Discussion

Improving Educational Quality: How Best to Evaluate Our Schools?

Thomas J. Kane*

While the academic debate has been preoccupied for much of the last decade with school vouchers, state policymakers have been moving in a very different direction, constructing elaborate incentive systems using school-level test-score measures. For instance, California spent nearly $700 million on school-level incentives in 2001, providing bonuses of up to $25,000 per teacher in schools with the largest increases in test performance between 1999 and 2000. Unfortunately, given that the discipline of economics has a long tradition of thinking about the design of incentives, economists have been largely absent from the debate accompanying the design of school accountability systems. For anyone seeking to catch up with the policy debate, the Hanushek and Raymond paper is extremely useful in categorizing the types of systems that have been created, in summarizing the fledgling literature on the impact of school accountability systems within states, and in providing some original evidence on the impact of state accountability policies using the National Assessment of Educational Progress (NAEP) test scores across states over time.

One of the great contributions of the paper is that it simply provides a clearer picture of the variety of systems that have been created. As described by the authors, most systems use a hybrid of one of three types of measures of test performance: status measures (mean levels of test performance), status-change measures (changes in the level of performance between cohorts over time), and gain score (the mean improvement in performance for a given cohort of students). Possibly because of

* Professor of Policy Studies and Economics at the University of California, Los Angeles.

the difficulty of tracking individual students' performance over time, most states have chosen to base their systems on either status measures or status-change measures.

Lessons from Optimal Incentives Literature in Economics

A number of potential lessons can be learned from the optimal incentives literature in economics. First, as is hinted by the authors, incentive systems based upon status-change measures inevitably are subject to "ratchet effects." Raising the bar in the future based upon performance today forces schools to choose between the payoff of improvements today and the increased cost of maintaining that level of performance in the future. It is particularly striking when Hanushek and Raymond point out that evaluations based on changes in performance are a component in most state accountability systems. When performance today has an impact on expectations tomorrow, schools may underinvest in reform. (This is particularly true in systems measuring status change for single grade levels, since those using multiple grade levels may continue to benefit from any pedagogical improvement for several years as a given cohort of students moves through several grade levels.) The "ratcheting" problem is exacerbated by the fact that rewards are usually discontinuous, stair-shaped functions of performance—meaning that the magnitude of one's reward is not a function of the distance by which a school might clear a given threshold. The authors note that in a system based upon status changes, a school may generate one-time improvements in performance by limiting the population of test takers, but will not necessarily increase its likelihood of success in future years. But the same may be true of many other worthwhile pedagogical reforms. (Consider what would happen if academics were rewarded based upon the increased number of articles published from one year to the next, rather than some average of the stock of accumulated work and the average output per year over their careers.)

Second, we know from the optimal-incentives literature summarized in Lazear (1995) and Milgrom and Roberts (1992) that imperfect measures of performance should receive less weight in an incentive framework. Test-score measures are imperfect measures of schools' output for at least four reasons.

First, test-score measures often include systematic, predictable factors that are outside schools' control. The easiest example of these factors is family background. Placing too great an implicit weight on family background and other factors affecting students' baseline performance encourages schools to exempt students from their testing programs. One partial solution to this problem is to focus on gain scores or value-added measures of achievement (it is only a partial solution since some students

not only start out with a lower baseline, but they may have a predictably flatter or steeper trajectory as well).

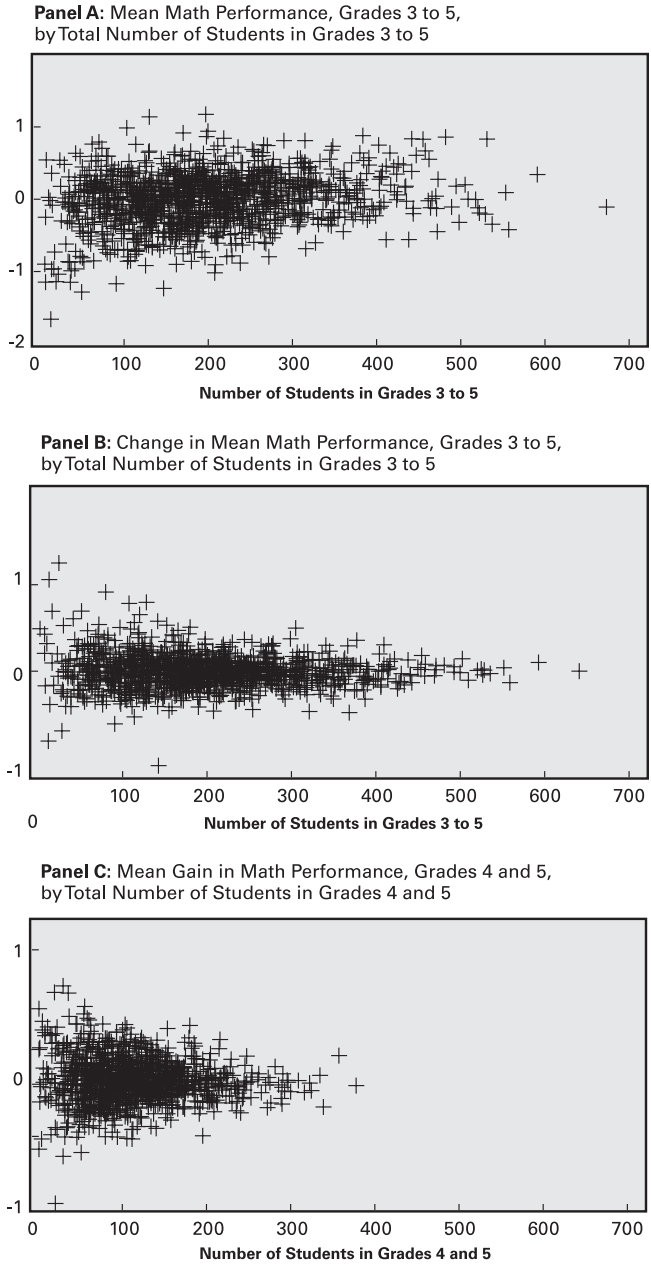
Second, as the authors note, the typical test-based measures are incomplete measures of school output. For example, most test-based accountability systems are based upon reading and math scores alone. As critics are wont to point out, civics and social tolerance are typically assigned zero value. However, it is also worthwhile to note that many "hard" skills—such as science, history, and social studies—are also excluded. The new federal No Child Left Behind Act of 2001 requires states to test reading and math skills in grades three through eight by the 2005–06 school year. Science tests will not be added until 2007–08. There are no plans to require states to test other skills.

Placing too great a weight on the measured outputs is likely to lead schools to substitute away from other valued, but difficult-to-measure domains. Whether intended or not, such rules are likely to tip the balance of instruction toward the subset of subject areas and concepts that are tested. For example, in the Kentucky accountability system in the early 1990s, science was tested in fourth grade and math was tested in fifth grade. Stecher and Barron (1999) found that teachers had reallocated their time so that they spent more time on science in fourth grade when students took the science test and more time on math in fifth grade when students took the math test. Jacob (2002) found that scores on science and social studies leveled off or declined in Chicago after the introduction of an accountability system that focused on math and reading performance.

Third, school-level test scores are also imprecise measures of the domains they are intended to measure. This fact is highlighted by Figure 1, which reports the distribution of different types of measures by school size, taken from a North Carolina sample. Panel A reports data on mean math performance in grades three through five by school size; Panel B reports data on changes in mean performance in grades three through five by school size; the final panel reports mean gains in performance at the individual student level in grades four and five. As is evident in the funnel-shaped patterns for all three distributions in Figure 1, one important source of imprecision is simple sampling variation. Given that the typical elementary school contains 60 students per grade level, a few particularly bright or particularly rowdy students can have a big impact on scores from year to year. Aggregating across several grades helps, but obviously does not eliminate this problem. Moreover, sampling variation appears to account for a larger share of the total variance for the change in performance from year to year and for the mean cohort gain across different schools than for levels.

Fourth, in addition to sampling variation, there is evidence of other one-time shocks to school performance (Kane and Staiger 2002a). These shocks may be due to other sampling-related causes—such as peer effects, testing artifacts generated by changes in test forms, school-wide distur-

Figure 1
Distribution of Levels, Changes, and Cohort Gains, by School Size



Note: Test-score measures on the vertical axes are standardized using student-level standard deviation and mean.
Source: Author's analysis of 1998 and 1999 North Carolina test-score data.

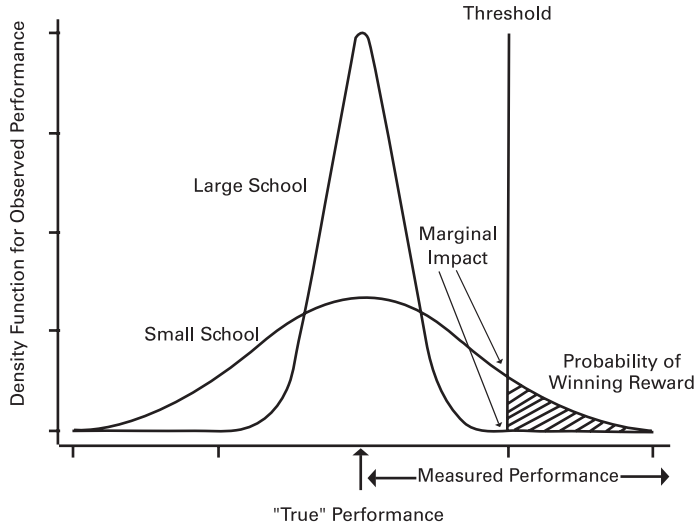
bances such as a dog barking in the parking lot on the day of the test, or other short-term impacts such as classroom chemistry. The pattern of such shocks suggests that there is a weak correlation in performance between test scores one year apart, but that correlation fades only gradually after that one year. Such one-time shocks are unlikely to be due to teacher turnover, since teacher turnover follows a very different pattern. After one year, about 20 percent of teachers turned over in the typical elementary school in North Carolina. However, after five years, about 50 percent of the teachers in a school had turned over. Teacher turnover may explain the pattern of declining correlation in the second year and beyond, but it cannot explain the dramatic fall-off in the first year.

Kane and Staiger (2002a) performed further analysis of the sources of variance in test scores in North Carolina, including decomposing the variance in status measures ("levels"), status-change measures ("changes"), and cohort-gain measures ("gains") into three parts: a persistent component, sampling variation, and other one-time shocks. They report four results worth noting: First, the between-school variance in mean test performance is small relative to the total variance in performance at the student level. Even including the effect of sampling variation, the between-school variance accounted for only 10 percent to 20 percent of the total variance in test scores. Despite the fact that there may be some very high-scoring schools and some very low-scoring schools, the differences in performance for students within the typical school tend to be much larger than the differences between schools.

Second, much of the difference in the test-score levels is persistent. Even among the smallest quintile of schools, nonpersistent factors account for only 27 percent of the variance between schools. Among the largest quintile of schools, such factors account for only 13 percent of the variance. However, since we are not adjusting for initial performance levels or for the demographic characteristics of the students, much of that reliability may be due to the unchanging characteristics of the populations feeding those schools and not necessarily from unchanging differences in school performance.

Third, although one might be tempted to rate schools by their improvement in performance or by the average increase in student performance over the course of a grade, such attributes are measured remarkably unreliably. More than half (56 percent) of the variance among the smallest quintile of schools in mean gain scores is due to sampling variation and other nonpersistent factors. Even among the largest quintile of schools, nonpersistent factors are estimated to account for 34 percent of the variance in gain scores. Changes in mean test scores from one year to the next are measured even more unreliably. More than 80 percent of the variance in the annual change in mean test scores among the smallest quintile of schools is due to one-time, nonpersistent factors.

Figure 2
Hypothetical Example: Marginal Payoff to Improving Performance for Small School and Large School



Fourth, 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 caused by nonpersistent factors was still over 60 percent.

Kane and Staiger (2002a) develop several implications of such imprecision for the design of accountability systems. Figure 2 illustrates the impact of imprecision in test-score measures on schools' incentives. Suppose a small school and a large school have the same expected performance next year. Each has a range of expected outcomes. Suppose that only those schools with scores above a threshold will win an award. The marginal incentive for each school is measured by the height of the density function where it crosses the incentive. For thresholds up at the extremes, more randomness can actually increase the strength of incentives. In this picture, when the threshold is at either extreme, small schools have a positive incentive to improve, while large schools have very little incentive. When the threshold is in the middle of the distribution, small schools with a greater variance in likely scores have the weaker incentive.

How Much Would an Increase in Performance Be Worth?

Critics of school accountability worry that current systems already place too great a weight on imperfect measures of academic achievement and, on net, may do more harm than good. To evaluate these concerns, one must have a sense of the potential value that we should place on an increase in student achievement.

Some simple calculations by Kane and Staiger (2002b) reveal that the monetary value of even a small improvement in academic achievement can have very large payoffs. Two recent papers provide estimates of the impact of test performance on the hourly wages of young workers. Murnane, Willett, and Levy (1995) estimate that a one-standard-deviation difference in math test performance is associated with an 8.0 percent hourly wage increase for men and a 12.6 percent increase for women. These estimates probably understate the value of test performance, since the authors also control for years of schooling completed. Neal and Johnson (1996), who do not condition on educational attainment, estimate that a one-standard-deviation improvement in test performance is associated with hourly wage increases of 18.7 percent for men and 25.6 percent for women. Using a discount rate of 6 percent, the present value at age 18 of a one-standard-deviation difference in test performance is worth roughly $62,000 per student using the Murnane, Willett, and Levy estimates and $146,000 per student using the higher estimates from Neal and Johnson.[1] Discounting these values back to age 9 (for example, fourth grade) would reduce the estimates to $40,000 and $94,000 per student.

Such estimates are quite large relative to the rewards offered to schools for increasing student test performance. For example, California paid elementary schools and their teachers an average award of $122 per student if their school improved student performance by an average of at least 0.03 student-level standard deviation.[2] Based on the calculations in

[1] I used the following calculation:

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

where β is the proportional rise in wages associated with a given test-score increase; w_i represents wages from age 18 through 64 estimated using full-time, year-round workers in the 2000 Current Population Survey; γ represents the general level of productivity growth, assumed to equal 0.01; and r is the discount rate, assumed to equal 0.06.

[2] The School Site Employee Bonus program provided $591 per full-time equivalent teacher to both the school and teacher, or $59 per student based on an average of 20 students per teacher. The Governor's Performance Award (GPA) program provided an additional $63 per student. The growth target for the average elementary school was 9 points on the state's academic performance index (API). Because the state did not publish a student-level

the preceding paragraph, the present value of such an increase in test scores to students in elementary school would be in the range of $1,200 to $2,800 per student (0.03 times $40,000, or $94,000), much more than the $122 paid by the state. In other words, the labor-market value of the test-score increase would have been worth roughly 10 times to 20 times the value of the incentive provided in 2001 by California—the state with the most aggressive financial incentive strategy in that year. (Budget cuts have subsequently led to declines in those incentive payments.)

This calculation suggests that even the most aggressive state is paying schools much less than the marginal payoff if we thought the test-score improvements reflected true achievement. Critics' concerns about relying on imperfect performance measures may already be reflected in small incentive payments. In fact, the strength of incentives for schools in California is similar to what Hall and Liebman (1998) found for CEOs: $1 in compensation for every $40 increase in firm valuation.

Empirical Estimates of the Impact of Accountability Incentives on Test Performance

The most intriguing part of the Hanushek–Raymond paper studies the relationship between state differences in the timing of adoption of test-based accountability and state performance on the NAEP. States with an accountability system in place in 2000 had achievement growth approximately 1 percent larger than states without such systems. The number of years a state had such a system in place was not related to NAEP performance.

Hanushek and Raymond report the impact in log units, not in student-level standard deviation units. A few simple calculations suggest that the impact was fairly modest when translated into standard deviation units. Between 1996 and 2000, the average growth in achievement on the state assessments between fourth and eighth grade was 52 points (from 222 to 274). A 1 percent difference, therefore, would represent a 0.5 point increase. The standard deviation in achievement in fourth or eighth grade was approximately 32 points. Therefore, a 1 percent improvement in the growth in performance from fourth to eighth grade would

---

standard deviation in the API scores, we had to infer it. A school's API score was a weighted average of the proportion of students in each quintile of the national distribution on the reading, math, language, and spelling sections of the Stanford 9 test. For elementary schools, the average proportion of students across the four tests in each quintile (from lowest to highest) was 0.257, 0.204, 0.166, 0.179, and 0.194, and the scores given to each quintile were 200, 500, 700, 875, and 1,000. Under the assumption that students scored in same quintile on all four tests, we could calculate the student-level variance as $0.257\ (200 - 620)^2 + 0.204\ (500 - 620)^2 + 0.166\ (700 - 620)^2 + 0.179\ (875 - 620)^2 + 0.194\ (1{,}000 - 620)^2 = 89{,}034$, implying a standard deviation of 298. This is nearly five times the school-level variance, which is roughly consistent with expectations.

represent a 0.016 student-level standard deviation improvement in performance for the average student. However, given the estimates above, even a small increase in performance may well be worthwhile. For elementary school students, a 0.016 student-level standard deviation increase would be worth $640 to $1,500 per student. Most states are spending much less than that on their accountability systems. Therefore, a more thorough cost-benefit analysis may yield quite large payoffs to creating an accountability system.

As Hanushek and Raymond acknowledge, accountability systems are weakened if an increasing number of students are excluded from taking the exams. We should be cautious in using the NAEP tests to study the impact of state accountability systems because there have been large increases over time in the proportion of students excluded from the state NAEP samples.[3] The NAEP test has traditionally excluded the test scores of students to whom the states have granted testing accommodations—such as allowing a longer time to take the test, having the questions read aloud, or having the test translated into a native language. The idea was to compare students in the same testing conditions. However, after the passage of the Individuals with Disabilities Act in 1996, many states began granting accommodations to a larger share of students. (There may be more nefarious reasons as well; now that NAEP scores are given a much higher profile, states have a stronger incentive to inflate their scores by excluding students.)

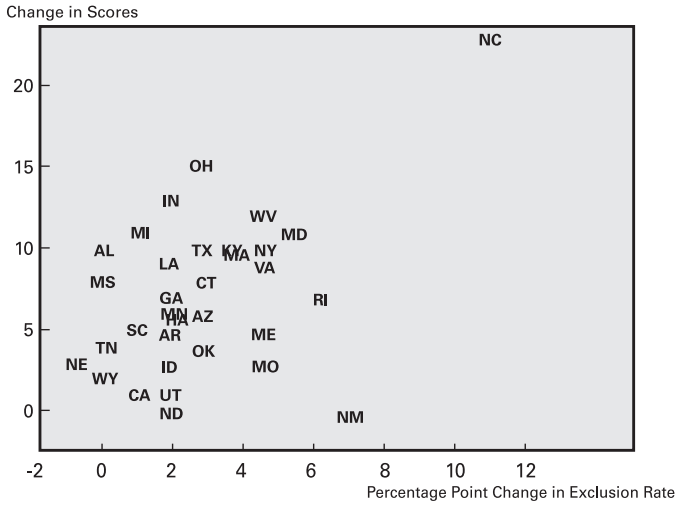
Moreover, the increases in exclusion rates seem to be particularly large in states that have been touted for their increases in NAEP performance. Figure 3 reports the change in eighth-grade math NAEP scores between 1992 and 2000 and the changes in the proportion of the sample excluded from the assessments. Between 1992 and 2000, the average state increased its exclusions by 3.5 percentage points, from 5 percent of sampled youth to 8.5 percent of sampled youth. One state, often cited as having an exemplary accountability system, North Carolina, increased its eighth-grade exclusion rate by 11 percentage points, more than in any other state.[4]

Because of this data problem, we may never be able to go back and assess the considerable experimentation with accountability initiatives that occurred in many states during the 1990s. Beginning with the 2000

[3] Grissmer and Flanagan (2002) note the same phenomenon.

[4] It is unlikely that the change in exclusion rates accounts for all of the change in North Carolina. The exclusion rates in fourth grade and in eighth grade together increased an average of 10 percentage points. If the distribution of test scores is normal at the student level, then raising the truncation point from the 3rd percentile to the 13th percentile would have raised test scores by only 0.17 standard deviation—much less than the observed increase in North Carolina. This is an extreme assumption since not all of the nontested students would have been in the bottom tail, so that the actual effect on NAEP scores is probably smaller.

Figure 3
Change in Eighth Grade NAEP Math Scores and in NAEP Math Exclusion Rate between 1992 and 2000



Source: U.S. Department of Education (2001), Tables A.7b and B.7.

assessment, the NAEP began reporting state-level results for the "no accommodations" sample as well as for a sample including the students with accommodations. In the future, then, it may be easier to track differences in improvements at the state level—although there will still be a tricky problem created by the fact that different states will continue to grant accommodations to different shares of their students.

Conclusion

Hanushek and Raymond provide an extremely useful description of state accountability schemes and review the developing literature on the impact of test-based accountability on academic achievement. Their analysis of the growth in NAEP scores in states with and without accountability systems suggests small, positive impacts on student performance. It is worthwhile noting that even a small increase in student performance would generate sufficient benefits to cover the moderate cost of operating an accountability system, given the value of academic achievement to students later in life.

Given the range of strategies used in different states—some states reward test-score levels, while other states reward changes in test scores, while still other states focus on cohort-gain scores—it is clear that we have a lot to learn about the relative payoffs of different approaches.

Therefore, it is unfortunate that the No Child Left Behind Act of 2001 imposes a new federal system of accountability that will inevitably conflict with many of the state rating systems in use. Under that system, states are allowed to define proficiency in whatever manner they choose. But once a state has defined proficiency, the minimum proficiency rate for all schools and for all racial and ethnic subgroups within schools will be equal to the proficiency rate of the 20th percentile school. Because the federal system will be based on status measures (or levels), while many states use status changes or gain scores, there will be many cases where schools are failing the federal definition while doing well using their state's metric. Many schools that fare well under California's system based on changes in test performance or under North Carolina's system using cohort-gain scores, even many of those achieving exemplary rankings, will be sanctioned under the new federal law. It remains to be seen whether the mixed signals created when the new federal accountability system is laid on top of state accountability systems will simply confuse schools and parents or whether it will spur them on to further improvements.

References

Grissmer, David and Ann Flanagan. 2002. "Tracking the Improvement in State Achievement Using NAEP Data." RAND Corporation, unpublished paper.

Hall, Brian J. and Jeffrey B. Liebman. 1998. "Are CEO's Really Paid Like Bureaucrats?" *Quarterly Journal of Economics* 113 (3): 653–91.

Jacob, Brian A. 2002. "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools." NBER Working Paper No. 8968 (May).

Kane, Thomas J. and Douglas O. Staiger. 2002a. "Volatility in School Test Scores: Implications for Test-Based Accountability Systems." In *Brookings Papers on Education Policy, 2002*, edited by D. Ravitch. Washington, DC: Brookings Institution.

———. 2002b. "The Promise and the Pitfalls of Using Imprecise School Accountability Measures." *Journal of Economic Perspectives* 16 (4): 91–114.

Lazear, Edward P. 1995. *Personnel Economics*. Cambridge, MA: MIT Press.

Milgrom, Paul and John Roberts. 1992. *Economics, Organization and Management*. Englewood Cliffs, NJ: Prentice Hall.

Murnane, Richard J., John B. Willett, and Frank Levy. 1995. "The Growing Importance of Cognitive Skills in Wage Determination." *Review of Economics and Statistics* 77 (2): 251–66.

Neal, Derek and William Johnson. 1996. "The Role of Premarket Factors in Black–White Wage Differentials." *Journal of Political Economy* 104 (5): 869–95.

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