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

THOMAS J. KANE *and*
DOUGLAS O. STAIGER

By the spring of 2000, forty states had begun using student test scores to rate school performance. Twenty states have gone a step further and are attaching explicit monetary rewards or sanctions to a school's test performance. For example, California planned to spend $677 million on teacher incentives in 2001, providing bonuses of up to $25,000 to teachers in schools with the largest test score gains. We highlight an underappreciated 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 sixty-eight students per grade level, the amount of variation stemming from the idiosyncrasies of the particular sample of students being tested is often large relative to the total amount of variation observed between schools. The second arises from one-time factors that are not sensitive to the size of the sample; for example, a dog barking in the playground on the day of the test, a severe flu season, a 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, because one would expect any idiosyncrasies in individual students' scores to average out. Although the averaging of students' scores does help lessen 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.[1] We estimate that the confidence interval for the average fourth-grade reading or math score in a school with sixty-eight 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. 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 Massachusetts Comprehensive Assessment System test scores were released 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.[2] 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 twenty-six students taking the test in tenth grade. Given the wide distribution of test scores among students in Massachusetts, any grouping of twenty-six students is likely to yield dramatic swings in test scores from year to year—that is, large relative to the distribution of between-school 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 how to proceed.

Sources of Data

We obtained math and reading test scores for nearly 300,000 students in grades three through five, attending elementary schools in North Carolina between the 1992–93 and 1998–99 school years. (The data were obtained from the North Carolina 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 date of birth, race, and gender.[3] 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 one other student in their school and grade, and 2 percent shared their birth date with two other people. (Less than 1 percent shared their birth date and gender with three 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 more than one person had the same school, birth date, race, and gender, we looked to see whether any unique matches could be made on parental education. If more than one person 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 fourth- and fifth-grade students in 1999. (The matching rate was similar in other years.) The data in table 1 compare the characteristics of the matched and the nonmatched 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 lower proportion black and Hispanic, and a slightly lower average parental education than the sample for which no match could be found.

We mostly 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 the 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

Table 1. Characteristics of the Matched and Nonmatched Sample of Fourth- and Fifth-Grade Students in 1999

Characteristic of sample	*Nonmatched*	*Matched*
Fourth- and fifth-grade students	34.2%	65.8%
Mean math score	153.8	156.5
Standard deviation in math score	11.1	10.5
Mean reading score	150.5	152.4
Standard deviation in reading score	9.5	9.1
Female	47.4%	50.1%
Black	35.1	27.7
Hispanic	5.4	2.2
Parental education		
High school dropout	16.6%	9.8%
High school graduate	47.1	43.7
Trade or 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 was statistically significant at the .05 level.

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 as above) 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 regressing a student's score on his or her score in the previous year, and then taking the residuals as a measure of student improvements. However, because the results were similar, we are reporting only the results using gain scores and test score levels.

In a previous work, we also adjusted each individual student's score for race, gender, and parental education.[4] That has the effect of removing between-school differences due to differences in race and parental education. We use the unadjusted test score data here. (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 (API) scores in 1998 through 2000. The Academic Performance Index is based upon school-level scores on the Stanford 9 tests. Schools receive 1,000 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 a weighted average of their scores in the reading, spelling, language, and mathematics portions of the Stanford 9 tests.[5] We use the 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.

Sources of Volatility in School-Level Test Scores

Three characteristics of school-level test score measures are vital to the design of test-based accountability systems. First, a considerable amount of variation in test scores exists 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 only sixty-eight students per grade level are in the typical elementary school, such variation can be substantial.

Second, other factors produce nonpersistent 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. We cannot estimate the magnitude of this source of variation directly without explicitly monitoring each influence on scores. However, 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 nonpersistent factors. Any nonpersistent variation in test scores that is not due to sampling variation we put into this category.

Third, 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 unreliable measures. Schools differ 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 nonpersistent—either because of 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-size school, roughly half of the variation between schools in gain scores (or value-added) for any given grade is also nonpersistent.

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, nearly one thousand schools in North Carolina had 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 fourth-grade reading scores to the total variance in reading scores was .89. That is, 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. It reflects a long-standing finding in educational assessment research. In their classic study of inequality of student achievement published in 1966, James S. Coleman and his colleagues estimated that only between 12 and 16 percent of the variance in verbal achievement among white third-grade students was due to differences across schools. The remainder was attributable to differences within schools.[6] 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 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, schools serving the elementary grades had sixty-eight students per grade level on average.[7] Dividing .87 and .89, respectively, by 68, we would expect a variance of .013, simply from the effect of drawing a new sample of students.

In North Carolina elementary schools near the national average in size (between sixty-five and seventy-five 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 fourth-grade math and reading test scores was due to sampling variation.

Gaining a strong intuitive sense for the magnitude of sampling variation with a proportion of variance calculation is sometimes difficult. 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 sixty-five and seventy-five 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 and Tennessee—rates its schools by focusing on the average gain in performance among students attending a particular school.[8] Advocates tout the value-added methodology as a fairer 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 third grade and the end of fourth 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 fourth-grade math and reading test scores). The variance in gains is smaller than the variance in test scores in fourth grade (or third grade), even though one imperfect measure of a child's performance is subtracted from another. 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 no relationship exists 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.

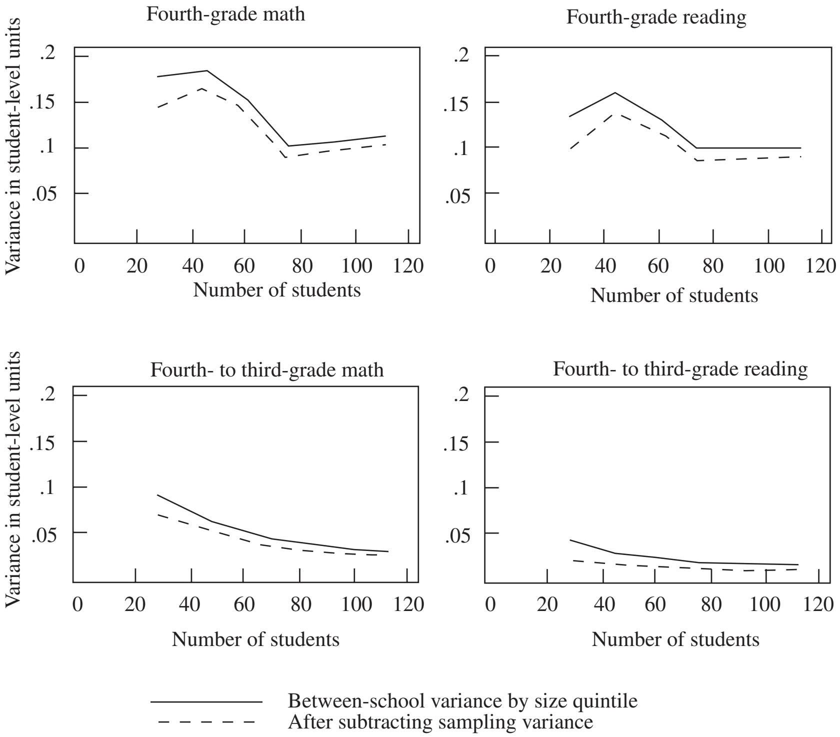
To calculate the variance in test scores we would expect to result from sampling variation for a school of average size, we would simply divide the within-school variance in gain scores (.331 and .343) by the sample size (68), yielding an estimate of .0049 for math and .0050 for reading. However, while the within-school variance in gains between third and fourth grades is four-tenths as large as the within-school variance in test scores in fourth 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 sixty-five to seventy-five students, the variance in reading scores was .015. Put another way, 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 fourth-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.[9]

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 five groups by school size. We then calculated the variance between schools in each quintile in test scores. We did so for fourth-grade math and reading and for gains in scores between third and fourth grade in math and reading.

Several facts are evident in figure 1. First, for each measure, we observed much more variance in test scores among smaller schools than among larger schools. For math and reading test scores, the variance between schools was roughly 50 percent larger for the smallest quintile of schools than for the largest quintile. For math and reading gain scores, the between-school vari-

Figure 1. Between-School Variances by School Size



ance 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 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.

We ignored any peer effects in our estimate of the sampling variance. 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, Caroline Hoxby identifies substantial negative impacts on student performance from having a disproportionate share of boys in one's cohort.[10] Any time that a characteristic of the sample has a direct effect on the performance of each individual in that sample, our estimates of the magnitude of variance due to sampling variation are likely to be understated.

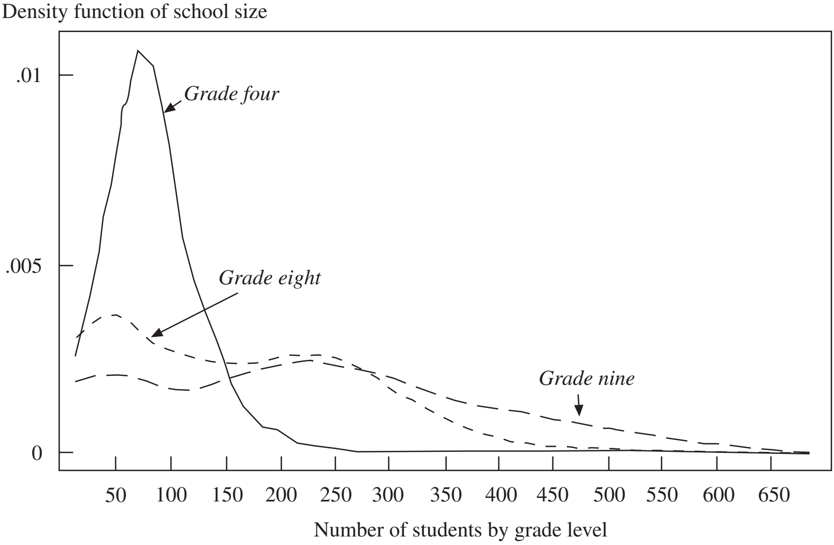
Third, very little variance existed between schools in the mean gain in reading scores between third and fourth grade. Even for the smallest quintile of schools, the between-school variance in the mean gain in reading performance was equal to .05 student-level standard deviations in fourth-grade reading 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 four. However, much more uniformity in school size is evident among elementary schools. While the size of the average middle school is larger than the size of the average elementary school, more heterogeneity is found in school size among middle schools. The same phenomenon is exaggerated at grade nine. High schools are generally much larger than elementary schools, but a number of small schools are enrolling ninth-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 the result of long-term differences among these schools (such as would occur if all

Figure 2. Distribution of School Size by Grade Level

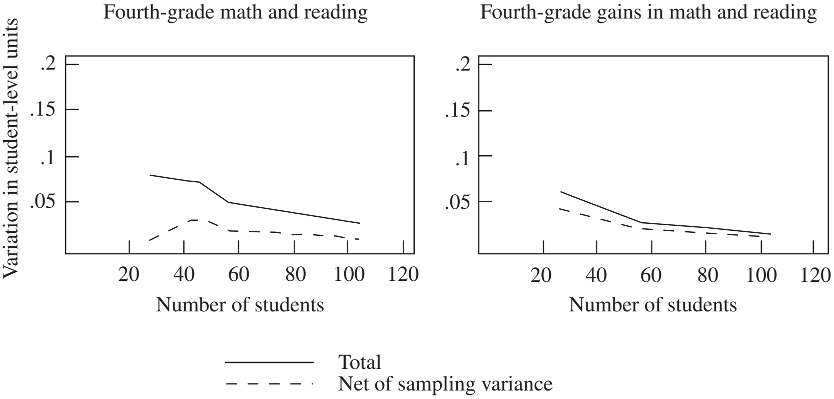


large schools were found in urban settings and if small schools contained a mixture of suburban and rural schools). Test scores also fluctuate much more from year to year among small schools than among large schools. Figure 3 plots the variance in the change in test scores between 1998 and 1999 by school size in North Carolina. The panel on the left portrays the variance in the change for fourth-grade math and reading scores and for gains in math and reading scores. The dotted line in both panels represents the result of subtracting our estimate of the contribution of sampling variation to the variance in the change. The variance in the change for fourth-grade test scores was three times as large among the smallest quintile of schools than among the largest quintile of schools (.079 versus .027). Moreover, the variance in the change for fourth-grade gain scores was five times as large among the smallest quintile of schools than among the largest quintile of schools (.060 versus .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. Sources of variation may be present at the classroom level, generated, for example, by teacher turnover, classroom chemistry between a teacher and the class, or the presence of a disruptive student

Figure 3. Between-School Variance in Annual Change



in a class. Sources of variation may affect a whole school, such as a dog barking in the parking lot on the day of the test or inclement weather, and could 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 grade 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 stemming from weather changes. We have neither. However, we provide a simple method for estimating that fraction of the variation in test scores over time that can be attributed to all such nonpersistent variation, even if we cannot identify the individual components as neatly. Subsequently, we will subtract our estimate of the sampling variation to form an estimate of these other sources of nonpersistent variation.

Suppose that some fixed component of school performance did not change over time and suppose that fluctuations in school test scores fell into two categories: those that persist and those that are transient or nonpersistent. One might describe a school's test performance, S_t, as being the sum of three factors: a permanent component that 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. That is,

$$S_t = \alpha + v_t + \varepsilon_t,$$
$$\text{where } v_t = v_{t-1} + u_t.$$

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

$$\Delta S_t = S_t - S_{t-1} = v_t - v_{t-1} + \varepsilon_t - \varepsilon_{t-1} = u_t + \varepsilon_t - \varepsilon_{t-1}$$
$$\Delta S_{t-1} = S_{t-1} - S_{t-2} = v_{t-1} - v_{t-2} + \varepsilon_{t-1} - \varepsilon_{t-2} = u_{t-1} + \varepsilon_{t-1} - \varepsilon_{t-2}.$$

Suppose that u_t , u_{t-1}, ε_{t-1}, and ε_t are independent.[11] Then the correlation between the change this year and the change last year could be expressed as

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

The numerator is the variance in the nonpersistent 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 algrebra, the above equation could be rearranged to produce

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

The expression on the right side of the equation describes the proportion of the change in test scores that is attributable to nonpersistent 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 –2. That is, 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 nonpersistent 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 of 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, suppose that the weather was particularly beautiful, the students were particularly well rested, and an unusually talented group of fourth-grade students was present on test day in 1999. Then the change in test scores for fourth-grade students between 1998 and 1999 would be large and positive. Because 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. We would expect scores in 2000 to be back to the average and 1999 to still appear as a stand-out year. In other words, if changes were nonpersistent, 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.

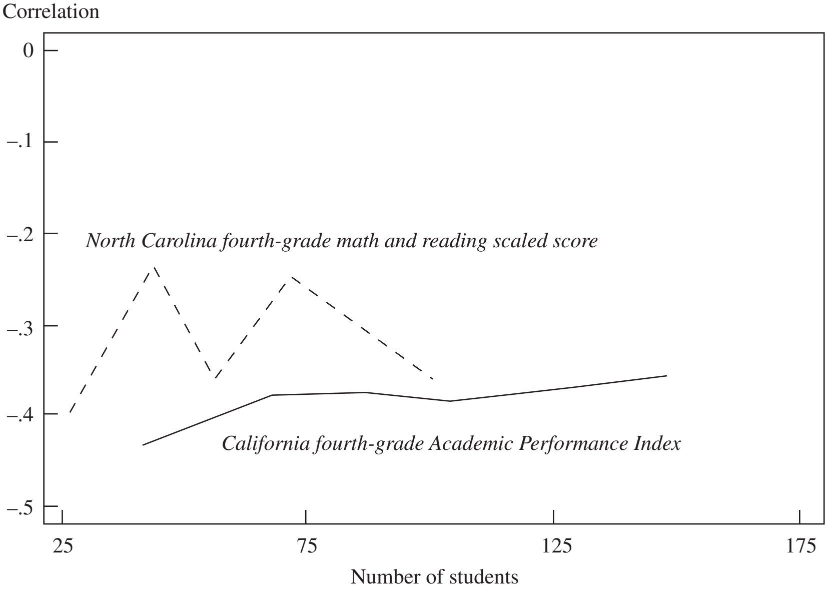
Suppose a school hired a new fourth-grade teacher in 1999 and improved facilities, thereby raising test performance. The school may make other such changes in the year 2000, but the magnitude of the changes one year provides no information about the expected magnitude of any such changes the next year. 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 instead of disappearing, we would expect a correlation of 0 in the change from one year to the next. If some changes are permanent, and some changes are purely transitory, one would expect a negative correlation between 0 and –.5.

The above estimator is focusing only on the transience of any changes in performance. Long-standing differences between schools 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, because we are interested in the proportion of change that persists, not the proportion of baseline differences that persist.

We calculated the mean fourth-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 Academic Performance Index scores in California for fourth-grade students and again calculated the correlation in the change in adjacent years. Figure 4 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 nonpersistent. 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 nonpersistent 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

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



adjacent years was –.43, implying that 86 percent of the variance in the changes between any two years is fleeting. For the largest fifth of schools, the correlation was –.36, implying that 72 percent of the variance in the change was nonpersistent.

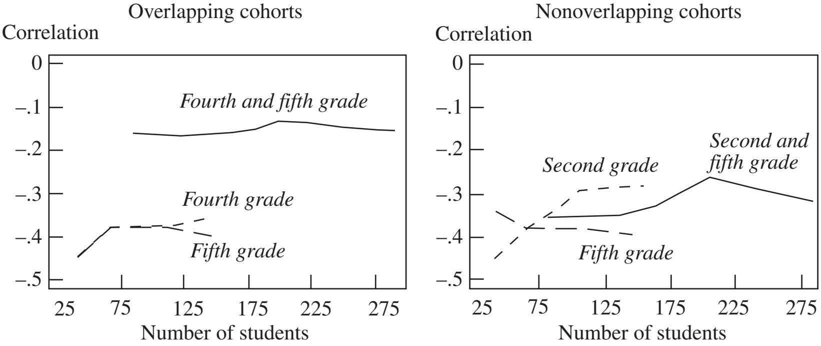
In California, the correlations clearly rise (become less negative) for the larger schools. This is what one would expect if a source of nonpersistence was sampling variability. In North Carolina, the pattern is less evident. However, this is presumably because of the smaller number of schools within each size quintile in North Carolina relative to California.

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, instead of on a single grade. The use of schoolwide averages has two primary effects. First, combining data from different grades increases the sample size and, therefore, reduces the importance of sampling variation. Second, considerable overlap exists in the 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 are. Consider an extreme example in which schools' long-term average performance does not change at all and any observed change in test performance is solely due to sampling variation. We would expect a correlation of –.5 in the change in performance in consecutive years for any given grade level, because any change would be nonpersistent. However, suppose we were using the change in a school's combined performance on fourth- and fifth-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 fourth-grade cohort from year *t*-1 is a particularly stellar group of kids. If we were only looking at fourth-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 fourth graders in year *t*-1 will repeat again as a stellar group of fifth graders in year *t*, any falloff 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 only after the random draw of students has been made in one year is there 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 considerable uncertainty.

Figure 5 portrays the correlation in changes in scores in consecutive years when combining two grades that would not overlap in three years, second and fifth grade, and when combining two grades that do overlap, such as fourth and fifth grade. Combining second- and fifth-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 nonpersistent. The correlation for the smallest quintile was –.37. However, when combining fourth- and fifth-grade scores, there is a discontinuous jump in the correlation. Instead of having a correlation of –.3, the correlation for all quintiles was close to –.15. Using schoolwide averages, combining test scores across grades, leaves the impression of greater stability.

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



Disaggregating the Variance in Scores into Persistent and Nonpersistent 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 nonpersistent variance. 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 nonpersistent 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 nonpersistent, we can use all these pieces of information to complete the puzzle and to generate an estimate of the variance due to nonpersistent factors other than sampling variation. The top panel of table 2 decomposes the variance in fourth-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 fourth-grade scores between years.

Three results in table 2 are worth highlighting. First, a school's average test performance in fourth grade can be measured reliably. Even among the smallest quintile of schools, nonpersistent 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 students' incoming performance, much of that reliability may result from the unchanging characteristics of the populations feeding those schools and not necessarily from unchanging differences in school performance.

Table 2. Decomposing Variance in School Test Scores due to Sampling Variation and Other Nonpersistent Factors

School size	*Average size*	*Total variance*	*Sampling variance*	*Other nonpersistent variance*	*Total proportion nonpersistent*
Combined reading and math scores in fourth grade					
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 third and fourth 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 fourth-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 fourth-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 nonpersistent factors was calculated as $-\rho_{\Delta_t\Delta_{t-1}}\sigma^2_{\Delta}-\sigma^2_{Samp}$ for each quintile, where $\rho_{\Delta_t\Delta_{t-1}}$ is the correlation in adjacent year changes for that quintile, σ^2_{Δ} is the variance in the change for that quintile, and σ^2_{Samp} is the estimated sampling variance for that quintile. Sampling variance and other nonpersistent variance for changes in test score levels were estimated by doubling the variances in the top panel. For an alternative estimator and standard errors, see Thomas J. Kane and Douglas O. Staiger, "Improving School Accountability Measures," Working Paper 8156 (Cambridge, Mass.: National Bureau of Economic Research, March 2001).

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 nonpersistent factors. Among schools near the median size in North Carolina, nonpersistent 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, nonpersistent 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 was roughly four times as large as the smallest quintile, the proportion of the variance in annual changes due to nonpersistent factors declined only slightly, from 79 percent to 73 percent. One might have the illusion of greater stability by combining multiple grades, but it is bought at the price of holding schools accountable for the past variation in the quality of incoming cohorts.

Instead of 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 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.

Implications for the Design of Incentive Systems

According to *Education Week*, forty-five states were providing annual report cards on their schools' performance in January 2001 and twenty states were providing monetary rewards to teachers or schools based on their performance.[12] However, the incentive systems have been designed with little recognition of the statistical properties of the measures upon which they are based. 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.

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 weak incentives for large schools.

Each year since 1997, North Carolina has recognized the twenty-five 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, 101 awards were given to schools ranking in the top twenty-five. (One year, two schools tied at the cutoff.) These 101 awards were won by 90 schools, with only 9 schools winning twice and only 1 school winning three times. No school was in the top twenty-five in all four years.

We have analyzed data for 840 elementary schools in North Carolina for which we had test score data for each year between 1994 and 1999. Of these schools, 59 were among the top twenty-five at some point between 1997 and 2000 (the top twenty-five each year included middle schools, which we are not analyzing here). Table 3 presents information on the mean gain scores in math in fourth and fifth grade, the variance in school mean gain scores, and the probability of winning a top twenty-five 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 five 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 twenty-five 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 twenty-five schools at some point over the four years the awards have been given. Even though their mean gains were not statistically different, the smallest schools were twenty-three times more likely to win a “Top 25” award than the largest schools (.277 / .012).

But, for the same reason, small schools are also overrepresented among those with extremely low test scores. Also beginning in 1997, the state assigned assistance teams to intervene in schools that had the poorest performance on the state tests and that also did not meet growth targets from the previous year. Table 3 also reports the proportion of schools in each school size decile that was 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 was 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 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

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

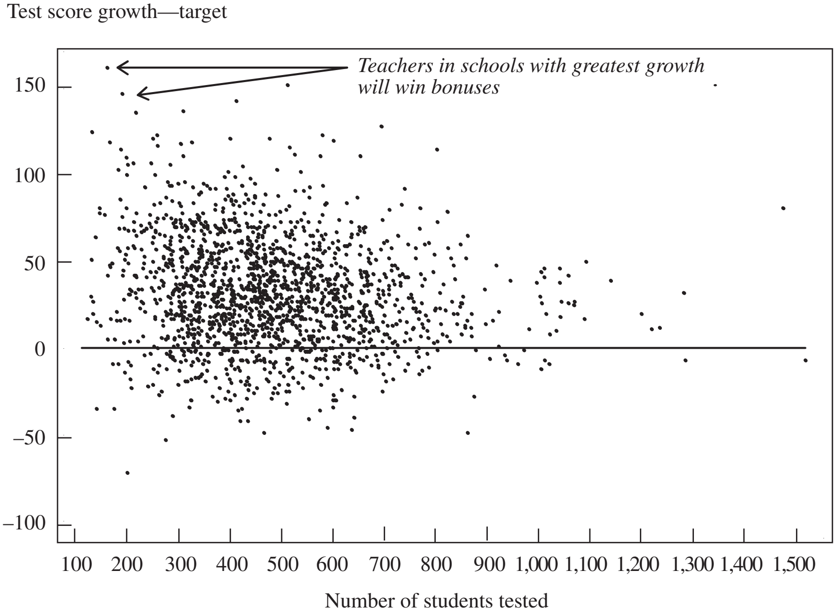
School size	*Mean gain in math*	*Between-school variance in mean gain in math*	*Percent ever "Top 25," 1997–2000*	*Percent ever assigned assistance team, 1997–2000*
Smallest decile	.020	.048	27.7	1.2
Second	-.007	.030	11.8	4.7
Third	.008	.028	8.2	7.1
Fourth	.009	.026	3.6	1.2
Fifth	-.002	.024	2.4	0
Sixth	.019	.018	3.6	0
Seventh	.007	.016	4.8	0
Eighth	.006	.016	7.1	0
Ninth	-.007	.015	0	1.2
Largest decile	-.011	.011	1.2	0
Total	.004	.023	7.0	1.5

Note: The table refers to the 840 regular public elementary schools for which the authors had data from 1994 through 2000. Charter schools are not included.

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.

In 2001 California planned 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.) 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 one hundred students.[13] Figure 6 plots the change in API scores by school size between 1999 and 2000 for those schools that met these requirements. One thousand teachers in schools with the largest improvements will receive $25,000 bonuses. Then, 3,750 teachers in schools with the next largest improvements in test scores will receive $10,000 bonuses. Finally, 7,500 teachers will receive $5,000 bonuses. The winners of the largest awards will generally be at smaller than median-size schools. Given the importance of sampling variation, this is hardly a surprise. Particularly when it comes to changes in test scores over time, the outlier schools will tend to be small schools.

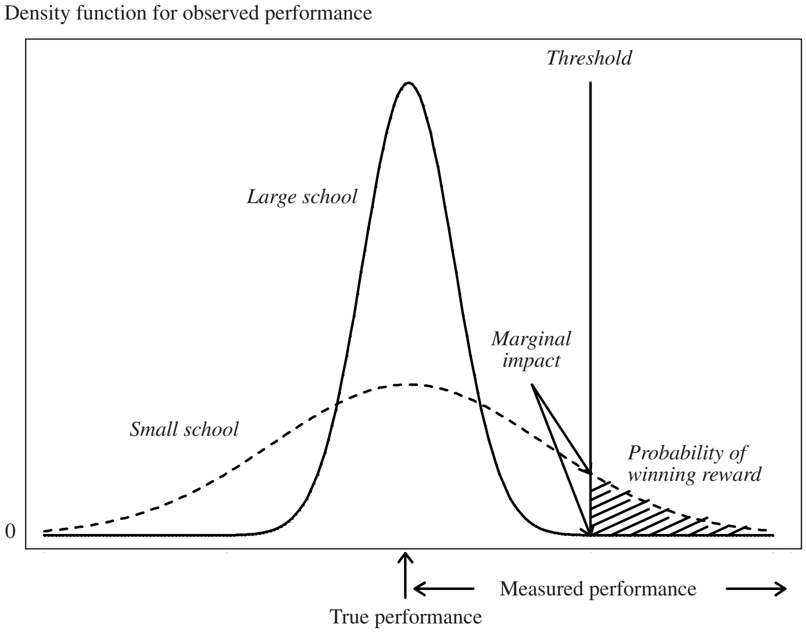
Figure 6. Improvements in Test Scores among California Schools Eligible to Win Teacher Bonuses by School Size



Note: Reflecting program rules, the figure has been drawn only for those elementary schools in the bottom five deciles of Academic Performance Index scores in 1999, with non-negative changes in test scores between 1998 and 1999, with at least one hundred students tested.

Rewards or sanctions for extreme test scores or large changes in test scores have little impact on large schools, because large schools have little chance of ever achieving the extremes. Figure 7 illustrates the point with a hypothetical example. Suppose a small and a large school had the same expected performance. But because of sampling variation and other factors that can lead to temporary 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, the large school will have little chance of winning the award if it does nothing and the small school will have a non-negligible chance of winning the award if it does nothing. Because 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 per-

Figure 7. Precision of Test Score Measures and Incentive Effects



formance 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 schools' 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, because the marginal effect of improving their performance on the likelihood of winning will be 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, the costs of coordinating the efforts of a larger number of teachers to implement a new curriculum likely would be larger.

A remedy would be to establish different thresholds for different size 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 percent 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.

Helen F. Ladd and Charles Clotfelter as well as David Grissmer and his colleagues report evidence suggesting that schools respond to incentives by raising student performance.[14] However, the long-term impacts of incentives may be substantially different from the short-term impacts. Even if school teachers are not sufficiently aware of the forces at work in an incentive system to analyze their incentives in a manner similar to that in figure 7, they may 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.

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

The accountability system in a number of states, including California and Texas, establishes 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 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, instead of a single toss.[15] 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, to be numerically significant, a group must represent at least 15 percent of the student body and contain more than thirty students, or represent more than one hundred students regardless of their percentage. There are eight different groups that 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, and socioeconomically disadvantaged students.[16]

Table 4. Proportion of California Elementary Schools Winning Governor's Performance Awards by School Size and Number of Numerically Significant Subgroups

	Number of numerically significant subgroups				
School size quintile	*1*	*2*	*3*	*4+*	*Total*
Smallest quintile					
Proportion winning	.824	.729	.587	.471	.683
Average growth in API 1999–2000	33.4	45.6	42.2	36.0	41.2
Number of schools	204	343	349	51	947
Second quintile					
Proportion winning	.886	.769	.690	.670	.749
Average growth in API 1999–2000	29.9	42.6	42.2	43.9	40.5
Number of schools	158	337	358	94	947
Third quintile					
Proportion winning	.853	.795	.708	.667	.756
Average growth in API 1999–2000	26.8	36.3	38.9	44.6	36.6
Number of schools	156	308	390	93	947
Fourth quintile					
Proportion winning	.903	.823	.776	.656	.799
Average growth in API 1999–2000	28.0	41.8	39.5	40.8	38.7
Number of schools	144	328	379	96	947
Largest quintile					
Proportion winning	.876	.776	.726	.686	.755
Average growth in API 1999–2000	29.5	37.9	36.9	40.5	37.0
Number of schools	89	370	387	102	948
Total					
Proportion winning	.864	.778	.699	.647	.749
Average growth in API 1999–2000	29.8	40.9	39.9	41.7	38.8
Number of schools	751	1,686	1,863	436	4,736

Note: API = Academic Performance Index. Reflecting the rules of the Governor's Performance Award program, the table was limited to elementary schools with more than one hundred students.

Table 4 reports the proportion of California elementary schools winning the Governor's Performance Award by school size quintile and number of numerically significant subgroups in each school. Among the smallest quintile of elementary schools, racially heterogeneous schools were almost half as likely to win a Governor's Performance Award as racially homogeneous schools: 47 percent of schools with four or more racial, ethnic, or socioeconomic subgroups won a Governor's Performance Award as opposed to 82 percent of similar-size 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 because of space limitations, such schools 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, homogeneous 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).

The data in table 4 have at least two important implications. First, under such rules, a district would have a strong incentive to segregate by race or ethnicity. For instance, suppose there were four 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 put 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 or 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).[17]

Although the costs of the subgroup targets are clear, the benefits are uncertain. Policymakers might want to know whether the rules force schools to focus more on the achievement of minority youth. If so, some consideration of the test scores of racial or ethnic subgroups may be worthwhile, despite the costs. One way to estimate this impact would be to compare the improvements for minority youth in schools where they are 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.[18] Despite the costs, the evidence does not suggest that such thresholds 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 unreliable. More efficient ways exist to pool information across schools and across years to identify those schools 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 to be. However, policymakers and school administrators often are uncertain (or, at least, they disagree) about what the right thing is. The state may also have an interest in helping to identify the schools that are worth emulating.[19] If the goal of an accountability system is not only to provide incentives, but also to help identify success, the absolute amount of imprecision and the amount of imprecision relative to the degree of underlying differences determine the likelihood of success in the search for exemplars.[20]

Building upon work by Mark McClellan and Douglas Staiger 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 performance.[21] 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 (3) ignore the school's scores and assume that student performance in the school is equal to the state average. 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 school performance did not appear to vary much across the state, then one would be tempted to choose the last option—place less weight on the school's scores because of their low signal-to-noise ratio and heavily weight the state average.

Alternatively, if that school performance seemed to change 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 considerable persistent change is evident 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 North Carolina elementary schools ranking in the top 10 percent in fifth-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 it reins in the prediction according to the amount of sampling variation and nonpersistent 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 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 twenty-five 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 fifth-grade math gains was .124 student-level standard deviations above the average in 1999. Among the twenty-five schools chosen using actual 1997 scores that 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 report the difference in mean scores moving across the first two columns or first two rows. Among those that were not identified as being in

Table 5. Performance of North Carolina Schools in 1999 Identified as in the Top 10 Percent in 1997, Based on Actual and Filtered Test Scores

		Based on actual 1997 score				
		School not in top 10 percent	*School in top 10 percent*	*Row total*	*Difference between top 10 percent and the rest*	*Expected difference*
Based on filtered prediction of 1999 score (from 1997)	School not in top 10 percent	-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 in top 10 percent	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 percent 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)		

Note: Within the box, the entries report the mean of the fifth-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 the top 10 percent and to calculate the expected difference between the top 10 percent and the rest. The rows of the table use filtered predictions of 1999 scores, based only on data from 1994–97, to assign schools to the top 10 percent.

the top 10 percent by the filtering method, knowing that they were in the top 10 percent on the actual 1997 score provided little information regarding test scores. The test scores were –.006 standard deviations lower on average holding the filtered prediction constant. In contrast, among those 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. 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 disappointed, because the actual difference in performance was only .115

standard deviations. 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 before 1998. The actual difference in performance for these schools was .160 standard deviations.

Table 6 compares the R^2 one would have obtained using three different methods to predict the 1998 and 1999 test scores of schools using only the information available before 1998. The first method is the filtering method. The second method is using the actual 1997 score as the prediction for the 1998 and 1999 scores. The third method uses the four-year average of math performance before 1998 (1994–97) to predict 1998 and 1999.

Whether one is trying to anticipate math or reading levels or gains in fifth grade, the filtering method leads to greater accuracy in prediction. The R^2 in predicting fifth-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 and simply predicting that performance in every school would be equal to the state average. 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 nonpersistent 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 nonpersistent

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

	Unweighted R^2 when forecasting 1998 and 1999 scores being predicted under alternative uses of 1993–97 data					
	Predicting scores in 1998 (one-year ahead forecast R^2)			*Predicting scores in 1999 (two-year ahead forecast R^2)*		
Test score being predicted	*Filtered prediction*	*1997 Score*	*Average score, 1994–97*	*Filtered prediction*	*1997 Score*	*Average score, 1994–97*
Adjusted score						
Fifth-grade math	0.41	0.19	0.29	0.27	-0.02	0.13
Fifth-grade reading	0.39	0.13	0.33	0.31	-0.05	0.24
Gain score						
Fifth-grade math	0.16	-0.27	0.09	0.12	-0.42	-0.01
Fifth-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 1993–97 data.

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 disadvantage of the filtering technique is that it is much less transparent.[22] 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, a number of mysterious calculations are 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 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.

North Carolina in 1997 identified fifteen elementary and middle schools with poor performance in both levels and gains and assigned assistance teams of three to five educators to work in these schools. The next year, all of the schools had improved enough to escape being designated as low-performing. 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.[23]

The value of the assistance teams was lauded in *Education Week*'s annual summary of the progress of school reform efforts in the states.[24] However, given the amount of sampling variation and other nonpersistent 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 thirty-five elementary schools that won a “Top 25” school award in either 1997 or 1998 as well as for ten 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. In the year of the award, their scores were .230 student-level standard deviations above the mean gain. However, in both the year before their award and the year after, their gain scores were slightly below the mean gain.

Moreover, the schools that were assigned assistance teams seem to have had a particularly bad year the year of their receiving the sanction. In the year before assignment, such schools had an average fourth-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

Table 7. Fourth-Grade Test Scores before and after Sanction or Reward in North Carolina

Group of schools	*Year before award or sanction*	*Year of award or sanction*	*Year after award or sanction*	*Year after– year before*	*Ratio of $S_{t+1}-S_{t-1}/ S_{t+1}-S_t$*
"Top 25" in 1997–98					
Math + reading gain score	-.003	.230	-.064	-.062[a] (.164)[b]	.211
Assistance team in 1997–98					
Math + reading test score	-.668	-.786	-.523	.145[a] (.059)[b]	.551
Math + reading gain score	-.078	-.134	.078	.156[a] (.006)[b]	.735

Note: Test scores are in units of student-level standard deviations. The mean test scores across all schools in each year have been subtracted. If a single school won an award more than once, we used its first award. Thirty-five elementary schools in our sample won a "Top 25" award in 1997 or 1998 based upon their gain scores. Ten elementary schools in our sample were assigned an assistance team in 1997 or 1998 based upon a combination of low test scores and low gain scores.

a. Difference.

b. *p*-value.

was even lower, .786 student-level standard deviations below the average school. The year after assignment, their scores seemed to rebound to .523 student-level 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 have improved their performance over time, by a fairly sizable .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.

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 want to focus on a school's mean value-added in the most recent year or on the changes in schoolwide scores over the most recent two years. However, 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 system 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.

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 significantly different from the short-term impacts. In 1967 psychologists Martin E. P. Seligman and Steven F. Maier published the results of an experiment in which one group of dogs was strapped into a harness and administered a series of electrical shocks through electrodes attached to their feet.[25] 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 lay 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 learned that they could do little to avoid the shocks. (A second group of dogs, which was 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. Caution must be taken about the lessons teachers and principals are learning about their ability to determine those outcomes and about how their efforts will be rewarded.

However, the results of our research should not be interpreted as implying that all accountability systems are necessarily flawed. 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 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 earlier research, we describe an estimator that does that in a more efficient way than simply taking a simple mean across as many years as possible.[26] 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.

Comment by David Grissmer

The issue addressed by Thomas J. Kane and Douglas O. Staiger is whether schools can reliably be chosen for rewards or sanctions based on year-to-year test score gains. The question is whether picking schools based on gains identifies good or bad schools, or lucky or unlucky schools. The authors' analysis convincingly concludes that methods relying on gain scores at a given grade are mostly identifying lucky and unlucky schools, not good and bad schools. The reason for misidentification is that the variance due to sampling and other sources of noise can be a significant portion of the variance in gains across schools. In this area, standards-based reform is far ahead of statistical reliability.

This paper required several readings to extract the nub of the argument and analysis. I think the exposition can be improved. Basically the focus is on five quantities and their relationship and relative size: between-school score variance in annual scores, between-school variance in score gains from grade to grade, between-school variance in year-to-year score changes at a given grade, sampling variance, and variance from other sources of noise.[27] The basic argument is that to reliably identify good and bad schools by any criteria requires that the signal be much greater than the sources of noise. The signal in this case is the portion of a score or score gain that can be attributed to school effort. The noise is caused by sampling variability from a hypothetical student population and other sources of random noise.

The paper estimates these parameters using data from North Carolina and California, and it shows that the sources of noise are too large relative to the signal to allow reliable identification of good or bad schools. More often than not, the reason a school ends up near the top or bottom of a ranking can be attributed to random factors and not real improvement. Also because the sampling variation decreases with school size, small schools are disproportionately represented at both the top and bottom part of rankings. Perhaps as important, systems that use criteria involving separate consideration of scores by race or ethnicity are likely to make even poorer identification.

I like the way the paper is designed. The authors develop a statistical model, make estimations for parameters in the model, make predictions from the model, and use the data to verify the predictions. The authors draw out the important policy implications and provide guidance on how to improve the identification process.

The parameters are estimated using third- and fourth-grade data from North Carolina. The method used to estimate random sources of noise outside sampling variability is neat. The model and parameter estimations lead to predictions that use of year-to-year gain scores leads to small schools being disproportionately identified as good or bad schools. This prediction is verified in two ways. First, small schools in North Carolina have over twenty times the probability of being identified in the top distribution of rewarded schools. Second, rarely are schools that are rewarded in one year also rewarded in the following years. This nonpersistence of performance implies that nonpersistence sources of error are probably a major component of the actual gains.

The situation gets even worse if rewards or sanctions depend on score gains by racial or ethnic groups within grades, which is an increasingly common practice. The identification is then based on even smaller sample sizes, and the chances of high gains of all racial or ethnic groups become even more dependent on chance. The policy implications cited by the authors include all the morale issues arising from having rewards or sanctions based on factors other than real performance to misidentifying the reasons that schools are improving by focusing on the wrong schools.

The authors analyze possible solutions to the problem of increasing the reliability of the identification process. They analyze pooling scores schoolwide instead of using individual grade scores. The increased reliability from this type of pooling is not as large as one would expect from the increased sample size because student characteristics persist from grade to grade as a cohort flows through grades. Using score gains averaged over longer time periods—that is, sustained high or low performance—can significantly improve reliability. The authors also suggest a more sophisticated statistical filtering technique designed to improve the signal-to-noise ratio.

Overall, the paper is an outstanding contribution at three levels: the methodology, the quality of results, and the policy implications. The methodology—used more often in physics, statistics, and information theory—brings a new perspective and set of tools to analysis of achievement data. More of this kind of analysis and its logical extensions likely will be seen in the future. The results are robust and almost beyond argument. This paper may be one of only a few that would generate widespread consensus among researchers. Finally, this paper has immediate and important policy implications for educational policy. The questions addressed are important ones currently being considered in national legislation and across many states. Following the advice in the paper will improve public policy in education.

My comments are directed to making the statistical model reflect the more complex aspects of the educational system and to place the results in a wider perspective on the results of accountability systems across states. The current statistical models underlying the analysis do not yet reflect much of the complexity of the system. At least three other factors affect the variance of gain scores besides sampling that change with the number of students in a grade in a school. The first factor is that the variance in teachers will be different in schools with more students at a given grade. Small schools will have one teacher per grade while larger schools will have several classes and teachers per grade. Assuming that teachers are randomly assigned and have differential effects on achievement, then this teacher effect would narrow variance in larger schools.

The second factor is that the likelihood of being in a small class also varies by size of school. Schools with fewer students are more likely to be in smaller classes than schools with more students. Class sizes are often determined by setting a limit on class size, which requires the creation of another class if that limit is exceeded. For instance, if the limit is twenty-five but thirty students enroll, then two classes of fifteen will be created. As the number of students increases, the average class size will approach the limit. Compelling experimental evidence exists to suggest class size affects achievement, implying that more variance is introduced in smaller schools than larger schools as a result of this effect.[28] Finally, teacher turnover is higher in urban areas where school sizes are larger. Higher teacher turnover will increase variance in gain scores for larger schools.

The first two effects are intrinsic characteristics of small and large schools and largely independent of educational policy. Like sampling variability, they increase the variance in small schools and exacerbate the effects described by the authors. A more complex statistical model could capture these effects. The effect on variance of teacher turnover probably can be influenced by educational policy and cannot automatically be classified as noise instead of part of a real signal.

The question of separating effects into persistent and nonpersistent is also more complex than modeled in the present analysis, but the models can be extended to include this complexity. Some effects can persist over several years and then decay, and persistence can be different depending upon types of students. The Tennessee experimental results seem to imply that the effects persist if students are in small classes for three to four years, but not for one to two years.[29] A good teacher may also have an effect not only on students in the present grade, but also in future grades. A 2000 study by David W. Grissmer and others suggested that persistence effects can be captured only by interaction terms between schooling conditions in early years and later years.[30] Persistence may be easier to achieve with early interventions, but much more difficult with later interventions. The complexity of persistence effects makes modeling only adjoining years to separate persistence and nonpersistence problematic.

The current results are not very damaging to the standards-based accountability movement for two reasons. First, much of the problem can be fixed by using gains over several years. Second, financial rewards are not central to successful accountability systems. Successful accountability relies primarily on having more and better quality data linked to standards that can be used to diagnose problems from the student to the teacher to the school to the school district. The availability and informed utilization of the data by parents, teachers, principals, school districts, and state policymakers provide increased achievement. Standards together with these data allow better resource allocation from the student to the state and represent the central component of accountability, not financial rewards.

Finally, this is one of the few papers that I have ever read in education where I thought that a consensus among researchers is possible. However, no consensus mechanism in educational research would allow this consensus to be recognized. Other areas of research such as health have consensus panels that are able to generate an important dialogue about research results. The absence of such mechanisms in education is a significant problem for improving the quality of research and informing public policy.

Comment by Helen F. Ladd

Most states now have educational accountability systems based on student achievement as measured by test scores. Such systems can direct attention to districts, schools, individual teachers or students, or some combination thereof. In their excellent paper, Thomas J. Kane and Douglas O. Staiger treat schools as the unit of accountability. That is, they focus on accountability programs in which states measure the effectiveness of individual schools and then use those measures as the basis for providing awards, imposing sanctions, giving assistance, or identifying exemplars.

As the authors emphasize, many states have introduced such programs without a full understanding of the underlying statistical characteristics of relevant measures. In light of this observation, the primary contribution of the paper is methodological. Kane and Staiger explain and document the importance of one basic characteristic, the volatility of the measures, using data from North Carolina and California elementary schools. They then spell out the implications of that volatility for the design and use of measures of school performance. By combining sophisticated, but intuitively understandable, statistical analysis with clear and compelling applications to the policy debate, the authors provide some powerful new insights into an important current policy issue.

The Problem of Volatility

A school's performance can be measured in at least three generic ways: as the mean of student test scores in the school, as the mean of the gains in student test scores during a year, or as the average annual change in test scores in the same grade from one year to the next. Regardless of the approach, the measures will be subject to what the authors refer to as nonpersistent variation across schools. This nonpersistent chance variation, or noise, refers to variation that does not reflect true differences in performance across schools. Such noise has two sources: sampling variation that arises because of the characteristics of the particular samples of students being tested and other nonpersistent variation that arises because of one-time idiosyncratic factors that influence test results in any particular year, such as a disruption in the classroom or in the school.

The relevant policy question is how large the nonpersistent variation is relative to the true signal or, in practice, relative to the total observed variation. Determining how much of the total variation in student performance across schools is noise is not straightforward and requires various assumptions that Kane and Staiger spell out. The basic idea is that the noise share can be estimated from information about how the changes in test scores in each school are correlated over time. The more negatively correlated these changes are, the greater is the share of nonpersistent variation. Using this method, Kane and Staiger conclude that the nonpersistent variation sometimes accounts for a large share of the total variation.

The magnitudes emerge most clearly in table 2, which is based on data from North Carolina elementary schools. That table indicates that the nonpersistent variation accounts for almost 15 percent of the total variation in levels of fourth-grade test scores across schools, almost 50 percent of the total variation in gains in scores during fourth grade, and a whopping 73 percent of the variation in annual changes in fourth-grade scores. The table also shows that the size of the school matters: The ratio of noise to total variation across schools is significantly larger for small schools than for larger schools.

Are these large noise ratios plausible? Do the patterns make sense? To answer these questions, it is useful to split the nonpersistent variation into the sampling variance and other persistent variance. Based on standard statistical theory, the sampling variance can be calculated as the average variance within schools divided by the average sample size. Hence the sampling variation is larger for small than for large schools. Other persistent variation is then calculated as the residual difference between the total nonpersistent variance and the sampling variance.

The numbers in table 2 indicate that the sampling variation alone accounts for about 11 percent of the total variation for a medium-size school when levels of test scores are used as the performance measure, about 13 percent when mean gains are used, and 55 percent when annual changes are used. These estimates all seem plausible and hard to refute. Of particular significance is the large share for the annual change measure of performance. Sampling variation is large for that approach because of the two different cohorts of students involved, those in the fourth grade one year and those in the fourth grade the following year. In addition, the table clearly demonstrates the larger sampling variance in the smaller schools.

For schools of average size, other nonpersistent variation accounts for about 35 percent of the variation in the mean gains of third to fourth graders across schools, a far larger share than for the other two performance measures. This large share reflects two factors. One is the relatively small variation in mean gains across schools, and the other is the fact that two test scores are involved for each student and hence two opportunities for idiosyncratic effects to emerge. The authors, however, may be overstating the problem of volatility in this case. While volatility of this form is demonstrably a problem when policymakers focus on the gains of a single cohort of students in a single grade, such volatility could be less of a problem when multiple grades are combined at the school level. Only if all the idiosyncratic factors operated at the school level (such as a commotion outside the school during the testing period) and not at the grade level would the move to the school level not reduce the overall variation.[31]

Lessons for the Design of Accountability and Incentive Systems

Kane and Staiger spell out four lessons for the design of test-based accountability systems that emerge from the presence of volatility. I generally agree with all of them, but with some qualifications. In addition, I add a few more of my own.

First, the authors assert that incentive systems that provide rewards or sanctions for schools at the extremes of the performance distribution primarily affect small schools and provide very weak incentives for large schools. Given the smaller sampling variation in performance measures for large than for small schools—and hence the smaller probability that a large school will be at the extremes of the distribution—it is certainly true that, for any given level of true performance, small schools have a higher probability of being rewarded or of being sanctioned than large schools. Stated differently, with extreme cutoffs, a small school has a higher probability of being miscategorized as a success or failure than a large school.

How this difference between small and large schools translates into the power of incentives for schools of different sizes to improve, however, is less clear. As Kane and Staiger point out, the long-term incentive impacts on schools could differ from the short-term impacts as school personnel in small schools have trouble perceiving a clear relationship between their effort and the school's performance. The fact that noise plays such a large role in the classification of small schools means that a small school will have difficulty determining what to continue doing if it is deemed a successful school or what to stop doing if it is deemed a failing school. Thus, I would be inclined to emphasize that an accountability system that rewards and sanctions schools at the extremes is an ineffective means of inducing even the small schools to improve over time.

The main way to justify such a system would be in terms of the general signal that it sends to all schools about the state's interest in improving student performance. The hope then would be that all schools, not just the schools with a chance of winning an award, would respond to the public pressure to improve student achievement in a positive way.

Second, Kane and Staiger show that incentive systems that establish performance thresholds for each separate racial or ethnic subgroup put racially or ethnically diverse schools at a disadvantage with respect to being rewarded and also encourage districts to establish racially homogeneous schools. Their analysis of the California data provides support for this conclusion. The authors provide further analysis that shows there may be few, if any, offsetting benefits of such disaggregation. Their observations on disaggregated measures seem valid and are central to current policy discussions about congressional proposals relating to accountability.

Third, the authors emphasize that annual test results, mean gains, or changes in test scores are flawed as a tool for identifying schools worth emulating, and they argue that filtered estimates based on a Bayesian approach would better serve that purpose.[32] These filtered estimates are based on a combination of the school's own test score, the state average, and the school's test scores from past years, other grades, and other subjects. As the authors document, the filtering system is more successful than the simpler gains approach in predicting which schools are likely to perform well in a subsequent year. The authors convincingly argue that state policymakers could and should use measures in that spirit for determining which schools are worth emulating. However, unless they can be simplified in a way that makes them transparent to the school officials whose behavior state policymakers are trying to influence, such measures would be less useful for the purpose of providing direct incentives for improving school performance.

Finally, Kane and Staiger warn policymakers that they should pay attention to the natural fluctuations in test scores in evaluating the impacts of any policy interventions. In this context, they are raising the standard problem of regression to the mean. This lesson is fine for—and is also well known to—policy analysts. However, from the perspective of policymakers who want policies to look successful, the schools with unusually low performance one year are precisely the ones they may want to target given such schools are more likely to improve the following year in any case. Perhaps Kane and Staiger's lesson is more relevant for the media, which, with more statistical sophistication, could hold policymakers more accountable for true policy impacts.

To this list, I would like to add a few more policy-relevant observations.

1. *If the purpose of the rewards and sanctions is to generate incentives for schools to improve, the thresholds or cutoff scores should not be at the extremes of the distribution.*

The policy recommendation that thresholds for recognition and rewards not be at the extremes of the distribution is implicit in the Kane and Staiger analysis but is never stated clearly given their tendency to focus on the differences between large and small schools. To be sure, even with thresholds closer to the mean performance measure across schools, the problem of noise does not go away. However, setting thresholds closer to the mean would give more schools a chance to win recognition or rewards and hence would extend direct incentives deeper into the distribution. Moreover, compared with a system that bases rewards on extreme values, the incentives for school improvement from such a system are likely to be more powerful given that the larger schools, whose chances of winning would be increased, have more control over their measured performance than do the small schools, whose performance measures are subject to so much noise.

2. *Financial awards should not be large, and for some decisions, information other than test scores should be brought to bear.*

The volatility in test scores across schools inevitably means than any incentive program is going to mislabel many schools. This is not just a problem at the extremes of the distribution, but also one that applies to all schools regardless of where they fall in the distribution. Such mislabeling could be acceptable provided that it does not distort school behavior in highly undesirable ways or lead to gross inequities among schools. One option to avoid those undesirable outcomes is to keep the financial rewards for positive performance relatively small. In the case of sanctions for low-performing schools, where the stakes may be high, policymakers would do well to supplement the information from the test-based measures of performance with other information, such as from site visits, about the performance of the school.

3. *Volatility of measures is important, but low volatility should not be the only criterion for deciding among approaches for measuring school performance.*

Some policymakers might be tempted to conclude from table 2 that measuring school performance using average test scores would be preferred to other approaches on the grounds that it leads to the lowest nonpersistent variation relative to the total variation across schools and hence generates the clearest signal about school performance. That conclusion, however, would be inappropriate. Among the three approaches examined by Kane and Staiger, average test scores generate the least valid measure of school performance. As has been well understood since the Coleman report of the 1960s, average test scores are highly correlated with the socioeconomic status of students in a school.[33] Hence, a school's average test score indicates more about the composition of students in the school than it does about the effectiveness of the school in imparting learning. Only if that measure were adequately corrected for the socioeconomic status of the student body would it be a valid measure of school performance, but such adjustments are hard to make. For that reason, some form of gain or change measure is preferred.

The annual change measure can and should be ruled out on the grounds that sampling variation from one year to another generates an unacceptably large amount of noise relative to true signal.[34] The only remaining question is whether the ratio of noise to total variation is also too high for the school performance measure based on mean gains of a given cohort of students. That is a judgment call. Certainly the amount of volatility reported in table 2 suggests there may be a problem, particularly for small schools. However, the aggregation of several grades would reduce the volatility somewhat, provided that the idiosyncratic effects on test scores were not schoolwide.

More generally, many issues other than volatility arise in developing a valid measure of school performance or school effectiveness. Kane and Staiger have made an important contribution by focusing on volatility, but that focus should not keep policymakers from asking the other hard questions related to the measurement of school performance, including the purposes for which the measure is to be used and whose behavior it is designed to change.[35]

Accountability in North Carolina

Kane and Staiger use North Carolina data to illustrate many of their points. Data from North Carolina are particularly useful because the state has administered statewide end-of-grade tests to all students in grades three to eight since 1993. Because Kane and Staiger's main purpose is methodological, they use the North Carolina data to highlight certain conclusions about volatility, not to discuss the broader set of issues related to accountability in that state. As a result, a reader might come away from the Kane and Staiger paper with a misleading sense of the North Carolina accountability system.

North Carolina's program is sophisticated and is less subject to some of the methodological problems discussed by Kane and Staiger than it would be if it were more like the California system that emphasizes the extremes of the performance distribution. Although North Carolina does publicly recognize the schools with the highest performance and those with the greatest gains, most of the accountability program, and all of the financial rewards, are directed toward a larger group of schools that are identified based on the gains in their test scores relative to the predicted gains for the school.[36] For elementary and middle schools, any school whose gains exceed its predicted gains by more than 10 percent is designated an exemplary school and financial bonuses are given to the teachers and staffs of such school. In 1997, after the first year of the program, about one in three schools met the exemplary status. By 1999, more than one in two schools were exemplary. Thus, the North Carolina accountability program is targeted at a much larger proportion of schools than would be true of one focused only on the very highest performing schools.

At the bottom end of the distribution, North Carolina uses test scores to identify low-performing schools that receive both additional scrutiny and attention from state assistance teams. The criteria for being a low-performing school are twofold. One is that the school did not meet its expected growth in test scores during that year, and the other is that less than 50 percent of the students were at grade level. Thus, even this cutoff is more complex, and may be subject to less volatility, than some of the measures discussed by Kane and Staiger.

North Carolina's school-based accountability system has had a powerful effect on the behavior of one set of key adults in the education system—school principals. This assertion is based on evidence from surveys of a random sample of elementary school principals in 1997 and 1999.[37] Analysis of the survey responses indicates that most principals, including both supporters and nonsupporters of the state's overall goals, responded to North Carolina's accountability program in ways that were consistent with the state's goal of focusing attention on the basic skills of reading, math, and writing. For example, by 1999, most principals had redirected resources to math and reading, incorporated math and reading into other courses, increased their work with teachers to prepare for the end-of-grade tests and to improve instruction, and incorporated math and reading into extracurricular activities.[38] In addition, the program induced many principals to focus more attention on test-taking skills or on other activities that would improve a school's rating but not necessarily student learning. Thus, Helen F. Ladd and Arnaldo Zelli conclude that a well-designed accountability program can be a powerful policy tool, and for that reason, they urge that policymakers use it cautiously.

A follow-up analysis of some of the data presented in Ladd and Zelli shows that Kane and Staiger's emphasis on the differential magnitude of the incentives facing small schools relative to large schools does not apply to the North Carolina program. To test for differential effects, the survey responses were divided by size of school and statistical tests undertaken to determine whether principals of small schools responded more strongly to the incentives of the program than principals of large schools. Out of fourteen specific comparisons, only one statistically significant difference emerged. In that case, the larger schools responded more strongly than the smaller schools.

As the state moves forward with its accountability system and its efforts to reduce the black-white gap in test scores, state policymakers would do well to heed Kane and Staiger's second lesson about the dangers of basing rewards on test scores disaggregated by subgroup within schools and their third lesson about the need for care in choosing schools to use as exemplars of outstanding performance.

Notes

[1] James S. Coleman and others, *Equality of Educational Opportunity* (Department of Health, Education, and Welfare, 1966).

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

[3] In addition, the survey contained information on parental educational attainment reported by students. We use these data when attempting to control for the socioeconomic background of students.

[4] Thomas J. Kane and Douglas O. Staiger, "Improving School Accountability Measures," Working Paper 8156 (Cambridge, Mass.: National Bureau of Economic Research, March 2001).

[5] For 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 (www.cde.ca.gov/psaa/api/yeartwo/base/apiinfogb.pdf).

[6] Coleman and others, *Equality of Educational Opportunity,* p. 326.

[7] In 1996–97, 51,306 public schools in the United States had a third grade. (See Department of Education, National Center for Education Statistics, *Digest of Education Statistic*s

(1998), p. 119, table 99.) This included 4,910 schools with grades prekindergarten, kindergarten, or first through third or fourth grade; 20,570 schools with prekindergarten, kindergarten, or first through fifth grade; 15,578 schools with prekindergarten, kindergarten, or first through sixth grade; 4,543 schools with prekindergarten, kindergarten, or first through eighth grade; and 5,705 schools with other grade-spans. Moreover, in the fall of 1996, 3.518 million students in the United States were enrolled in third grade. (See *Digest of Education Statistics*, p. 58, table 43.)

8. Darcia Harris Bowman, "Arizona Ranks Schools by 'Value-Added' to Scores," *Education Week*, February 9, 2000.

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

10. Caroline Hoxby, "Peer Effects in the Classroom: Learning from Gender and Race Variation," Working Paper 7867 (Cambridge, Mass.: National Bureau of Economic Research, August 2000).

11. In 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. (We observed no obvious evidence of ceiling effects in the data.) However, 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, "Improving School Accountability Measures."

12. Ulrich Boser, "Pressure without Support," *Education Week,* January 11, 2001, pp. 68–71, table.

13. Schools also had to meet targets for each numerically significant racial or ethnic group.

14. See Helen F. Ladd and Charles Clotfelter, "Recognizing and Rewarding Success in Public Schools," in Helen F. Ladd, ed., *Holding Schools Accountable* (Brookings, 1996), pp. 23–64; and David Grissmer and others, *Improving Student Achievement: What State NAEP Test Scores Tell Us* (Santa Monica, Calif.: RAND, 2000). Nevertheless, whether the improvement in performance is real or the result of teaching to the test is unclear. See Daniel Koretz, "Limitations in the Use of Achievement Tests as Measures of Educators' Productivity," *Journal of Human Resources* (forthcoming), paper initially presented at "Devising Incentives to Promote Human Capital," National Academy of Sciences Conference, December 1999. Kane and Staiger, in "Improving School Accountability Measures," report that the schools in North Carolina that showed the greatest improvement on fifth-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 an hour of homework, or the proportion of students watching less than six hours of television), even though these characteristics were related to gain scores in the base year.

15. However, 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.

16. A 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.

17. The 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 or ethnic differences

in spending is not precisely correct, because 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.

18. See Thomas J. Kane, Douglas O. Staiger, and Jeffrey Geppert, "Assessing the Definition of 'Adequate Yearly Progress' in the House and Senate Education Bills," University of California at Los Angeles, School of Public Policy and Social Research, July 2001.

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

20. A similar point is made in David Rogosa, "Myths and Methods: 'Myths about Longitudinal Research' plus Supplemental Questions," in John Mordechai Gottman, ed., *The Analysis of Change* (Mahwah, N.J.: Lawrence Erlbaum Associates, 1995), pp. 3–65.

21. Mark McClellan and Douglas Staiger, "The Quality of Health Care Providers," Working Paper 7327 (Cambridge, Mass.: National Bureau of Economic Research, August 1999). See also Kane and Staiger, "Improving School Accountability Measures." The estimator is an empirical Bayes estimator in the spirit of Carl Morris, "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, vol. 381, no. 78 (1983), pp. 47–55. However, the method employed for estimating the variance components is less computationally intensive than that proposed in Anthony Bryk and Stephen Raudenbush, *Hierarchical Linear Models* (Newbury Park, Calif.: Sage Publications, 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.

22. For a discussion of the merits of transparency, see Ladd and Clotfelter, "Recognizing and Rewarding Success in Public Schools."

23. North Carolina Department of Public Instruction, "ABCs Results Show Strong Growth in Student Achievement K–8," August 6, 1998 (www.ncpublicschools.org/news/abcs_results_98.html).

24. Kathleen Kennedy Manzo, "North Carolina: Seeing a Payoff, " *Education Week,* vol. 18, no. 17 (January 11, 1999), p. 165.

25. Martin E. P. Seligman and Steven F. Maier, "Failure to Escape Traumatic Shock," *Journal of Experimental Psychology*, vol. 74, no. 1 (1967), pp. 1–9.

26. Kane and Staiger, "Improving School Accountability Measures."

27. Sampling error in the context of sampling variance means something different from choosing a survey sample. The authors envision in each district a population of parents who produce children at random times making the mix of children in a given grade different from year to year. The variance is due to both intrafamily differences that are mainly random genetic differences (between siblings) and interfamily differences.

28. Jeremy Finn and C. Achilles, "Tennessee's Class Size Study: Findings, Implications and Misconceptions," *Educational Evaluation and Policy Analysis*, vol. 20, no. 2 (Summer 1999), pp. 97–109; and A. B. Krueger, "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics,* vol. 114 (1999), pp. 497–532.

29. Barbara Nye, Larry V. Hedges, and Spyros Konstantopoulos, "The Long-Term Effects of Small Classes: A Five-Year Follow-up of the Tennessee Class Size Experiment," *Educational Evaluation and Policy Analysis*, vol. 20, no. 2 (Summer 1999), pp. 127–42.

30. David W. Grissmer and others, *Improving Student Achievement: What State NAEP Scores Tell Us,* MR–924–EDU (Santa Monica, Calif.: RAND, 2000).

31. The authors argue that the use of multiple grades often does not reduce volatility as much as one might expect. Their example, however, relates to a different situation than the one discussed here in that it focuses on annual changes in which there are overlapping cohorts of students.

32. For the theory underlying the estimates, see Kane and Staiger, "Improving School Accountability Measures."

33. Coleman and others, *Equality of Educational Opportunity*.

34. Federal policymakers appear to have missed this basic point as is evident from the House and Senate versions of education bills passed during the summer of 2001. Those bills are designed to hold schools accountable for "adequately yearly progress," defined as changes in test scores from one year to the next.

35. Some of these issues are discussed, for example, in Ladd and Clotfelter "Recognizing and Rewarding Success in Public Schools"; Robert H. Meyer, "Comments on Chapters Two, Three, and Four," in Helen F. Ladd, ed., *Holding Schools Accountable: Performance-Based Reform in Education* (Brookings, 1996), pp. 137–45; and Helen F. Ladd and Randall Walsh, "Implementing Value-Added Measrues of School Effectiveness: Getting the Incentives Right," *Economics of Education Review* (forthcoming).

36. Kane and Staiger allude to this fact but do not elaborate. For a description of the state's methodology, see Ladd and Walsh, "Implementing Value-Added Measures of School Effectiveness"; and Helen F. Ladd and Arnaldo Zelli, "School-Based Accountability in North Carolina: The Responses of School Principals," Working Paper (Sanford Institute, 2001).

37. Ladd and Zelli, "School-Based Accountability in North Carolina."

38. Comparisons between some of the 1999 responses and the 1997 responses allowed the researchers to isolate the effects of the accountability system. See Ladd and Zelli, "School-Based Accountability in North Carolina."