Failing to account for natural fluctuations in test scores could undermine the very idea of holding schools accountable for their efforts—or lack thereof

by THOMAS J. KANE, DOUGLAS O. STAIGER, AND JEFFREY GEPPERT

randomly accountable

THE ACCOUNTABILITY DEBATE TENDS TO DEVOLVE INTO A BATTLE between the pro-testing and anti-testing crowds. But when it comes to the design of a school accountability system, the devil is truly in the details. A well-designed accountability plan may go a long way toward giving school personnel the kinds of signals they need to improve performance. However, a poorly designed scheme, which ignores the statistical properties of schools' average test scores, may do more harm than good.

The recent debate over the reauthorization of the federal Elementary and Secondary Education Act (ESEA) is a case in point. From his first days in office, President Bush promised to make education reform a centerpiece of his administration, using the reauthorization of the ESEA as an opportunity to give the state-led accountability movement a dramatic shove forward. Within six months of his taking office, both houses of Congress had passed bills that imposed new federal standards for the states' accountability efforts.

However, both bills were seriously flawed. They created standards that, over time, would have identified nearly every school in the nation as "low performing," forcing them to spend precious resources developing unnecessary school-improvement plans. A tide of paperwork would have crowded out time for learning. This almost turned the most significant federal foray into education policy in decades into an embarrassment. Changes were made by a House-Senate conference committee, so the law, as enacted, remedied the most glaring problems, but created others. The saga illustrates the difficulties of designing an effective accountability system.

PHOTOGRAPH BY TINA WEST/WWW.IMAGES.COM

The House and Senate Bills

At the heart of both bills was a detailed formula for determining when a school is making "adequate yearly progress." The consequences for schools that failed to meet their performance targets were progressively severe—after one year, districts would be required to offer public school choice to all the students in a school; after several years, districts would be required to replace school staff, convert the school into a public charter school, or hand the school over to a private contractor.

Single-year changes in test performance are very unreliable indicators of where a school is headed over the long term.

The problem is that such consequences place too much weight on single-year changes in test scores at the school level. Either bill would have required an increase in the proportion of students scoring above the proficient level in both math and reading, each and every year. However, test scores at the school level often fluctuate for reasons other than any underlying change in a school's performance. Such volatility arises from two sources. The first is variation due to differences in the groups of students being tested each year. Even if the students are being drawn from the same families and the same neighborhoods, the average performance of a school can fluctuate from year to year depending on the attitudes and abilities of the students in each cohort. The average elementary school contains only 68 students per grade level. With a sample this small, having five particularly bright students (or a few students with undiagnosed learning disabilities) in any one year can lead to large fluctuations in a school's test scores from one year to the next. The Department of Labor measures the monthly unemployment rate with a sample of nearly 60,000 households. Congress was proposing that the Department of Education measure the performance of the typical elementary-school grade with a sample nearly 1/1000 the size.

The second source of variation is one-time factors that lead to temporary fluctuations in test performance. Some of these factors are likely to be unrelated to the educational practices of a school. For instance, a dog barking on the day of the test, a severe flu season, or one particularly disruptive student in class could cause scores to fluctuate. There may be other sources of volatility that are more related to the educational mission of a school, such as the favorable chemistry between a teacher and a particular group of students or teacher turnover. Whatever the source of variation, single-year changes in test performance are very unreliable indicators of where a school is headed over the long term.

Consider the examples of North Carolina and Texas. Between 1994 and 1999, these states were the educational envy of the nation, raising proficiency rates in math and reading by 2 to 5 percentage points in the average year. However, the vast majority of schools in those states exhibited much less consistent progress: less than 2 percent of schools witnessed an increase in math and reading proficiency each and every year for those five years. Indeed, we estimate that between 98 and 100 percent of the elementary schools in North Carolina and Texas would have failed the House and Senate's initial definitions of annual yearly progress at least once between 1994 and 1999.

Furthermore, both bills would have compounded the error by requiring annual increases in test scores for every racial subgroup in a school. The intent was admirable: to ensure that schools do not ignore minority children. But this provision was likely to have harmed its intended beneficiaries, by arbitrarily sanctioning schools that enroll students from several different racial or ethnic subgroups. Suppose that a school is solidly on the path to improvement, with a 70 percent chance of increasing the proficiency of any racial subgroup in a given year. A school with two racial subgroups in its student body would have a less than 50-50 chance of achieving an increase for both groups in a given year—because the year-to-year fluctuations are nearly independent for each racial group (therefore the probability is .70 times .70, or .49). The odds would be even longer for a school with three racial subgroups (.70 times .70 times .70, or .34). Since African-American and Latino students are more likely to attend schools with more than one racial group, they are more likely to see their education disrupted arbitrarily.

A number of states have established accountability programs that track the performance of racial and ethnic subgroups separately. For example, California requires schools to meet certain growth targets for all "numerically significant" subgroups in a school. In order to be numerically significant, a group must either represent at least 15 percent of the student body and have more than 30 students or have more than 100 students regardless of what percentage they are. There are eight different groups that can qualify as numerically significant, depending on 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.

We calculated the likelihood of a California school's winning a Governor's Performance Award by size and by the number of numerically significant subgroups. Among the smallest quintile of elementary schools, 47 percent of racially heterogeneous schools (those with four or more racial subgroups) won performance awards, versus 82 percent of similarly sized but racially homogeneous schools. This is particularly ironic given the fact that overall growth in performance was slightly higher for more-integrated schools between 1999 and 2000. Moreover, the reason for a school's failure to win an award was often not that African-American and Latino students were lagging behind,

but that white non-Hispanic students experienced slower growth in achievement: the average school with multiple racial subgroups witnessed larger gains for African-American and Latino students than for white students.

Separate achievement targets for racial and ethnic subgroups seem to be neither necessary nor especially effective in coaxing schools to focus on the performance of racial and ethnic minorities. In North Carolina, where there are no separate racial targets, African-American and Latino students experienced slightly higher improvements in proficiency than white non-Hispanic youth. Until this year, the rating system in Texas specified separate targets for racial subgroups that accounted for more than 10 percent of the student body (and more than 30 students). However, African-American and Latino students saw the same improvements in their test scores whether or not they attended schools with enough minority students to require a separate racial target.

Remaining Problems

The conference committee's compromise bill remedied some problems but created new ones. Earlier versions of the legislation rated schools according to their year-to-year improvements in the share of their students who achieve a certain proficiency level. Now schools will simply need to have a certain minimum percentage of their students (and of racial, ethnic, and socioeconomic subgroups within each school) deemed "proficient" each year. The initial minimum proficiency rate will be the greater of the proficiency rate of the 20th-percentile school or the average statewide proficiency rate of the lowest-scoring subgroup. In many states, the effective minimum will be the proficiency rate of the 20th-percentile school. However, more than 20 percent of all schools are likely to fail, because the threshold will apply not only to the school as a whole but also to all the subgroups in a school. If any racial, ethnic, or socioeconomic subgroup within a school fails, the school fails. As a result, a disproportionate number of the schools that enroll disadvantaged minority subgroups are likely to fail. The minimum proficiency rate that schools are required to meet will be raised gradually to 100 percent over the next 12 years.

The main beneficiaries of the conference committee's changes will be suburban schools whose initial rates of proficiency are above the minimum, for they will no longer be penalized for temporary downward fluctuations in scores. The primary losers will be schools with initially low levels of proficiency for any subgroup. They will now be required to achieve a 10-percentage-point increase in proficiency for those subgroups to avoid the sanctions in a given year.

One flaw in the new formula is that it provides a strong incentive for states to lower the score students must exceed on their state tests in order to achieve "proficiency." The problem is that redefining proficiency simply because of the new federal requirements may create a credibility problem for the standards movement in a number of states.

Another problem not remedied in the final bill is that any federal definition of adequate yearly progress is likely to conflict with at least one of the state accountability plans that are already in place. There are three common variants in state accountability systems: some states, such as North Carolina, Arizona, and Tennessee, rate their schools with a measure of a school's value-added, using the growth in performance for a given group of students since the end of the preceding school year; other states, such as Texas and Illinois, rate their schools on the percentage of students scoring above certain thresholds; still other states, such as California, rate their schools based on their change in test scores from one year to the next. (A fourth category of states rates schools based on some mixture of value-added, levels, or changes.) Thus states that have been rewarding schools based on value-added measures or on changes in scores may be required to sanction the very schools they have been rewarding.

The next battleground is likely to be the issue of how many students it takes to create a separate racial, ethnic, or socioeconomic subgroup for accountability purposes. The legislation only requires that there be a sufficient number of students to yield statistically reliable information in order for the subgroup to count separately. The higher the threshold—say, requiring a subgroup to represent at least 15 percent of the student body, as opposed to 5 or 10 percent—the lower the failure rate will be for schools with small percentages of disadvantaged minority students.

Design Principles

State and federal officials ought to keep three basic principles in mind in designing test-based accountability systems:

- *Multiple years of data are required to measure improvements in performance reliably.*

Children arrive at school with widely varying levels of preparation. Even a mediocre school can expect high test scores if its students come from wealthy backgrounds. As a result, policymakers in many states have attempted to level the playing field by focusing on improvements in test scores. However, improvements are very difficult to discern with a couple of years' worth of data, for two reasons. First, schools differ much less in the extent to which they improve test scores from year to year than they do in their beginning level of performance. Second, any measure of change in performance is likely to amplify the effect of sampling variation and other one-time factors that lead to fluctuations in performance. In other words, identifying improvements in performance as opposed to levels of performance in a single year is like looking for a smaller needle in a bigger haystack. If policymakers intend to measure and reward improvements in test performance at the school level, they will need to rely on multiple years of data.

Improvements can be measured in two basic ways: the

improvement in performance for a given group of students from one year to the next (known as a value-added approach), or the improvement in performance across different groups of students (which we will refer to as cross-cohort changes). The improvement of scores for at least two contiguous grades (for example, grades 4 and 5) from one year to the next is a mixture of value-added changes (the 4th grade students who become 5th graders) and cross-cohort changes (the 4th grade students this year are a different group from the 4th grade students last year).

Kane and Staiger have analyzed the statistical properties of value-added and cross-cohort changes in test scores, using data from North Carolina (see Figure 1). (Full citations are available at www.educationnext.org.) We measured value-added with the average change in combined reading and math scores for a school's students between the end of 3rd grade and the end of 4th grade; we measured cross-cohort changes with the change in 4th grade scores from one year to the next. Among median-size schools in North Carolina, roughly half of the variance between schools in value-added in 4th grade math and reading was due to sampling variation and other one-time factors. For the smallest quintile of schools, the percentage of variance due to non-persistent factors was even higher (58 percent), while for the largest quintile of schools the percentage was somewhat lower (29 percent). Cross-cohort changes in mean test scores from one year to the next were measured even more unreliably. More than three-quarters of the variance in the annual change in mean test scores among the smallest quintile of schools was due to one-time, non-persistent factors. This percentage was only slightly smaller (73 percent) for the largest quintile of schools. Such volatility can wreak havoc when rewards and punishments are doled out on the basis of changes in test scores; school personnel are at risk of being punished or rewarded for results that are beyond their control.

Therefore, when policymakers seek to reward schools for improvements in test scores, they should do so based on multiple years rather than a single year of data. Moreover, while a simple arithmetic average of improvements over multiple years would be an improvement, there are even more efficient ways to pool information over time. For instance, building on work by McClellan and Staiger (1999) in rating hospital performance, we have proposed a simple technique for pooling information over time, which improves on a simple arithmetic mean by taking into account the amount of "signal" and "noise" in a given measure of performance. For instance, for large schools, for which we would expect less noise in any given year's measure, the proposed method would place more weight on more recent scores; for small schools, the method would place more equal weights on each of several years' worth of scores.

With test scores being so volatile, school personnel are at a substantial risk of being punished or rewarded for results that are beyond their control.

- *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—affect primarily small schools and provide very weak incentives for large schools.*

Each year since 1997, North Carolina has recognized the 25 elementary and middle schools in the state with the highest scores on the "growth composite," a measure reflecting the average gain in performance among students enrolled at a school. Winning schools receive financial awards.

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

Of the 840 elementary schools we analyzed, 59 were among the top 25 at some point between 1997 and 2000 (the top 25 each year included middle schools, which we are not analyzing here). Among all the schools, the average gain score was not strongly related to school size, but 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 compared with .011). As a result, schools in the smallest decile were much more likely to be among the top 25 schools at some point over the period: Even though their mean gains were not statistically different, the smallest schools were 23 times more likely to win a top-25 award than the largest schools.

For the very same reason, small schools are also overrepresented among those with extremely low test scores. Beginning in 1997, the state assigned assistance teams to intervene in schools that performed poorly on state tests and failed to meet their growth targets from the previous year. All but one of the elementary schools assigned an assistance team were among the smallest 40 percent of schools. (The smallest decile of schools would have received an even larger share of the assistance teams, except for a rule requiring that the proportion of students scoring below grade level be statistically significantly less than 50 percent.)

This year, the state of California distributed $100 million to teachers in schools that started with test scores in the bottom half of schools in 1999 and achieved large gains in performance between 1999 and 2000. A thousand teachers in schools with the largest improvements received $25,000 bonuses on average.

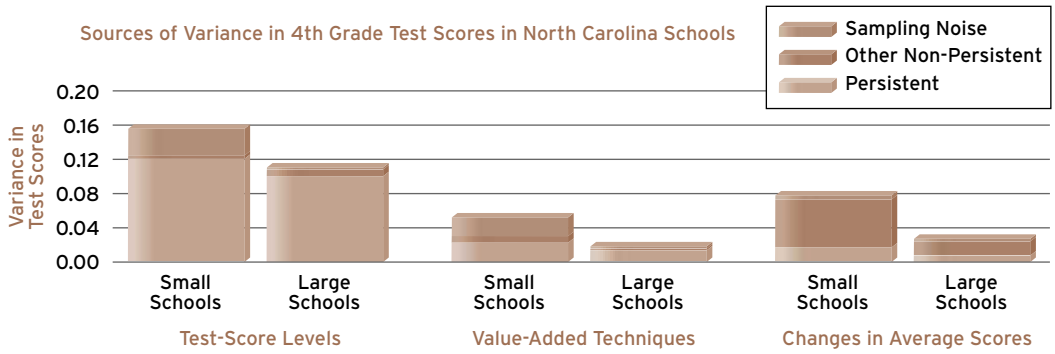
Small schools in California were considerably more likely to win one of these awards than were larger schools. Given the importance of sampling variation and the fact that the largest bonuses were reserved for teachers in schools with the most extreme increases in test scores, this is hardly a surprise.

A threshold at either extreme is likely to be irrelevant for large schools, since they are unlikely to experience such large swings in performance regardless of their efforts. If the marginal costs of improving are also higher at large schools, the problem of weak incentives for large schools would only be compounded. A remedy would be to establish different thresholds for schools of different sizes. 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 would 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 disparity for large and small schools is less extreme.

Helen Ladd and Charles Clotfelter in 1996 and David Grissmer et al. in 2000 reported evidence suggesting that schools respond to incentives by raising student performance. However, the long-term effects of incentives may be quite different from their short-term effects. Even if teachers are not sufficiently aware of the statistical forces at work to recognize their rather limited influence on test scores in the short run, they may well become aware of this 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 negative opinions regarding the value of their efforts.

Value-Added and Change Scores Are Less Reliable (Figure 1)

Of three ways of measuring a school's performance—the level of its test scores, value-added techniques, and changes in the average level—test-score levels are the most reliable. This is because schools differ more in their level of performance than they do in their relative rates of change, and a smaller portion of differences in levels is due to sampling noise or other non-persistent factors, such as a disruption on the day of the test. Reliability problems are most acute for small schools because of the smaller number of observations.



SOURCE: Thomas J. Kane and Douglas O. Staiger, 2001

• *When evaluating the impact of policies on changes in test scores over time, the natural fluctuations in test scores must be accounted for.*

In 1997, North Carolina identified 15 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 "low performing." The state Department of Public Instruction ascribed the improvements to the efforts of the assistance teams; the assistance teams were lauded in *Education Week*'s annual summary of the progress of school reform efforts in the states as well. However, given the amount of sampling variation and other non-persistent fluctuations in test-score levels and gains, schools with particularly low test scores in one year would be expected to bounce back in subsequent years.

The schools that were assigned assistance teams seem to have had a particularly bad year the year they received the sanction. In the year before assignment, such schools had an average 4th grade combined reading and math test score that was .67 student-level standard deviations below the average school. This reveals that they were weak schools the year before being sanctioned. However, in the year of assignment, their average score was even lower, .79 student-level standard deviations below the average school. The year after assignment, their scores seemed to rebound to .52 student standard deviations below the mean. One is likely to greatly overestimate the impact of assistance teams by taking the change in performance in the year after assignment.

There are real differences in performance at the school level. And schools that are not improving should be identified for intervention. However, one year's worth of test-score data is insufficient to discern such differences in a meaningful way. States should be allowed to experiment until the nation finds the ideal way to determine which schools are making adequate yearly progress. We understand the impulse to create a system that requires specific remedies sooner rather than later. However, impatience is an insufficient excuse for bad education policy.

–Thomas J. Kane is a professor of policy studies and economics at the School of Public Policy and Social Research at the University of California at Los Angeles. Douglas O. Staiger is an associate professor of economics at Dartmouth College. Jeffrey Geppert is a senior research analyst at the National Bureau of Economic Research in Palo Alto, California.
A portion of this article is drawn from a chapter that will appear in the Brookings Papers on Education Policy 2002 *(Brookings, 2002).*