indeed, the variance in gains is roughly four-tenths as large as the variance in fourth grade scores within schools (.331/.87 and .343/.89). If there were no relationship between a student’s third grade score and fourth grade score, we would expect the variance to double when taking the difference. However, because third and fourth grade performance for a given student has a correlation coefficient of approximately .8, the variance in the gain is only roughly four-tenths as large as the variance in the test score levels.

To calculate the variance in test scores we would expect to result from sampling variation for a school of average size, we would simply divide the within-school variance in gain scores (.331 and .343) by the sample size (68), yielding an estimate of .0049 for math and .0050 and

⁶Darcia Harris Bowman, “Arizona Ranks Schools by ‘Value-Added’ to Scores”, *Education Week*, February 9, 2000.

reading. However, while the within-school variance in gains between 3rd and 4th grades is four-tenths as large as the within-school variance in test scores in 4th grade, the amount of variance between-schools drops even more when moving from mean test scores to mean gains-- at least for reading scores. Among schools with 65 to 75 students, the variance in reading scores was .015. In other words, the between-school variance in mean student gains among schools of roughly the average size is only one-fifth as large as the between-school variance in mean 4th grade scores. Yet, the variance between schools due to sampling variation is two-fifths as large. As a result, the share of variance between schools in mean reading gain scores that is due to sampling variation is double that seen with mean reading score levels. Sampling variation makes it much harder to discern true differences in reading gain scores across schools.⁷

Sampling Variation in Small and Large Schools

In all of the above calculations, we limited the discussion to schools close to the national average in size. Sampling variation will account for a larger share of the between school variance for small schools and a smaller share for large schools. For Figure 1, we sorted schools in North Carolina by the number of test-takers and divided the sample into 5 groups by school size. We then calculated the variance between schools in each quintile in test scores. We did so for 4th grade math and reading and for gains in scores between 3rd and 4th grade in math and reading.

Several facts are evident in Figure 1. First, for each measure, we observed much more

⁷The impact on math gain scores is less pronounced, given greater variability in mean math gain scores between schools.

variance in test scores among smaller schools than among larger schools. For math and reading test scores, the variance between schools was roughly fifty percent larger for the smallest quintile of schools than for the largest quintile. For math and reading gain scores, the between school variance was roughly three times as large for the smallest quintile of schools than for the largest quintile.

Second, the dotted line in each panel of Figure 1 identifies the between-school variance in each quintile after subtracting our estimate of the sampling variation. The sampling variation we estimated above accounts for some portion of the greater variation among smaller schools, but even after subtracting our estimates of the sampling variation, the between-school variance is greater for smaller schools.

It is worthwhile noting that we ignored any “peer effects” in our estimate of the sampling variance above. In other words, we assumed that having a disproportionate number of high- or low-test score youth would have no direct effect on the performance of other students in the class. However, if there were peer effects (for instance, if having a disproportionate share of low-performing youth pulls down the average performance of others or having a large number of high performing youth raises the performance of all students through the quality of class discussions), we might expect the effects of any sampling variation to be amplified. If peer effects exist, we are understating the importance of sampling variation.

The “peer effect” need not operate through student test scores, however. A similar phenomenon would occur if any other characteristic that varied across samples had a direct effect on student test scores. For instance, Hoxby (2001) identifies substantial negative impacts on student performance from having a disproportionate share of boys in one’s cohort. Any time that

a characteristic of the sample has a direct effect on the performance of each of the individuals in that sample, our estimates of the magnitude of variance due to sampling variation are likely to be understated.

Third, there was very little variance between schools in the mean gain in reading scores between 3rd and 4th grade. Indeed, even for the smallest quintile of schools, the between-school variance in the mean gain in reading performance was equal to .05 student-level standard deviations in 4th grade reading scores. Moreover, a large share of this is estimated to have been due to sampling variation.

Small sample size is a particularly large problem for elementary schools. However, the problem is not unique to elementary schools. Figure 2 portrays the distribution of sample sizes by grade in North Carolina. School size is generally smaller in grade 4. However, there is much more uniformity in school size among elementary schools. While the size of the average middle school is larger than the size of the average elementary school, there is also more heterogeneity in school size among middle schools. The same phenomenon is exaggerated at grade 9. High schools are generally much larger than elementary schools, but there are still quite a few small schools enrolling 9th grade students. In other words, elementary schools tend to be smaller than middle schools and high schools. However, they are also more uniform in size, meaning that schools have a more similar likelihood of having an extremely high or extremely low score due to sampling variation. Middle schools and high schools have larger sample sizes on average, but there is greater heterogeneity between schools in the likelihood of seeing an extremely high or extremely low test score due to sampling variation.

Variation in the Change in Test Scores over Time

The greater variability in test scores among small schools is not simply due to long-term differences among these schools (such as would occur if all large schools were found in urban settings and if small schools contained a mixture of suburban and rural schools). Indeed, test scores also fluctuate much more from year-to-year among small schools than among large schools. Figure 3 plots the variance in the change in test scores between 1998 and 1999 by school size in North Carolina. The panel on the left portrays the variance in the change for 4th grade math and reading scores and for gains in math and reading scores. The dotted line in both figures represents the result of subtracting our estimate of the contribution of sampling variation to the variance in the change. The variance in the change in 4th grade test scores was 3 times as large among the smallest quintile of schools than among the largest quintile of schools (.079 vs. .027). Moreover, the variance in the change in 4th grade gain scores was 5 times as large among the smallest quintile of schools than among the large schools (.060 vs. .013).

A Measure of the Persistence of Change in School Test Scores

Sampling variation is only one reason that a school might experience a change in test scores over time. There may be sources of variation at the classroom level, generated by teacher-turnover, classroom chemistry between a teacher and her class, the presence of one particularly disruptive students in a class. Indeed, there may be sources of variation that affect a whole school, such as a dog barking in the parking lot on the day of the test or inclement weather, that would generate temporary fluctuations in test performance. We can estimate the amount of variation due to sampling variation by assuming that the succession of cohorts within a particular

grades is analogous to a random sampling process. However, we have no similar method of modeling these other sources of variation and anticipating *a priori* how much variation to expect from these other sources. For instance, we would need a model of the time series process affecting weather over time and have an estimate of the effect of such weather on student test scores to approximate the variation in test scores due to weather changes. We have neither. However, in this section, we provide a simple method for estimating that fraction of the variation in test scores over time which can be attributed to all such non-persistent variation, even if we cannot identify the individual components as neatly. Later, we will subtract off our estimate of the sampling variation to form an estimate of these other sources of non-persistent variation.

Suppose that there were some fixed component of school performance that did not change over time and suppose that there were two categories of fluctuations in school test scores: those that persist and those that are transient or non-persistent. One might describe a school's test performance, S_t , as being the sum of three factors: a permanent component which does not change, α , a persistent component, v_t , which starts where it left off last year, but is subject to a new innovation each year, u_t , and a purely transitory component that is not repeated, ε_t :

$$S_t = \alpha + v_t + \varepsilon_t$$

$$\text{where } v_t = v_{t-1} + u_t$$

One could write the changes from year t-2 to year t-1 and from year t-1 to year t as follows:

$$\begin{aligned}\Delta S_t &= S_t - S_{t-1} = v_t - v_{t-1} + \varepsilon_t - \varepsilon_{t-1} = u_t + \varepsilon_t - \varepsilon_{t-1} \\ \Delta S_{t-1} &= S_{t-1} - S_{t-2} = v_{t-1} - v_{t-2} + \varepsilon_{t-1} - \varepsilon_{t-2} = u_{t-1} + \varepsilon_{t-1} - \varepsilon_{t-2}\end{aligned}$$

Suppose that u_t , u_{t-1} , ε_{t-1} and ε_t are independent.⁸ Then the correlation between the change this year and the change last year could be expressed as below:

$$\rho = \frac{-\sigma_\varepsilon^2}{\sigma_u^2 + 2\sigma_\varepsilon^2}$$

In the above expression, the numerator is the variance in the non-persistent component (with a negative sign attached) and the denominator is the total variance in the change in test scores from one year to the next. With a little algebra, the above equation could be rearranged to produce the equation below:

$$-2\rho = \frac{2\sigma_\varepsilon^2}{\sigma_u^2 + 2\sigma_\varepsilon^2}$$

⁸In this section, we have sacrificed some generality for intuitive appeal. In some cases, it may not be reasonable to expect u_t and u_{t-1} or ε_{t-1} and ε_t to be independent. For instance, if there were “ceiling effects” such that a change one year bumped up against a limit (u_t and u_{t-1} and ε_{t-1} and ε_t were negatively correlated), we would overstate the amount of transience. (It should be noted that we observed no obvious evidence of ceiling effects in the data.) On the other hand, if there were schools that were consistently improving in a systematic way (u_t and u_{t-1} were positively correlated), we would understate the amount of transience. For a more general treatment of the issue, see Kane and Staiger (2001).

The expression on the right side of the equation describes the proportion of the change in test scores that is attributable to non-persistent factors. The expression on the left side of the equation is simply the correlation in the change in test scores in two consecutive years multiplied by negative two. In other words, given an estimate of the correlation in changes in test scores in two consecutive years, we can estimate the proportion of the variance in changes that is due to non-persistent factors by multiplying that correlation by -2 . If the correlation were zero, we would infer that the changes that occur are persistent; if the correlation were close to $-.5$, we would infer that nearly 100 percent the changes that occur are purely transitory, such as sampling variation or a dog barking in the parking lot on the day of the test or inclement weather.

To explore the intuition behind the expression above, suppose that the weather was particularly beautiful, the students were particularly well-rested and one had a unusually talented group of 4th grade students present on test day in 1999. Then the change in test scores for 4th grade students between 1998 and 1999 would be large and positive. Since these factors were one-time phenomena that were unlikely to be repeated in 2000, we would expect a smaller than average change between 2000 and 1999, since we would expect scores in 2000 to be back to the average and because 1999 will still appear as a stand-out year. In other words, if changes were non-persistent, we would expect a negative correlation between the change this year and the change next year. In fact, if all change were transitory, we would expect a correlation of $-.5$.

Now, consider the opposite case, where any change reflected a more permanent improvement in a school's performance. For instance, suppose a school hired a new 4th grade teacher in 1999 and improved facilities, thereby raising test performance. The school may make other such changes in the year 2000, but the magnitude of the change one year provides no

information about the expected magnitude any such changes the next year. They may improve again, and they may decline, but to the extent that all changes are persistent, one would have no reason to *expect* any backsliding. If change in performance serves as the basis for subsequent improvements or declines rather than disappearing, we would expect a correlation of 0 in the change from one year to the next. On the other hand, if some changes are permanent, and some changes are purely transitory, one would expect a negative correlation between 0 and -.5.

To avoid confusion, it is important to keep in mind that the above estimator is focusing only on the transience of any *changes* in performance. There are certainly longstanding differences between schools that do persist over time. But because any fixed trait of a school (α) drops out when we are focusing on changes, any unchanging characteristics are being excluded from our calculations. That is only fitting though, since we are interested in the proportion of change that persists, not the proportion of baseline differences that persist.

We calculated the mean 4th grade scores in North Carolina (combining the scaled scores for math and reading) and calculated the correlation in the change in adjacent years, 1997-98 and 1998-99. We also calculated the mean API scores in California for 4th grade students and again calculated the correlation in the change in adjacent years. Figure 3 reports those correlations for each school size quintile in North Carolina and California. In North Carolina, the correlations ranged between -.25 and -.4. Using the reasoning above, this would imply that between 50 and 80 percent of the variance in the change in mean fourth grade scores is non-persistent. In other words, if one were to look for signs of improvement by closely tracking changes in mean scores from one year to the next, 50 to 80 percent of what one observed would be temporary-- either due to sampling variation or some other non-persistent cause.

Although the California schools tend to be larger, the data reveal slightly more volatility in the California Academic Performance Index for any given school size. For the smallest fifth of schools, the correlation in the change in adjacent years was $-.43$, implying that 86 percent of the variance in the changes between any two years are fleeting. For the largest fifth of schools, the correlation was $-.36$, implying that 72 percent of the variance in the change was non-persistent.

In California, the correlations clearly rise (become less negative) for the larger schools. This is what one would expect if one of the sources of non-persistence were sampling variability. In North Carolina, the pattern is less evident. However, this is presumably because of the smaller number of schools within each size quintile in NC relative to CA.

Schoolwide Scores, Overlapping Cohorts and the Illusion of Stability

Some states, such as California, reward schools based upon changes in the average performance across all grades in a school, rather than on a single grade, as we analyzed above. The use of school-wide averages has two primary effects: First, combining data from different grades increases the sample size and, therefore, reduces the importance of sampling variation. Second, there is considerable overlap in the actual sample of students in a school over a three year period. Failing to take account of such overlap can create the illusion that school improvements are more stable than they actually are. Consider an extreme example in which schools' long-term average performance does not change at all, and where any observed change in test performance is solely due to sampling variation. If we were to perform the exercise above, we would expect to a correlation of $-.5$ in the change in performance in consecutive years for any

given grade level, since any change would be non-persistent. However, suppose we were using the change in a school's combined performance on 4th and 5th grade tests in two consecutive years (the change between years $t-1$ and $t-2$ and the change between year t and $t-1$). Now suppose that the 4th grade cohort from year $t-1$ is a particularly stellar group of kids. If we were only looking at 4th grade students, we would expect that the change from year $t-1$ to t would be smaller than the change from $t-2$ to $t-1$, because a great group of students is unlikely to appear two years in a row. However, because that stellar group of 4th graders in year $t-1$ will repeat again as a stellar group of 5th graders in year t , any fall-off in performance is likely to be muted, because that group is still being counted in a school's test score. When one combines test scores from consecutive grades, one will have an illusion of stability in the year to year improvements, but only because it takes a while for a particularly talented (or particularly untalented) group of students to work their way through the educational pipeline. It is an illusion because it is only after the random draw of students has been made in one year, that there is less uncertainty for the change the subsequent year. A school is either doomed or blessed by the sample of students who enrolled in previous years, but before those cohorts are observed, there is actually a lot of uncertainty.

Figure 4 portrays the correlation in changes in scores in consecutive years when combining two grades that would not overlap in 3 years, 2nd and 5th grade, and when combining two grades that do overlap, such as 4th and 5th grade. Combining 2nd and 5th grade scores is like expanding the sample size. The consecutive year changes are less negatively correlated. The correlation for the largest quintile of schools was approximately $-.3$, implying that 60 percent of the variance in annual changes is non-persistent; the correlation for the smallest quintile was

-37. However, when combining 4th and 5th grade scores, there is a discontinuous jump in the correlation. Rather than having a correlation of -.3, the correlation for all quintiles was close to -.15. Using school-wide averages, combining test scores across grades, leaves the impression of greater stability.

Disaggregating the Variance in Scores into Persistent and Non-Persistent Variation

Table 2 disaggregates the variation in school test scores into two parts: that due to sampling variation and that due to other sources of non-persistent variance. To gain some insight into our accounting, one must recognize that we observe the total variation in mean test scores and mean gain scores among schools of different sizes; we also see the variance in their changes from one year to the next; we have an estimate of the proportion of the change that is due to non-persistent variation; and we have an estimate of the amount of variation we would expect to result from sampling variation. Because sampling variation is by definition non-persistent, we can use all these pieces of information to complete the puzzle and to generate an estimate of the variance due to non-persistent factors other than sampling variation. The top panel of Table 2 decomposes the variance in 4th grade scores in a single year, the middle panel decomposes the variance in the mean gain in scores for students in a particular school and the bottom panel decomposes the variance in the *change* in mean 4th grade scores between years.

Three results in Table 2 are worth highlighting. First, a school's average test performance in 4th grade can be measured rather reliably. Even among the smallest quintile of schools, non-persistent factors account for only 20 percent of the variance between schools. Among the largest quintile of schools, such factors account for only 9 percent of the variance. However,

when using mean test score levels unadjusted for student's incoming performance, much of that reliability may be due to the unchanging characteristics of the populations feeding those schools and not necessarily due to unchanging differences in school performance.

Second, in contrast, mean gain scores or annual changes in a school's test score are measured remarkably unreliably. More than half (58 percent) of the variance among the smallest quintile of schools in mean gain scores is due to sampling variation and other non-persistent factors. Among schools near the median size in North Carolina, non-persistent factors are estimated to account for 49 percent of the variance. Changes in mean test scores from one year to the next are measured even more unreliably. More than three-quarters (79 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, increasing the sample size by combining information from more than one grade will do little to improve the reliability of changes in test scores over time. Even though the largest quintile of schools were roughly 4 times as large as the smallest quintile, the proportion of the variance in annual changes due to non-persistent factors declined only slightly, from 79 percent to 73 percent. As mentioned above, one might have the illusion of greater stability by combining multiple grades, but the illusion is bought at the price of holding schools accountable for the past variation in the quality of incoming cohorts.

Summary

Rather than holding schools accountable for the level of their students' performance in a given year, a growing number of states are rewarding or punishing schools on the basis of

changes in test scores or on mean *gains* in performance. Although either of the latter two outcomes may be closer *conceptually* to the goal of rewarding schools based upon their value-added or rewarding schools for improving student performance, both outcomes are difficult to discern. Schools simply do not differ very much in terms of the change in their performance over time or in terms of the mean gain in performance achieved among their students. Moreover, changes over time are harder to measure. As a result, attempting to find such differences is like searching for a smaller needle in a bigger haystack. In the remainder of the paper, we explore the implications for the design of test-based accountability systems.

IV. Implications for the Design of Incentive Systems

According to *Education Week*, 45 states were providing annual report cards on their schools' performance in January 2001 and 20 states were providing monetary rewards to teachers or schools based on their performance.⁹ However, the incentive systems have been designed with little recognition of the statistical properties of the measures upon which they are based. As we argue below, failure to take account of the volatility in test score measures can lead to weak incentives (or, in many cases, perverse incentives), while sending confusing signals to parents and to schools about which educational strategies are worth pursuing. We draw four lessons for the design of test-based incentive systems.

⁹Ulrich Boser, "Pressure Without Support" *Education Week*, Vol. XX, Number 17, January 11, 2001. See table on pp. 68-71.

Lesson 1: Incentives targeted at schools with test scores at either extreme-- rewards for those with very high scores or sanctions for those with very low scores-- primarily affect small schools and imply very weak incentives for large schools.

Each year since 1997, North Carolina has recognized the 25 elementary and middle schools in the state with the highest scores on the “growth composite”, a measure reflecting the average gain in performance among students enrolled at a school. Winning schools are honored at a statewide event in the fall, are given a banner to hang in their school and receive financial awards.

One indicator of the volatility of test scores is the rarity of repeat winners. Between 1997 and 2001, there were 101 awards for schools ranking in the top 25. (One year, two schools tied at the cut-off.) These 101 awards were won by 90 different schools, with only 9 schools winning twice and only 1 school winning three times. No school was in the top 25 in all 4 years.

We have analyzed data for 840 elementary schools in North Carolina for whom we had test score data for each year between 1994 and 1999. Of these schools, 59 were among the top 25 at some point between 1997 and 2000 (the top 25 each year included middle schools, which we are not analyzing here). Table 3 presents information on the mean gain scores in math in 4th and 5th grade, the variance in school mean gain scores and the probability of winning a top 25 award by school size decile. Several results in Table 3 are worth highlighting. First, the mean gain score is not strongly related to school size. Although the mean gain score over the period 1997 through 2000 among the smallest decile of schools was .032 student-level standard deviation units larger than the largest decile of schools (.021-(-.011)), that difference was not statistically significant. Second, although mean performance varied little with school size, the variance between schools was much larger for small schools. The variance in mean gain scores

among schools in the smallest size decile was nearly 5 times the variance among the largest decile of schools (.048/.011). Third, as a result of this variability, schools in the smallest decile were much more likely to be among the top 25 schools at some point over the period. 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, even though their mean gains were not statistically different, the smallest schools were 23 times more likely to win a "Top 25" award than the largest schools (.277/.012)!

But, for the same reason, small schools are also over-represented among those with extremely low test scores. Also beginning in 1997, the state assigned assistance teams to intervene in schools with the poorest performance on the state tests, who also did not meet growth targets from the previous year. Table 3 also reports the proportion of schools in each school size decile that were assigned an assistance team because of extremely low test scores in a given year. 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 less than 50 percent.)

The North Carolina accountability system provides other rewards that do not operate solely at the extremes. For example, 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. Therefore, we highlight the "Top 25" award not to characterize the North Carolina system as a whole, but to cite an example of the type of award program that is particularly susceptible to sampling variation.

This year, California is planning to spend a total of \$677 million on school and teacher bonuses. One component of the accountability system will provide bonuses of up to \$25,000 to teachers in schools with the largest improvements in test scores between 1999 and 2000. (The state is expecting to spend \$100 million on this component of the system alone.) Each school was given an overall target, based upon their 1999 scores. (Schools with lower 1999 scores faced higher targets for improvement.) In order to be eligible for the largest bonuses, a school had to have schoolwide scores below the median school in 1999, have no decline in test scores between 1998 and 1999 and have at least 100 students.¹⁰ Figure 5 plots the change in API scores by school size between 1999 and 2000 for those schools that met these requirements. Teachers in schools with the largest improvements will receive \$25,000 bonuses. These awards will be awarded to 1000 teachers. Then, 3750 teachers in schools with the next largest improvements in test scores will receive \$10,000 in bonuses. Finally, 7500 teachers will receive \$5,000 bonuses. It is clear from Figure 5 that the winners of the largest awards will all be smaller than the median sized school. Given the importance of sampling variation, this is hardly a surprise. Particularly when it comes to changes in test scores over time, all of the outlier schools will tend to be small schools.

Rewards or sanctions for extreme test scores or large changes in test scores have very little impact on large schools, because large schools have very little chance of ever achieving the extremes. Figure 7 illustrates the point with a hypothetical example. Suppose there were a small and a large school with the same expected performance. But because of sampling variation and

¹⁰Schools also had to meet targets for each “numerically significant” racial/ethnic group, an issue we will discuss further below.

other factors that can lead to temporarily changes in scores, each school faces a range of possible test scores next year, even if they do nothing. As portrayed in Figure 7, the range of potential test scores is likely to be wider for the small school than for the larger school. Suppose the state were to establish some threshold, above which a school won an award. If, as in Figure 7, the threshold is established far above both schools expected performance, there will be very little chance that the large school will win the award if they do nothing and a non-negligible chance that the small school will win the award if they do nothing. Since the probability of winning the award is represented by the area to the right of the threshold in the graph, the marginal effect of improving one's expected performance on the likelihood of winning the award is measured by the height of the curve as it crosses the threshold. In the hypothetical example portrayed in Figure 7, the marginal incentive is essentially zero for the large school and only slightly larger for the small school. (Note that the opposite would be true if the threshold were established close to both school's expected performance and that large schools would have a stronger incentive.)

A single threshold at either extreme is likely to be irrelevant for schools that are large, since the marginal effect of improving their performance on the likelihood of winning will be quite small. If the marginal costs of improving are also higher at large schools, the problem of weak incentives for large schools would only be compounded. While we do not observe the marginal costs of improving, it seems plausible that the costs of coordinating the efforts of a larger number of teachers to implement a new curriculum would be larger.

A remedy would be to establish different thresholds for different sized schools, such that the marginal net payoff to improving is similar for small and large schools, or offer different payoffs to small and large schools. For example, grouping schools according to size (as is done

in high school sports) and giving awards to the top 5% in each size class tend to even out the incentives (and disparities) between large and small schools. An alternative solution would be to establish thresholds closer to the middle of the test score distribution, where the differential in marginal payoffs is less extreme.

Ladd and Clotfelter (1996) and Grissmer et.al. (2000) report evidence suggesting that schools respond to incentives by raising student performance.¹¹ However, the long-term impacts of incentives may be quite different from the short-term impacts. Even if a school teacher is not sufficiently aware of the forces at work in an incentive system to draw a graph similar to that in Figure 7, they may well infer the magnitudes of the marginal incentives from their own experience over time. If their best efforts are rewarded with failure one year and less work the following year is rewarded with success, they are likely to form their own estimates of the value of their effort. Even if they do not fully recognize the statistical structure underlying their experience, teachers and principals are likely to learn over time about the impact of their efforts on their chances of winning an award. As a result, the long-term impacts on schools could be different from the short-term impacts.

¹¹ Nevertheless, whether the improvement in performance is real or the result of “teaching to the test” is unclear (see Koretz (2000)). Kane and Staiger (2001) report that the schools in North Carolina that showed the greatest improvement on 5th grade math and reading gains did *not* improve more on measures of student engagement (such as student absences, the proportion of students reporting less than hour of homework or the proportion of students watching less than 6 hours of television), even though these characteristics were related to gain scores in the base year.

Lesson 2. Incentive systems establishing separate thresholds for each racial/ethnic subgroup present a disadvantage to racially integrated schools. In fact, they can generate perverse incentives for districts to segregate their students.

The accountability system in a number of states, including Texas and California, establish separate growth expectations for racial or ethnic subgroups. The presumed purpose of such rules is to maintain schools' incentive to raise the performance of all youth, and to raise the cost to teachers and administrators of limiting their efforts to only one racial group. However, because the number of students in any particular racial group can be quite small, scores for these students are often volatile. For a racially integrated school, winning an award is analogous to correctly calling three or four coin tosses in a row, rather than a single toss.¹² As a result, at any given level of overall improvement, a racially integrated school is much less likely to win an award than a racially homogeneous school.

In California, in order 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 8 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.¹³

Table 4 reports the proportion of California elementary school's winning their Governor's

¹²However, because the threshold is higher, winning an award is probably a more accurate measure of true improvement for racially heterogeneous schools than for homogeneous schools.

¹³A socioeconomically disadvantaged student is 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.

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: 47 percent of schools with 4 or more racial/ethnic/socioeconomic subgroups won a Governor's Performance Award as opposed to 82 percent of similarly sized schools with only one numerically significant group. This is particularly ironic given that the more integrated schools had slightly larger overall growth in performance between 1999 and 2000 (36.0 API points versus 33.4 points). Moreover, although the results are not reported in Table 4, due to space limitations, such school actually witnessed larger gains on average for African American and Latino students than for white students.

Because any numerically significant subgroups will be larger in size (and, as a result, their scores less volatile), the gap between homogeneous and heterogeneous schools is slightly smaller among larger schools. Among schools in the largest size quintile, homogenous schools were 28 percent more likely to win a governor's performance award (.876/.686), even though, the more heterogeneous schools had greater improvements in overall test scores (40.5 API points as opposed to 29.5).

Table 4 has at least three important implications: First, under such rules, a district would have a strong incentive to segregate by race/ethnicity. For instance, suppose there were 4 small schools in a district, each being 25 percent African American, 25 percent Latino, 25 percent Asian American and 25 percent white, non-Hispanic. According to the results in Table 4, a district could nearly double each school's chance of winning an award simply 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 rules have the ironic effect of putting the average school enrolling minority students at a disadvantage in the pursuit of award money. For instance, in Table 4, the addition of each racial/ethnic subgroup lowers a school's chance of winning an award by roughly 9 percentage points on average. The average number of subgroups in the schools attended by African American student was 2.8; the average number of subgroups in the schools attended by white non-Hispanic students was 2.2. If each school had an equal chance of winning an award, the average school attended by an African American youth would have a 74.9 percent probability of winning an award. Therefore, a rough estimate would suggest that the measure has the effect of taking 7 percent of the money that would otherwise have gone to schools attended by African American youth and handing it to schools enrolling white, non-Hispanic youth $((2.8-2.2)*(.09/.749)=.072)$.¹⁴

Third, although the program reduces any temptation to focus on one group to the exclusion of others, it lowers the marginal payoff to improvement for all groups in an integrated school. For instance, suppose a homogenous school raised its probability of winning an award by a given amount if it improved its true performance by 10 points. A similarly sized school with more than one racial/ethnic group would have to improve its performance by more than 10 points to achieve the same impact on their likelihood of winning an award. When an award is

¹⁴The number of subgroups in the average school attended by Latino students was 2.5, African American students 2.8, Asian students 2.7, American Indian students 2.5, Pacific Islanders 2.7, Filipino students 2.8, socially disadvantaged students 2.6 and white, non-Hispanic students 2.2. The rough estimate of the impact of the rule on racial/ethnic differences in spending is not precisely correct, since the marginal impact of the number of subgroups on a school's chances of winning an award depends upon the number of subgroups and the size of the school.

contingent on an increase for each and every one of a school's racial subgroups, there is a greater chance that true progress will be spoiled by a chance decline for one group in a given year.

Although the costs of the subgroup targets are clear, the benefits are uncertain.

Policymakers might want to know whether the rules actually do force schools to focus more on the achievement of minority youth. If so, some consideration of the test scores of racial/ethnic subgroups may be worthwhile, despite the costs described above. One way to estimate this impact would be to compare the improvements for minority youth in schools where they just above and just below the minimum percentage required to qualify as a separate subgroup. We have done so with data from Texas. The trend in test scores for African American and Latino youth in schools where they were insufficiently numerous to qualify as a separate subgroup (in Texas, between 5 and 10 percent) was identical to the trend for African American and Latino youth in schools where their percentage of enrollment was high enough to qualify for a separate standard.¹⁵ In other words, despite the costs, there is no evidence that such thresholds actually force schools to focus on the performance of disadvantaged minority youth.

Lesson 3: As a tool for identifying best practice or fastest improvement, annual test scores are generally quite unreliable. There are more efficient ways to pool information across schools and across years to identify those schools that are worth emulating.

When designing incentive systems to encourage schools to “do the right thing”, one cares about the absolute amount of imprecision in school test score measures and how that imprecision may vary by school size. The more imprecise the measures are, the weaker the incentives tend

¹⁵See Kane, Staiger and Geppert (2001).

to be. However, there is also a lot of uncertainty (or, at least, disagreement) about what the “right thing” is, among policymakers and school administrators. The state may also have an interest in helping to identify the schools that are worth emulating.¹⁶ If the goal of an accountability system is not only to provide incentives, but to help identify success, it is not only the absolute amount of imprecision that matters, but the amount of imprecision relative to the degree of underlying differences that determines the likelihood of success in the search for exemplars.¹⁷

Building upon work by McClellan and Staiger (1999) in rating hospital performance, we have proposed a simple technique for estimating the amount of signal and noise in school test score measures and to use that information to generate “filtered” estimates of school quality that provide much better information about a school’s actual performance. (See Kane and Staiger (2001)¹⁸.) Many of the key ideas behind that filtering technique are illustrated through a simple example. Suppose that a school administrator is attempting to evaluate a particular school’s performance based on the mean test scores of the students from that school in the most recent two years. Consider the following three possible approaches: (1) use only the most recent score for a school, (2) construct a simple average of the school’s scores from the two recent years, and

¹⁶Presumably, that is the point of identifying the “Top 25” schools in North Carolina and giving them a banner to identify that fact.

¹⁷Rogosa (1995) makes a similar point.

¹⁸The estimator is an empirical Bayes estimator in the spirit of Morris (1983). However, the method employed for estimating the variance components is less computationally intensive than that proposed in Bryk and Raudenbush (1992) and can incorporate information on multiple outcomes and multiple years. Moreover, the filtering technique based upon these estimates is linear, offering additional computational advantages.

(3) ignore the school's scores and assume that student performance in the school is equal to the state average. In order to minimize mistakes, the best choice among these three approaches depends on two important considerations: the signal-to-noise ratio in the school's data, and the correlation in performance across years. For example, if the average test scores for the school were based on only a few dozen students, and one had reason to believe that school performance did not vary much across the state, then one would be tempted to choose the last option and place less weight on the school's scores because of their low signal-to-noise ratio and heavily weight the state average. Alternatively, if one had reason to believe that school performance changed very slowly over time, one might choose the second option in hopes that averaging the data over two years would reduce the noise in the estimates by effectively increasing the sample size in the school. Even with large samples of students being tested one might want to average over years if idiosyncratic factors such as the weather on the day of the test affected scores from any single year. Finally, one would tend to choose the first option, and rely solely on scores from the most recent year, if such idiosyncratic factors were unimportant, if the school's estimate was based on a very large sample of students, and if there seems to be a lot of persistent change over time.

Our method of creating filtered estimates formalizes the intuition from this simple example. The filtered estimates are a combination of the school's own test score, the state average, and the school's test scores from past years, other grades, or other subjects. Table 5 compares the mean performance in 1999 for NC elementary schools ranking in the top 10 percent in 5th grade math gains on two different measures: the simple means of math gains in 1997 and the filtered prediction that would have been made of a school's performance in 1999 using all of the data available through 1997. Thus, both predictions use only the data from 1997 or before.

However, the filtered prediction incorporates information from reading scores and from prior years, and “reins in” the prediction according to the amount of sampling variation and non-persistent fluctuation in the data.

Table 5 reports the mean 1999 performance, cross-tabulated by whether or not the school was in the top 10 percent using the filtering technique and using the naive estimate based upon the actual 1997 scores. Sixty-five (65) schools were identified as being in the top 10 percent as of 1997 using both the naive and the filtered predictions, and these schools scored .15 student level standard deviations higher than the mean school two years later in 1999. However, among the schools where the two methods disagreed, there were large differences in performance. For instance, among the 25 schools that the filtering method identified as being in the top 10 percent that were not in the top 10 percent on the 1997 actual scores, the average performance on 5th grade math gains was .124 student-level standard deviations above the average in 1999. On the other hand, among the 25 schools chosen using actual 1997 scores who were not chosen using the filtering technique, scores were .022 standard deviations *lower* than the average school in 1999. The next to last column and row in Table 5 reports the difference in mean scores moving across the first two columns or first two rows. Among those who were not identified as being in the top 10 percent by the filtering method, knowing that they were in the top 10 percent on the actual 1997 score provided very little information regarding test scores. In fact the test scores were -.006 standard deviations lower on average holding the filtered prediction constant. In contrast, among those were not identified as being in the top 10 percent on actual 1997 scores, knowing that they were selected using the filtering method was associated with a .140 standard deviation difference in performance. Apparently, the filtering method was much more successful

in picking schools that were likely to perform well in 1999.

Moreover, the filtering technique provides a much more realistic expectation of the magnitude of the performance differences to expect. As reported in the last column of Table 5, the schools in the top 10 percent on the actual test in 1997 scored .453 standard deviations higher than the average school in 1997. If we had naively expected them to continue that performance, we would have been quite disappointed, since the actual difference in performance was only .115 standard deviations. On the other hand, among those who were chosen using the filtering method, we would have predicted that they would have scored .180 standard deviations higher than the average school in 1999 based upon their performance prior to 1998. The actual difference in performance for these schools was .160 standard deviations.

Table 6 compares the R^2 one would have obtained using 3 different methods to predict the 1998 and 1999 test scores of schools using only the information available prior to 1998. The first method is the “filtering method” described above. The second method is using the actual 1997 score as the prediction for the 1998 and 1999 scores. The third method uses the 4-year average of math performance prior to 1998 (1994-1997) to predict 1998 and 1999.

Whether one is trying to anticipate math or reading, levels or gains in 5th grade, the filtering method leads to greater accuracy in prediction. The R^2 in predicting 5th grade math levels was .41 using the filtering method, .19 using the 1997 score and .29 using the 1994-97 average. The filtering method also calculates a weighted average using the 1994-97 scores, but it adjusts the weights according to sample size (attaching a larger weight to more recent scores for large schools) and uses both the math and reading score histories in predicting either. In so doing, it does much better than a simple average of test scores over 1994-97.

In predicting math or reading gain scores in 1998, the second column reports *negative* R^2 when using the 1997 scores alone. A negative R^2 implies that one would have had less squared error in prediction by completely ignoring the individual scores from 1997 score and simply predicting that performance in every school would be equal to the state average. Of course, one could probably do even better by not ignoring the 1997 score, but simply applying a coefficient of less than 1 to the 1997 score in predicting future scores. That is essentially what the filtering method does, while recognizing that the optimal coefficient on the 1997 score (and even earlier scores) will depend upon the amount of non-persistent noise in the indicator as well as the school size.

Although it performs better than either the 1997 score or the 1994-97 average in predicting 1998 and 1999 gains, the R^2 using the filtering method is only .16 on math gains and .04 on reading gains. This hardly seems to be cause for much celebration, until one realizes that even if the filtering method were completely accurate in predicting the persistent portion of school test scores, the R^2 would be less than 1 simply because a large share of the variation in school performance is due to sampling variation or other non-persistent types of variation. Because of these entirely unpredictable types of error, the highest R^2 one could have hoped for would have been .75 in predicting math levels, .60 in predicting reading levels, .55 for math gains and .35 in reading gains. For math gains, for instance, the filtering method was able to predict 16 percentage points of the 55 percentage points that one ever had a hope of predicting, implying an R^2 for the systematic portion of school test scores of $.16/.55=.29$.

One obvious disadvantage of the filtering technique is that it is much less transparent.¹⁹

¹⁹See Ladd and Clotfelter (1996) for a discussion of the merits of transparency.

The average parent, teacher or school principal is likely to be familiar with the idea of computing an arithmetic mean of test scores in a school; the average parent or principal is certainly unlikely to be familiar with empirical Bayes techniques. However, it is important to remember that there are a number of mysterious calculations involved in creating a scale for test scores that are currently well tolerated. To start, parents are likely to have only a very loose understanding of the specific items on the test. (Admittedly, teachers and principals are probably better informed about test content.) Moreover, any given student's test score is generally not a "percent correct", but a weighted average of the individual items on the test. Parents and all but a few teachers are unfamiliar with the methods used to calculate these weights. The filtering technique we are proposing could be used to provide an index of school performance, and beyond an intuitive description of the techniques involved, it might well be as well-tolerated as the scaling process is already.

Lesson 4: When evaluating the impact of policies on changes in test scores over time, one must take into account the fluctuations in test scores that are likely to occur naturally.

In 1997, North Carolina identified 15 elementary and middle schools with poor performance in both levels and gains and assigned "assistance teams" of 3-5 educators to work in these schools. The next year, all of the schools had improved enough to escape being designated as a "low-performing" school. In summarizing the results of that first year, the state Department of Public Instruction claimed an important victory:

"Last year, the assistance teams of 3-5 educators each worked in 15 schools, helping staff to align the instructional program with the Standard Course of Study, modeling and

demonstrating effective instructional practices, coaching and mentoring teachers and locating additional resources for the schools. As a result of this assistance and extra help provided by local school systems, nearly all of these schools made exemplary growth this year and none are identified as low performing.” (emphasis added)

NC Department of Public Instruction, “ABCs Results Show Strong Growth in Student Achievement K-8”, August 6, 1998

Indeed, the value of the assistance teams was lauded in *Education Week*'s annual summary of the progress of school reform efforts in the states.²⁰ However, given the amount of sampling variation and other non-persistent fluctuations in test score levels and gains, schools with particularly low test scores in one year would be expected to bounce back in subsequent years.

We had test score data from 1994 through 1999 for 35 elementary schools that won a “Top 25” school award in either 1997 or 1998 as well as for 10 elementary schools that were assigned an assistance team in 1997 or 1998. Table 7 reports fourth grade test scores the year before, the year after and the year that each school either won the award or sanction. (For those assigned assistance teams, the help did not arrive at the school until the year after their low scores merited the assignment.)

For the average school winning a “Top 25” award, the year of the award is clearly an aberrant year. The year of the award, their scores were .230 student-level standard deviations above the mean gain. However, both the year before their award and the year after, their gain scores were actually slightly below the mean gain.

Moreover, the schools that were assigned assistance teams seem to have had a particularly

²⁰ Kathleen Kennedy Manzo, “North Carolina: Seeing a Payoff” *Education Week* Vol. 18, No. 17, p. 165.

bad year the year of their receiving the sanction. In the year before assignment, such schools had an average 4th grade combined reading and math test score .668 student-level standard deviations below the average school. This reveals that they were weak schools the year before being sanctioned. However, in the year of assignment, their average score was even lower, .786 student-level standard deviations below the average school. The year after assignment, their scores seemed to rebound to .523 student standard deviations below the mean.

Because the year of assignment was a bad year and because change is volatile, one is likely to greatly overestimate the impact of assistance teams by taking the change in performance in the year after assignment. In Table 7, we estimate the impact of the assistance teams by taking the difference in scores in the year after assignment relative to the scores in the year before assignment. That estimate suggests that schools that were assigned assistance teams may well have improved their performance over time, by a fairly sizeable .145 student level standard deviations. (Their mean gain also improved by .156 student-level standard deviations.) Both such estimates would be considered statistically significant at the .059 and .006 levels. However, as reported in the last column of Table 7, such an estimate of the impact is between only 55 and 73 percent as large respectively as one would have seen using the year of assignment as the base year.

V. Conclusion

To date, school accountability systems have been designed with little recognition of the statistical properties of the measures upon which they are based. For instance, if there were little sampling variation and if changes in performance were largely persistent, one might well want to

focus on a school's mean value-added in the most recent year or on the changes in school-wide scores over the most recent two years. However, as we describe above, such reasoning ignores an important trade-off: changes in performance and mean value-added are very difficult to recognize and reward with only two years of test score data. An accountability systems that seems reasonable in a world of persistent rates of change and easy to discern differences in value-added, may generate weak or even perverse incentives when implemented in the world of volatile test scores in which we live.

The long-term effects on the morale and motivation of school personnel remain to be seen. Given the apparent role of chance in some of the incentive regimes being implemented, those effects could be quite different from the short-term impacts. In 1967, psychologists Martin Seligman and Steven Maier (1967) published the results of an experiment in which one group of dogs were strapped into a harness and administered a series of electrical shocks through electrodes attached to their feet. Needless to say, the dogs developed a strong aversion to such treatment. Later, the same dogs were transferred into a room in which they were administered similar shocks through the floor. The dogs merely had to jump over a shoulder-height barrier to escape from the shocks. However, rather than flee, the dogs simply laid down on the floor and accepted the shocks. Why the apparently self-destructive behavior? In addition to learning that they did not like being shocked, the first group of dogs apparently had learned that there was little they could do to avoid the shocks. (A second group of dogs, that were able to stop the shocks during the first stage of the experiment by tapping a paddle, did flee the shocks in the second stage by jumping over the barrier.) In states' efforts to encourage school personnel to focus on student performance, it is not sufficient to create desirable rewards or noxious sanctions

attached to student performance. We must also be cautious of the lessons teachers and principals are learning about their ability to determine those outcomes and be sure that they infer that their efforts will be rewarded.

However, the results of this paper should not be interpreted as implying that all accountability systems are necessarily flawed. Rather, in Section IV, we provided four simple principles for improving existing systems. First, rewards and bonuses should not be limited to schools with extreme scores. To preserve incentives for large schools, states should either establish separate thresholds for schools of different sizes or, slightly less effective, provide smaller rewards to schools closer to the middle of the test score distribution. Second, rules making any rewards contingent on improvement in each racial group present a great disadvantage to integrated schools, and generate a number of perverse incentives that may actually harm rather than help minority students. Third, when seeking to identify schools that are improving the most or to identify schools with the highest mean value-added, one can generate much more reliable estimates by pooling information across years and across outcomes. In Kane and Staiger (2001), we describe an estimator that does that in a more efficient way than simply taking a simple mean across as many years as possible. Finally, when evaluating the impact of policies that operate on schools at either extreme of the distribution, one has to recognize the importance of volatility and be careful about the choice of a baseline.

References:

- Bryk, Anthony and Stephen Raudenbush, *Hierarchical Linear Models* (Newbury Park, CA: Sage Publications, 1992).
- California Department of Education, Policy and Evaluation Division, “2000 Academic Performance Index Base Report Information Guide”, January 2001.
- Coleman, James S., Ernest Campbell, Carol Hobson, James McPartland, Alexander Mood, Frederic Weinfeld and Robert York, *Equality of Educational Opportunity* (Washington, DC: U.S. Department of Health Education and Welfare, 1966).
- Grissmer, David, Ann Flanagan, Jennifer Kawata and Stephanie Williamson, *Improving Student Achievement: What State NAEP Test Scores Tell Us* (Santa Monica, CA: Rand, 2000).
- Hoxby, Caroline “Peer Effects in the Classroom: Learning from Gender and Race Variation” National Bureau of Economic Research Working Paper No. 7867, August 2000.
- 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., Douglas O. Staiger and Jeffrey Geppert, “Assessing the Definition of ‘Adequate Yearly Progress’ in the House and Senate Education Bills” Unpublished paper, UCLA School of Public Policy and Social Research, July 2001.
- 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*.)
- Ladd, Helen F. and Charles Clotfelter, “Recognizing and Rewarding Success in Public Schools” in Helen F. Ladd (ed.), *Holding Schools Accountable* (Washington, DC: Brookings Institution, 1996).
- McClellan, Mark and Douglas Staiger, “The Quality of Health Care Providers” National Bureau of Economic Research Working Paper No. 7327, August 1999.
- Morris, Carl. “Parametric Empirical Bayes Inference: Theory and Applications” *Journal of the American Statistical Association*, (1983) Volume 381, No. 78, pp. 47-55.

Rogosa, David “Myths and Methods: ‘Myths about Longitudinal Research’ plus Supplemental Questions” in John Mordechai Gottman (ed.) *The Analysis of Change* (Mahwah, New Jersey: Lawrence Erlbaum Associates, 1995).

Seligman, Martin E. P. and Steven F. Maier, “Failure to Escape Traumatic Shock” *Journal of Experimental Psychology* (1967) Vol. 74, No. 1, pp. 1-9.

Tarcy, Brian “Town’s Scores the Most Improved” *Boston Globe*, December 8, 1999, p. C2.

Table 1.
Characteristics of the Matched and Non-Matched
Sample of 4th and 5th Grade Students in 1999

	Non-Matched	Matched
% of 4th and 5th Grade Students	34.2	65.8
Mean Math Score	153.8	156.5
S.D. in Math Score	11.1	10.5
Mean Reading Score	150.5	152.4
S.D. in Reading Score	9.5	9.1
Percent Female	47.4%	50.1%
Percent Black	35.1	27.7
Percent Hispanic	5.4	2.2
Parental Education:		
H.S. Dropout	16.6%	9.8%
H.S. Graduate	47.1	43.7
Trade/Business School	4.6	5.3
Community College	11.3	14.2
Four-Year College	16.5	21.9
Graduate School	3.9	5.1
Sample Size	69,388	133,305

Note: Each of the differences above were statistically significant at the .05 level.

**Table 2. Decomposing Variance in School Test Scores
Due to Sampling Variation and Other Non-Persistent Factors**

Combined Reading and Math Scores in 4th Grade:

School Size:	Average Size:	Total Variance:	Sampling Variance:	Other Non-Persistent Variance:	Total Proportion Non-Persistent
Smallest Quintile	28	0.156	0.028	0.003	0.198
Middle Quintile	56	0.137	0.015	0.005	0.144
Largest Quintile	104	0.110	0.008	0.002	0.092

Combined Reading and Math Gains between 3rd and 4th Grade:

Smallest Quintile	28	0.053	0.008	0.022	0.575
Middle Quintile	56	0.031	0.004	0.011	0.486
Largest Quintile	104	0.019	0.002	0.003	0.286

Annual Change in 4th Grade Combined Reading and Math Scores:

Smallest Quintile	28	0.078	0.056	0.005	0.793
Middle Quintile	56	0.055	0.030	0.009	0.728
Largest Quintile	104	0.027	0.017	0.003	0.733

Note: All variances are expressed in units of student-level variances for 4th grade scores. Sampling variance was calculated by dividing the average within-school variance (calculated separately for each school size quintile) by the sample size. The variance due to other non-persistent factors was calculated as $-\rho_{\Delta_t, \Delta_{t-1}} \sigma_{\Delta}^2 - \sigma_{Samp}^2$ for each quintile where $\rho_{\Delta_t, \Delta_{t-1}}$ is the correlation in adjacent year changes for that quintile, σ_{Δ}^2 is the variance in the change for that quintile and σ_{Samp}^2 is the estimated sampling variance for that quintile. Sampling variance and other non-persistent variance for changes in test score levels was estimated by doubling the variances in the top panel. See Kane and Staiger (2001) for an alternative estimator and standard errors.

Table 3.
Awards and Sanctions Among Elementary Schools in North Carolina

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

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

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

Table 5.
Performance of North Carolina Schools in 1999
Identified as Being in the “Top 10%” in 1997
Based on Actual and Filtered Test Scores
5th Grade Math Gains

		<u>Based on actual 1997 Score</u>				
		School not in Top 10%	School is in Top 10%	Row Total	Difference between Top 10% and the rest	Expected difference
<u>Based on filtered prediction of 1999 Score (from 1997)</u>	School not in Top 10%	-0.016 (0.007) [N=779]	-0.022 (0.066) [N=25]	-0.016 (0.007) [N=804]	-0.006 (0.043)	0.385 (0.034)
	School is in Top 10%	0.124 (0.050) [N=25]	0.151 (0.026) [N=65]	0.144 (0.023) [N=90]	0.027 (0.052)	0.236 (0.036)
	Column Total	-0.012 (0.007) [N=804]	0.103 (0.027) [N=90]	0 (0) [N=894]	0.115 (0.024)	0.453 (0.019)
	Difference between top 10% and the rest	0.140 (0.042)	0.173 (0.059)	0.160 (0.023)		
	Expected difference	0.147 (0.013)	0.095 (0.012)	0.180 (0.007)		

Notes: Within the box, the entries report the mean of the 5th grade math gain score in 1999, along with standard errors of these estimates and the sample size in each cell. The columns of the table use actual scores in 1997 to assign schools to “top 10%” and to calculate the expected difference between the top 10% and the rest. The rows of the table use filtered predictions of 1999 scores, based only on data from 1994-1997, to assign schools to “top 10%”.

Table 6.
Comparing the Accuracy of Alternative Forecasts
of 1998 and 1999 Test Scores using North Carolina Data

Test Score Being Predicted:	Unweighted R ² when Forecasting 1998 and 1999 Scores under Alternative Uses of 1993-97 Data					
	Predicting Scores in 1998 (1-year ahead forecast R ²)			Predicting Scores in 1999 (2-year ahead forecast R ²)		
	“Filtered” Prediction	1997 Score	Average Score 1994-97	“Filtered” Prediction	1997 Score	Average Score 1994-97
Adjusted Score						
5 th Grade Math	0.41	0.19	0.29	0.27	-0.02	0.13
5 th Grade Reading	0.39	0.13	0.33	0.31	-0.05	0.24
Gain Score						
5 th Grade Math	0.16	-0.27	0.09	0.12	-0.42	-0.01
5 th Grade Reading	0.04	-0.93	-0.12	0.04	-0.85	-0.20

Note: The “filtered” prediction is an out-of-sample prediction, generated using only the 1993-1997 data.

Table 7.
Fourth Grade Test Scores Before and After
Sanction or Reward in North Carolina

	Year Before Award/ Sanction	Year of Award/ Sanction	Year After Award/ Sanction	Year After- Year Before	Ratio of $S_{t+1}-S_{t-1}/$ $S_{t+1}-S_t$
Top 25 in 1997-1998				Diff. (p-value)	
Math+Reading Gain Score	-.003	.230	-.064	-.062 (.164)	.211
Assist Team in 1997-1998					
Math+Reading Test Score	-.668	-.786	-.523	.145 (.059)	.551
Math+Reading Gain Score	-.078	-.134	.078	.156 (.006)	.735

Note: Test scores are in units of student-level standard deviations. In this table, the mean test scores across all schools in each year has been subtracted. If a single school won an award more than once, we used their first award. There were 35 elementary schools in our sample that won a Top 25 award in 1997 or 1998 based upon their gain scores. There were 10 elementary schools in our sample that were assigned an assistance team in 1997 or 1998 based upon a combination of low test scores and low gain scores.

Figure 1:

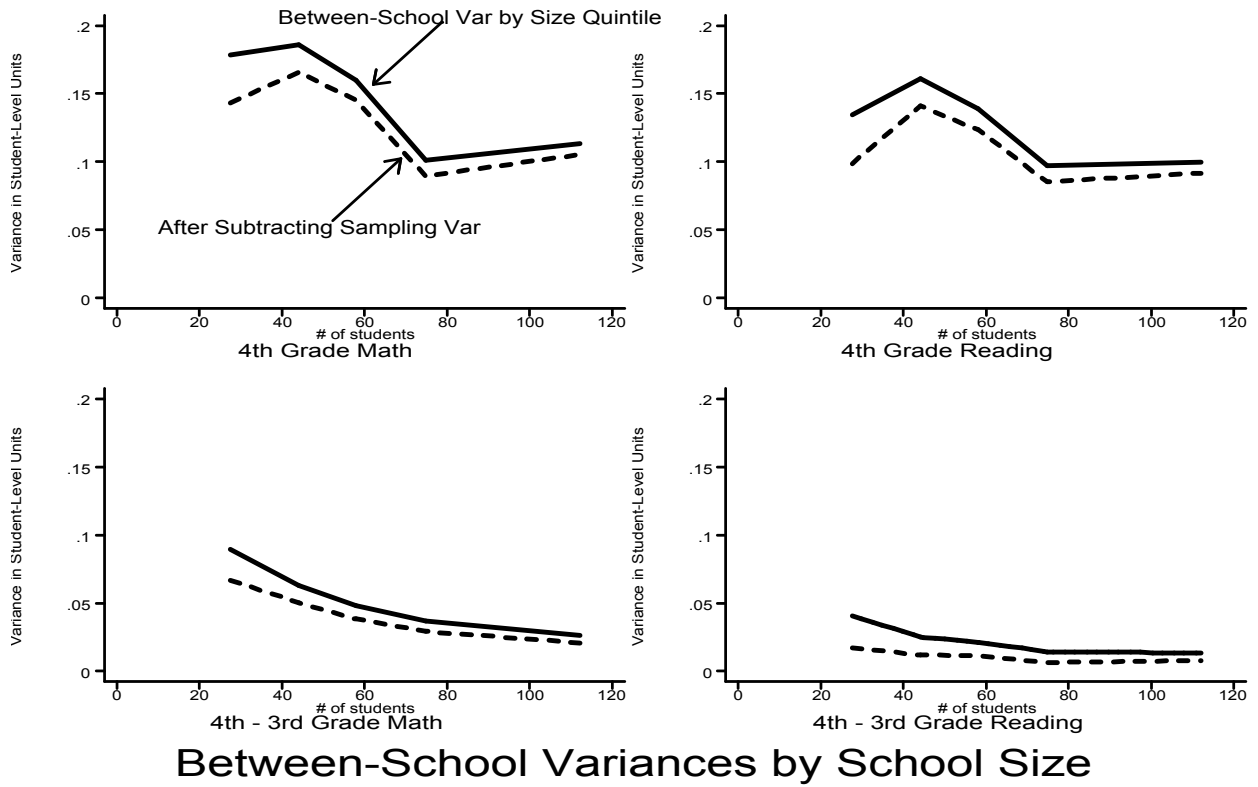


Figure 2:

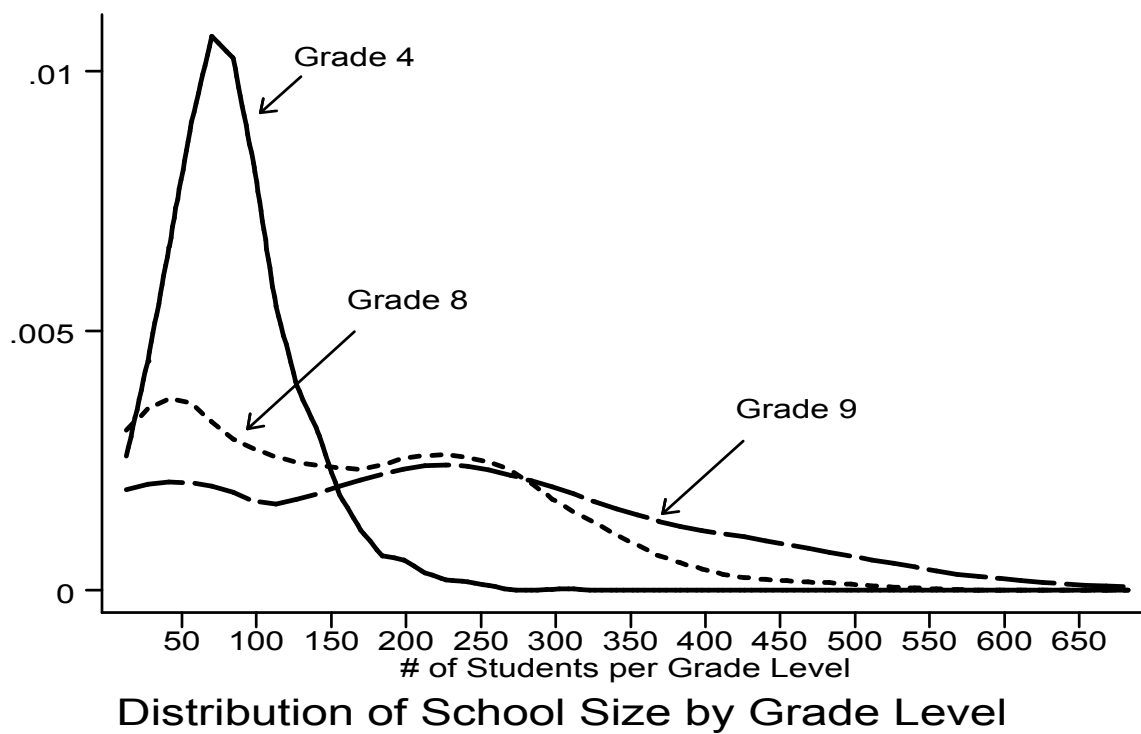
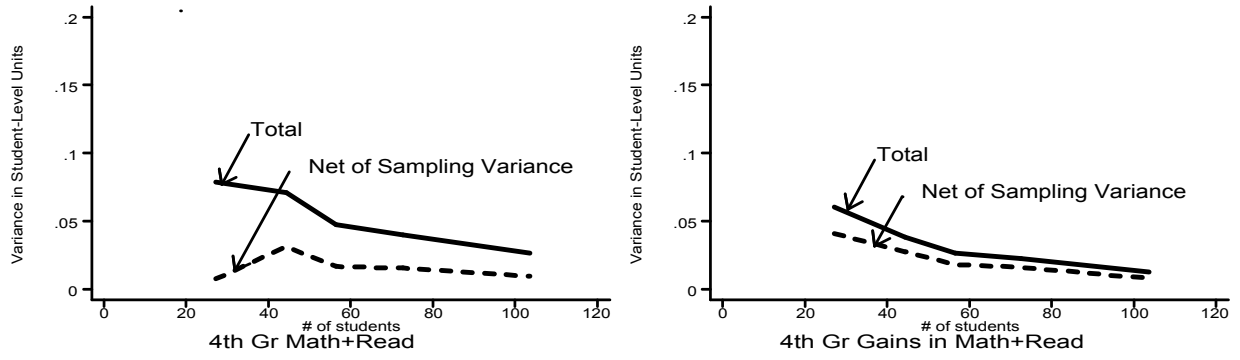


Figure 3.



Between-School Variance in Annual Change

Figure 4. Correlation in the Change in Scores in Consecutive Years by Size of School in North Carolina and California

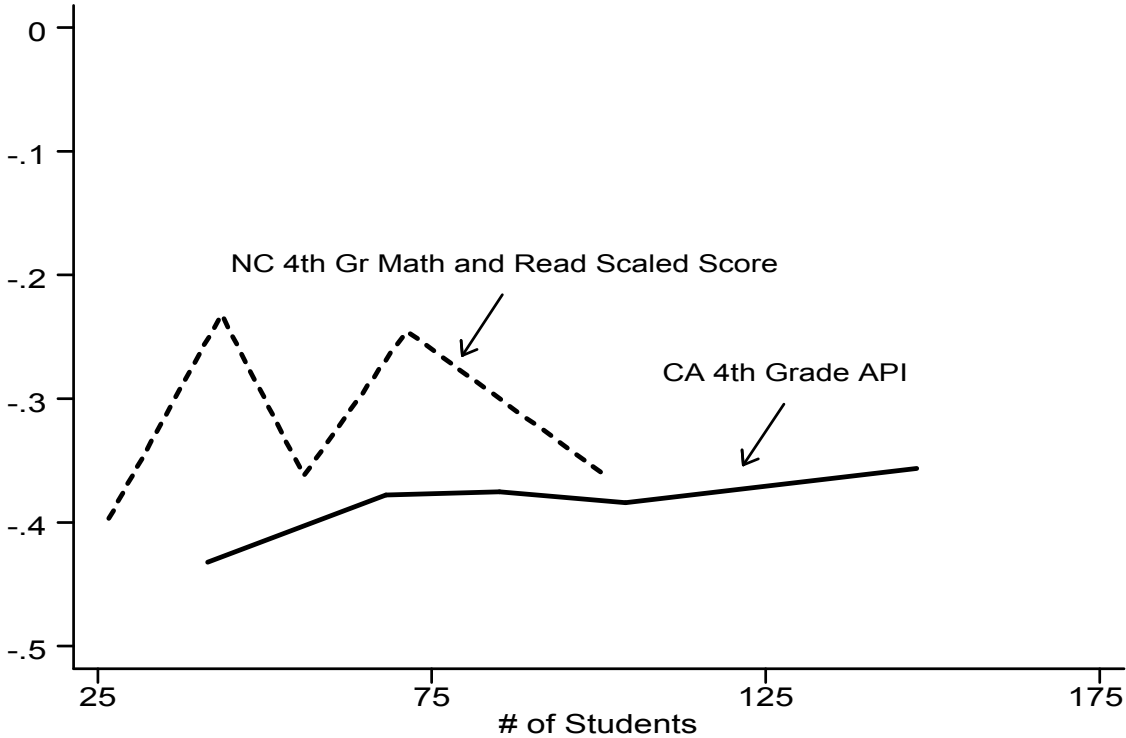


Figure 5.
Correlation in Change in Scores in Consecutive Years
with Overlapping and Non-Overlapping Cohorts

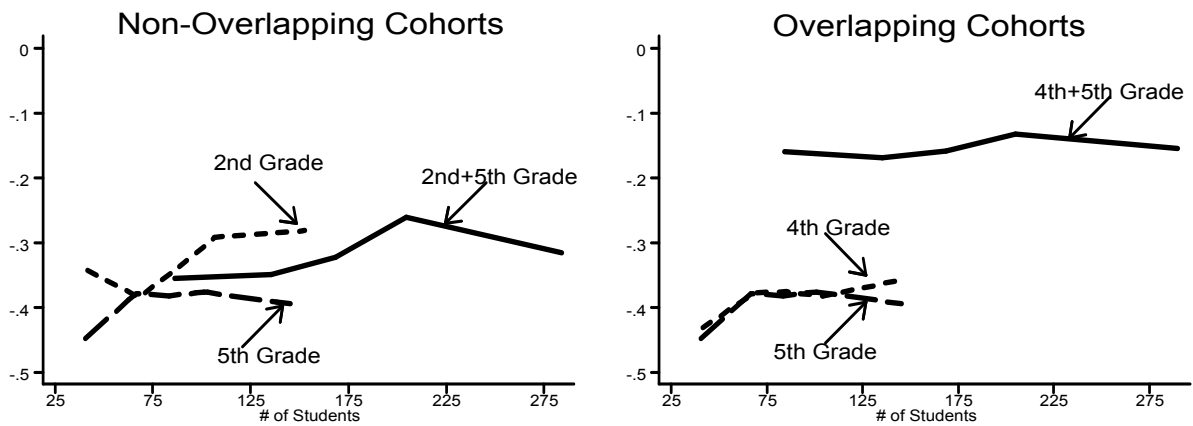
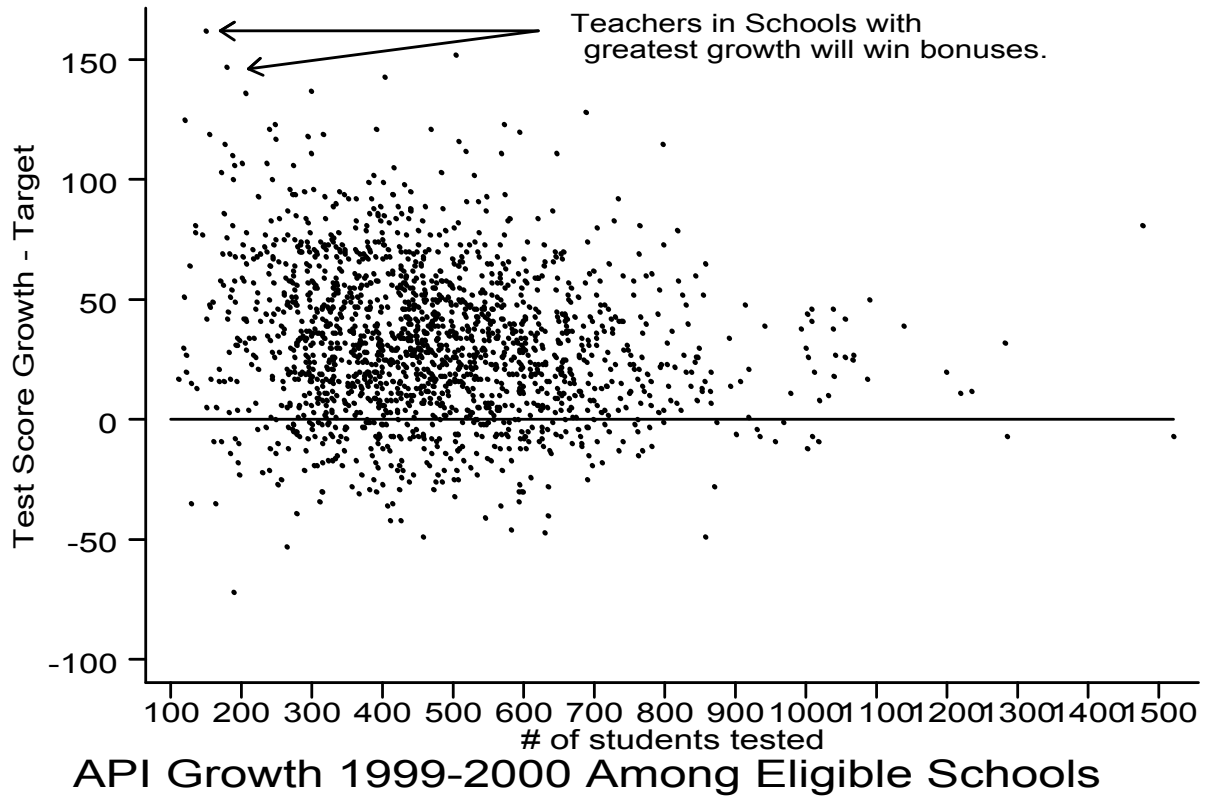


Figure 6.
Improvements in Test Scores among California Schools
Eligible to Win Teacher Bonuses by School Size



Note: Reflecting program rules, the above figure has been drawn only for those elementary schools in the bottom 5 deciles of API scores in 1999, with non-negative changes in test scores between 1998 and 1999, with at least 100 students tested.

Figure 7.
The Precision of Test Score Measures and
Incentive Effects

