

What Test Scores Can and Cannot Tell Us About the Quality of Our Schools

BY THEODORE M. CRONE

How to best judge the quality of our schools is a thorny issue. Now, the No Child Left Behind Act, which was signed into law in January 2002, mandates standardized testing in math and reading for students in grades three through eight. The test scores will then be used both to gauge the students' level of proficiency in these subjects and to evaluate the schools' performance. But emphasizing test scores as a measurement of the quality of schools raises several questions. In this article, Ted Crone looks at some of these questions and warns us to be cautious in how we use test scores.

On January 8, 2002, President Bush signed into law the No Child Left Behind Act (NCLB), the latest reauthorization of the Elementary and Secondary Education Act. When fully implemented, the new law will require that students in grades three through eight take statewide standardized tests every year in math and reading. The scores on these tests will be used to determine whether students have achieved the required level of proficiency in these subjects for their grade level. Schools will be evaluated and

rewarded or penalized on the basis of the test results. Since states are obligated under the law to release annual report cards on the schools, the general public is also likely to view these test scores as the primary measure of school quality.

This increased emphasis on standardized test scores as a measure of school quality and a tool for accountability raises the issue of what test scores can and cannot tell us about the quality of our schools. Should we be looking at average test scores or changes in test scores as the measure of quality? How much of a difference in either of these measures is significant? Finally, how can we distinguish the school's contribution to these test scores from the effects of the students' innate abilities, their family, social, and economic backgrounds, and the

abilities and backgrounds of their peers in the classroom?

LARGE-SCALE TESTING IS NOT NEW TO THE U.S. EDUCATION SYSTEM

The beginning of large-scale external testing in the U.S., that is, tests developed outside the schools in which they are used, goes back to the mid-19th century.¹ Initially, such testing was limited. But the use of standardized tests increased significantly in the two decades after the development and publication of the Stanford Achievement Test in 1923. Between World War II and the 1960s, standardized tests were primarily used to evaluate students and curricula; they were not commonly used to hold schools accountable for student performance. Except for tests such as the SAT, which is used for college admissions, there were few direct consequences for the students or the schools associated with the scores on standardized tests.

In the 1960s two new programs at the national level expanded the role for large-scale testing. Title I of the Elementary and Secondary Education Act, enacted in 1965, provided federal funds to schools with a large percentage of low-income students. The act required the periodic testing of students in the program to assess its effectiveness. Also, in the late 1960s, the Education Commission of the States sponsored the first set of tests



Ted Crone is a vice president and head of the regional and urban economics section in the Research Department of the Philadelphia Fed.

¹ For a brief history of large-scale testing in the U.S., see the report from the U.S. Congress, Office of Technology Assessment. See also the article by Laura Hamilton and Daniel Koretz, and the two articles by Koretz.

known as the National Assessment of Educational Progress (NAEP). These tests have been administered periodically since 1969 to a random sample of 9-, 13-, and 17-year-olds in reading, math, and science to measure progress over time. A parallel set of NAEP tests was developed in the 1980s to be given in specific grades rather than to students based on their age. Besides being given to a national sample, this set of tests is given every two years to a sample of fourth and eighth graders in participating states and provides a basis of comparison among the states.

At the state level, large-scale standardized testing took on a new role in the 1970s. Many states adopted minimum-competency testing as a

holding students accountable; as such, they require that each student take the test and that a cutoff score be established to determine who meets the minimum competency level. Results from these types of tests are likely always to be considered the best measure of academic competency for primary and secondary students.

The role of large-scale standardized tests expanded again in the 1980s and 1990s. Besides using standardized tests to hold *students* accountable, states began to use them to hold *schools* accountable, rewarding or penalizing schools based on test scores (the so-called high-stakes testing).³ The first wave of reform began in the early 1980s and was given momentum

cism for being too narrowly focused and not testing higher-level skills. A second wave of reform in the 1990s introduced standardized tests that were not as dependent on the multiple-choice format and that emphasized a broader range of skills. These were sometimes referred to as “tests worth teaching to.” Assessment programs also began to rely on other measures *in addition to* test scores to evaluate students’ achievement levels (for example, portfolios of students’ work, presentations, and longer term projects).

Despite the reform efforts in the 1980s and 1990s and some improvement in national scores, achievement levels of U.S. students remained unacceptably low at the beginning of this century, giving rise to the testing requirements of NCLB. (See *Achievement Levels of U.S. Students*.) NCLB mandates yearly testing in math and reading in grades three through eight no later than the 2005-06 school year.⁴ States are allowed to develop and administer their own tests, but they must specify what constitutes the acceptable level of proficiency for each grade. A sample of fourth and eighth graders from each state must also participate in the state-level NAEP tests every other year to provide a basis of comparison with the state’s own tests.

NCLB requires that all students in each school reach the state-designated proficiency level on the state’s own tests by the end of the 2013-14 school year. Prior to the



requirement for promotion or graduation or as a benchmark for assigning students to remedial programs. Prior to 1975, only two states had mandated any kind of minimum competency test; by 1980, however, 29 states had mandated such tests.² Minimum competency tests are essentially a tool for

² See U.S. Congress, Office of Technology Assessment, p. 59.

by the publication of *A Nation at Risk*, a critical report on the state of American education by the National Commission on Excellence in Education. By the end of the decade many of the standardized tests introduced in this first wave of reform came under criti-

³ Test scores are also the most frequently used output measure in studies that estimate the effects of various school inputs; see the 1997 article by Eric Hanushek.

⁴ NCLB continues the previous requirement that students be tested at least once in these two subjects in grades 10-12. By the 2007-08 school year, states will be required to test students in science at least once in the grade spans 3-5, 6-9, and 10-12.

2013-14 school year, schools that have not reached 100 percent proficiency must make adequate yearly progress toward that goal.⁵ Adequate progress must be made for all students and for major subgroups of students (by race, ethnicity, income, and disability). The penalties for not achieving adequate progress become progressively severe.

For students in any school that fails to make adequate progress for two consecutive years, the district must provide them with a choice of public schools they can attend, and the state may be required to spend up to 5 percent of its federal funds under Title I to pay for that option. Subsequent years of inadequate progress result in further penalties. After five consecutive years of inadequate progress, districts are required to set up an alternative governance structure for the school. This could include reconstituting it as a charter school, turning over management to a private company, or having the state run the school. Thus, NCLB has significantly raised the stakes for schools based on student performance on standardized tests.

TO WHAT EXTENT DO STANDARDIZED TEST SCORES MEASURE SCHOOL QUALITY OR PERFORMANCE?

Individual Student Scores.

Scores on standardized tests are primarily a measure of student achievement or competence in the subject being tested. They provide a better basis of comparison between students in different classrooms or different schools than scores on teacher-generated tests or course grades. Standardized test scores are not a perfect measure of achievement or competence, however. A written test cannot capture the full range

⁵ There is some exception to the 100 percent goal for the learning disabled.

of a student's abilities, and every test involves a certain amount of measurement error.⁶ The reliability of a test is measured by the standard error of measurement or the degree to which the scores would spread out around the average score if the same student took the test many times. The measurement error on standardized tests can stem from a number of random factors, such as the student's health on the day of the test, the form of the test the student receives, or how

The measurement error on standardized tests can stem from a number of random factors, such as the student's health on the day of the test, the form of the test the student receives, or how well the student slept the night before.

well the student slept the night before. A mark of a well-designed test is that the measurement error is small relative to the range of scores on the test. For example, scores for the SAT I test used for college admissions range from 200 to 800, and the standard error of measurement is 30 points. In practice, this means that a test-taker could be 68 percent sure that her score on the test is within 30 points either way of her "true score" or average score if she took the test many times. She could be 95 percent sure that her score on the test is within 60 points either way of her true score.

The existence of measurement error raises a serious issue for minimum competency tests. There will be some misclassification in *both directions* when cutoff scores are used to determine which students meet the minimum level of proficiency. Each time the test is given, some students

⁶ See the article by Vi-Nhuan Le and Stephen Klein.

who are above the required achievement level are likely to score below the minimum and vice versa. For this reason most states that require a minimum proficiency score allow the students to take the test several times.⁷ Unless students are allowed to take minimum competency tests more than once, the temptation will always be to lower the cutoff score to account for the measurement error in a single test score. The incentive to lower proficiency levels on tests is compounded

by the fact that the national legislation provides no national standard for proficiency. Each state is allowed to set its own proficiency levels.

Average School Scores.

Although tests are primarily a measure of individual student achievement, average scores or the percent of students scoring above a certain level are increasingly being used as measures of school quality and accountability. Usually a school will have some students with high scores and some with low scores, so the average score for the school will be somewhere in between, and the range of average scores across schools is much narrower than the range of individual scores for all students. For example, on the math and reading tests administered to fifth, eighth, and 11th graders as part

⁷ See Chapter 3 of the report from the Center on Education Policy. Of the 19 states that had adopted a high-school exit exam in 2003, all but one allowed students to take the test two or more times. The one exception was Washington State, and the state's minimum competency requirement had not yet gone into effect.

of the Pennsylvania System of School Assessment (PSSA) in 2002, the range for school scores was only 50 percent to 60 percent as wide as the range for individual student scores in the state.⁸ Researchers have consistently found that most of the variation in test scores is accounted for by the variation in individual students' scores within schools rather than by the variation between schools.⁹ Thomas Kane and Douglas Staiger (2002b) report that the variation in fourth-grade math and reading scores in a typical North Carolina school is about 90 percent as large as the variation among all the state's fourth graders. The large variation in scores *within* schools is an argument for the NCLB requirement that not only schools as a whole but also major subgroups within each school make adequate yearly progress toward proficiency.

Since average scores for schools will be reported in states' annual reports, it is important to understand how reliable these average scores are and how well they measure the quality of the school. Average scores for a class are a more reliable measure of the "true" class average than is any individual's score of his true average. Many of the random factors that

⁸ For example, the individual math scores for the 11th grade ranged from 700 (at the first percentile) to 1893 (at the 99th percentile); the average school scores ranged from 770 (at the first percentile) to 1460 (at the 99th percentile). See the report from the Pennsylvania Department of Education.

⁹ In general, the variation among schools' average scores accounts for only 10 percent to 20 percent of the total variation in test scores; the rest is due to variation among individual students within schools. The Coleman report in the mid 1960s found that nationally about 16 percent or less of the variation in reading and math scores on achievement tests for sixth, ninth, and 12th graders could be attributed to variation across schools. David Figlio (February 13, 2002) reported that only 14 percent to 15 percent of the variation in math and reading scores in two Florida school districts could be attributed to the variation between schools.

affect individual students' scores (for example, a student's health or the form of the test) tend to cancel out when scores are averaged across an entire class. However, some random factors, such as a distraction in the classroom, poor lighting, or imprecise instructions from the teacher, can affect average scores for the class as a whole. In their study of math and reading scores for fourth graders in North Carolina from 1992 to 1999, Kane and Staiger estimated that these types of random factors accounted for a relatively small

ing a different sample of students from the neighborhood each year (Table, column 3, row 1).¹¹

Thus, for the typical school random factors and cohort effects (an abnormal number of good or poor students) account for about 15 percent of the variation in school scores. But these factors influence the average scores of smaller schools more than those of larger schools. As a result a greater percentage of smaller schools tend to be at the top and bottom of the distribution of average scores in a given

Most of the variation in the average *level* of test scores from school to school is persistent; that is, it is not due to factors that change on a yearly basis.

percentage of the variation in average school scores — only 3.6 percent of the total variation among mid-size schools (Table, column 2, row 1).¹⁰

How much of the remaining variation in average test scores is due to differences in the instructional quality of the schools? Certainly, not all of it. Some of the variation in average scores across schools is due to differences in the quality of the students. The quality of students differs not only across schools and school districts but also across cohorts or age groups within the same school. In any given year, the students in a particular grade may be brighter than the students in other years *even though* they come from the same neighborhood. In their North Carolina sample, Kane and Staiger estimated that almost 11 percent of the variation in the combined reading and math scores for the fourth grade among mid-size schools is due to draw-

year. Kane and Staiger estimated that the combined effect of random factors and different cohorts accounts for less than 10 percent of the variation in average fourth-grade reading and math scores among the largest schools and almost 20 percent of the variation in scores among the smallest schools.¹² NCLB recognizes the problems with the high variability in scores for small samples by not requiring that average scores be reported for subgroups when the number of children is small.

Most of the variation in the average *level* of test scores from school to school is persistent; that is, it is not due to factors that change on a yearly

¹¹ This does not include differences in the student population across neighborhoods serviced by different schools.

¹² The largest quintile of schools in North Carolina has an average of 104 students in the fourth grade, and the smallest quintile has an average of 28 students in the fourth grade. The greater variability in test scores for smaller schools than larger schools due to these transitory effects has also been documented for Chile. See the paper by Kenneth Chay, Patrick McEwan, and Miguel Urquiola.

¹⁰ See Kane and Staiger, 2002b. Mid-size schools are those in the middle quintile by size; on average they have 56 fourth-grade students.

basis. Kane and Staiger's analysis of the variation in fourth-grade math and reading scores in North Carolina suggests that about 85 percent of the variation in the *level* of test scores across mid-size schools is persistent (Table, column 1, row 1). Evidence from the PSSA also shows that the relative differences in scores across schools are persistent. The correlation of 11th-grade scores for public high schools for consecutive years between 1998 and 2002 is approximately 0.88 for math and 0.80 for reading.¹³

¹³ Although schools' relative PSSA test scores are fairly stable across years, in every year between 1999 and 2002, there are examples of the average 11th grade math or reading score for a school moving from the 25th to the 75th percentile in the state or vice versa.

Who's Responsible for These Persistent Differences in Test Scores? Are these persistent differences a measure of the quality of the school or a measure of the abilities and backgrounds of the students? Economic studies of the educational process have identified three major influences on student achievement besides the quality of the school: the student's innate ability and family characteristics and the characteristics of the student's classroom peers.¹⁴

Teachers are well aware of the wide range of student abilities from the learning disabled to the gifted. But it

¹⁴ See the article by Byron Brown and Daniel Saks, and the 1979 and 1986 articles by Hanushek.

is difficult to get a pure measure of the innate ability of students. Initial test scores are not a pure measure of innate ability; by the time students enter the school system their achievement levels have been influenced by a number of environmental and social factors. Moreover, as students progress through the school system, their achievement levels are the result of their cumulative educational experience.

Family characteristics, such as the parents' education and income, can also affect the level of student achievement and test scores. Education takes place not only in the classroom but also at home; in general, students whose parents are more highly educated have a better educational environment in the home. For example,

TABLE

Sources of Variation in Fourth-Grade Test Scores for Mid-Size Schools in North Carolina (Average number of fourth graders = 56)

| | % of Variance Due to Persistent Characteristics of the School | % of Variance Due to Purely Random Factors | % of Variance Due to Differences in Cohorts Within the School |
|---|---|--|---|
| Combined Reading and Math Scores for Fourth Grade | 85.4% | 3.6% | 10.9% |
| Change in Combined Reading and Math Scores for Fourth Grade from One Year to the Next (Variation is 40 percent as great as variation in scores across schools) | 29.1% | 16.4% | 54.5% |
| Change in Combined Reading and Math Scores from Third Grade to Fourth Grade (Variation is 23 percent as great as variation in scores across schools) | 51.6% | 35.5% | 12.9% |

Source: Author's calculations from Kane and Staiger, 2002b, Table 2. Mid-size schools are those in the middle quintile based on size. Time frame: 1992-99.

students who are exposed to a richer vocabulary in their family conversations are more likely to perform better in language arts than students who are exposed to a limited vocabulary. Income matters as well. Higher income families can better afford certain aids to education, such as computers or private tutors. In addition, students from higher income families are more likely to have more educational experiences like foreign travel.

Finally, a number of studies have found that the achievement levels and other characteristics of fellow students in the classroom can have an effect on a student's own achievement and test scores — the so-called peer-group effect.¹⁵ Peers can provide motivation for a student. They can contribute to learning through direct interaction, or they can affect the learning process in the classroom. A disruptive student clearly hinders the learning process for his peers, but a bright student can aid the process by asking questions that help other students as well.¹⁶ A good set of peers in the classroom can increase the quality of the school by enhancing the learning environment, but in the U.S. public school system, classroom peers are largely determined by the families who choose to live in the neighborhood, not by the school.

The cumulative effect of students' innate abilities, family backgrounds, and peers will be reflected

¹⁵ See the articles by Anita Summers and Barbara Wolfe; Vernon Henderson, Peter Mieszkowski, and Yvon Sauvageau; Ron Zimmer and Eugenia Toma; Caroline Hoxby; and Erick Hanushek, John Kain, Jacob Markman, and Steven Rivkin. Peer-group effects are difficult to isolate and hard to separate from school effects, since students generally attend school for some years with most of their classroom peers. Joshua Angrist and Kevin Lang found only very weak evidence of peer-group effects in their study.

¹⁶ See the article by Edward Lazear and the one by Hanushek et al.

in tests scores. Schools, however, have little or no influence over these factors, raising the issue of how test scores can be used to judge the school's contribution to learning. One suggestion is to use *changes* in test scores rather than the *level* of scores to measure school quality and performance.¹⁷

ARE CHANGES IN TEST SCORES A BETTER MEASURE OF SCHOOL QUALITY?

Changes in Scores for a Given Grade. There are several ways to measure changes in test scores in a system of school accountability.¹⁸ The first is to compare this year's score for

Random factors such as a large number of students with a cold on the day of the test can also contribute to the change in scores for a given grade from one year to the next.

a given grade with last year's score for that grade, for example, the change in the average test score for fourth grade. Of course, we are not comparing the same students in this exercise, and so it is difficult to know how much any increase in the average score represents an improvement in student achievement. The school's contribution to this increase in the score is equally difficult to assess. Kane and Staiger estimate that more than 50 percent of the variation in the annual change in fourth-

¹⁷ See Hanushek's 1986 article, and the article by Hanushek and Lori Taylor.

¹⁸ See the article by Laura Hamilton and Daniel Koretz.

grade reading and math scores across mid-size schools in North Carolina is due to the fact that a different cohort of students is being tested each year, and each cohort has a different average level of ability (Table, column 3, row 2).

Besides these cohort effects, random factors such as a large number of students with a cold on the day of the test can also contribute to the change in scores for a given grade from one year to the next. According to Kane and Staiger, the combination of cohort effects and these kinds of random factors accounts for more than 70 percent of the variation in the annual change in fourth-grade scores in North Carolina.¹⁹

Changes in Scores for a Given Cohort of Students. A partial solution to the problem of comparing two different cohorts of students is to compare this year's average fourth-grade score with last year's third-grade score. But this is only a partial solution for two reasons: The composition of the class may have changed as some students enter or leave the class, and even if there has been no change in the composition of the class, different cohorts of students advance at different rates. If a cohort of particularly able students has moved from third to fourth grade this year, a larger than average increase in scores may not be due to the school at all. The change in scores from one grade to the next tends to be considerably less variable than the change in scores for a given grade. But even in this case, only about half the variation in the change in scores can be attributed to differences

¹⁹ See Kane and Staiger, 2002b. In a study of test scores in Florida, David Figlio and Marianne Page found that the correlation between *changes* in average test scores in consecutive years for a given grade at a school was negative, supporting the notion that changes in test scores are a noisy measure of school quality.

in the quality of the schools (Table, column 1, row 3). Kane and Staiger estimate that in mid-size schools in North Carolina about 13 percent of the variation in the average change in test scores from third to fourth grade is due to the cohort that is advancing that year, and more than 35 percent is due to purely random factors (Table, columns 2-3, row 3).

Changes in Individuals' Test Scores. A more refined measure of the value added by a school is the improvement in individual student scores over time rather than the improvement in class scores from one year to the next.²⁰ Students' test scores are highly correlated from one year to the next, so it may take a longer period to capture meaningful changes in a student's scores compared to the average change.²¹ But data on individual students' test scores are difficult to maintain over time, especially for students who are very mobile. Furthermore, tracking individual students does not solve all the issues of identifying the school's contribution to any improvement in scores. Family background and innate ability influence not only the level of scores at a point in time but the rate of change as well. A student whose father or mother has a graduate degree in engineering is likely to get more help on his algebra homework and, therefore, advance more quickly than the student whose parents did not graduate from high school. Kane and Staiger found that students in North Carolina whose parents had a higher

²⁰ Figlio and Page found that the ranking of Florida schools based on the improvement in individual scores is very different from the ranking based on the average level of scores in a given year. But the available data only allowed them to calculate the change in individual reading scores from the fourth to the fifth grade.

²¹ In Kane and Staiger's study of North Carolina schools, the correlation between individual students' standardized third- and fourth-grade scores was 0.80.

level of education had greater gains in test scores from the end of third grade to the end of fourth grade. Whether we compare schools based on the level of test scores or some measure of change in scores, the school's contribution has to be determined in light of the innate abilities and backgrounds of the students.

Teaching to the Test. Another word of caution has to be raised about changes in scores on high-stakes tests whose results have serious consequences for the school. No matter how we measure changes in test scores, there is a tendency in the early years after a new high-stakes test is introduced for scores to rise rapidly. Daniel Koretz provides a striking example of inflation in high-stakes test scores.²² He and his colleagues tracked student

²² See the article by Daniel Koretz, Robert Linn, Stephen Dunbar, and Lorrie Shepard, and both articles by Koretz.

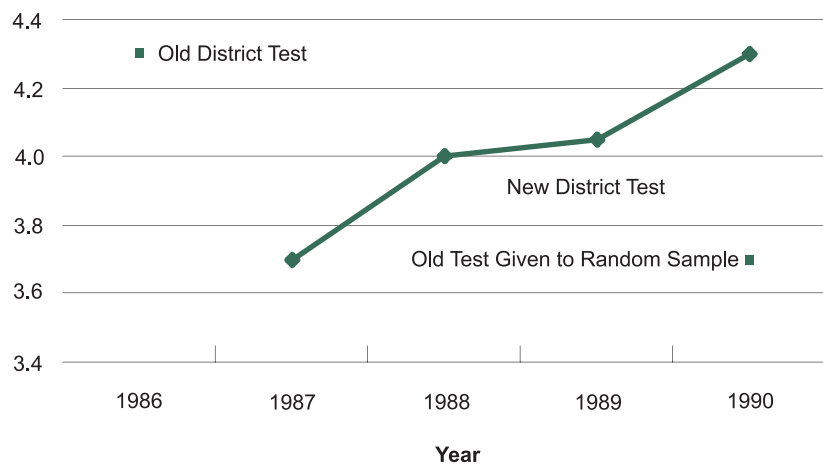
performance in two school districts on several tests; some were high-stakes tests and some were not. The figure illustrates what happened to third-grade math scores in one district that changed its high-stakes test between 1986 and 1987. In the final year in which the old test was given, the median grade equivalent was 4.3 for the third graders in the district. In the first year of the new test, the grade equivalent dropped to 3.7, but by the fourth year of administering the new test, the median grade had climbed back to 4.3. In the fourth year of the new test, Koretz and his colleagues administered the old test to a random sample of third graders. Their median score on the old test was 3.7. Scores on the new district-wide test had increased substantially in the four years, and scores on the old test had dropped.

The initial rapid rise in high-stakes test scores is often attributed to the practice of "teaching to the test." There is evidence that teachers do

FIGURE

Median Grade Equivalent (Old Test and New Test) Third-Grade Mathematics

Grade Equivalents



Source: Daniel M. Koretz, *Journal of Human Resources*, 2002. Used with permission of the author.

spend more time on the subjects tested in their grade than on other subjects.²³ In any given subject, teachers can emphasize the material they know will be covered on a high-stakes test. These are not necessarily negative consequences of high-stakes testing. If high-stakes tests adequately cover the essential material to be learned in each grade, these practices can enhance the teaching in the classroom. Teaching to a well-designed set of tests can improve both test scores and student achievement. But we cannot assume that every improvement in test scores is an improvement in overall academic achievement. Some classroom practices improve test scores on high-stakes tests but have little or no effect on achievement levels. For example, teachers learn over time how to administer tests with less confusion, and they prepare students for the format of the new high-stakes test.

One check on whether higher test scores are measuring true gains in achievement or simply reflect score inflation is to compare the improvement in scores on high-stakes tests with improvement in other test scores. Researchers have compared gains in several state-mandated tests with gains in the NAEP tests taken in the state. The results are mixed. Gains in test scores from the Kentucky Instructional Results Information System (KIRIS) were not matched by gains in the state's NAEP scores.²⁴ In the first two years of the program, fourth-grade reading scores increased dramatically on the KIRIS tests but did not increase at all on the NAEP tests. In the first four years of the program, fourth- and eighth-grade math scores increased three-and-a-half to four times more

²³ See the article by Brian Stecher.

²⁴ See the paper by Daniel Koretz and Sheila Barron.

on the KIRIS tests than on the NAEP tests. Perhaps the most publicized high-stakes testing program in the country has been the Texas Assessment of Academic Skills (TAAS). Stephen Klein and his associates compared gains in TAAS scores from 1994 to 1998 with gains in the Texas NAEP scores.²⁵ Both sets of tests showed gains in reading and math in fourth and eighth grade, but the gains on the TAAS tests were much larger than those on the NAEP tests. Moreover, other educational

One check on whether higher test scores are measuring true gains in achievement or simply reflect score inflation is to compare the improvement in scores on high-stakes tests with improvement in other test scores.

outcomes such as graduation rates or plans to attend college have not improved with the gains in the Texas test scores.²⁶ Unlike the situation in Kentucky and Texas, increases in scores on the North Carolina state tests were about the same as increases in the NAEP scores in the state. This may be because the North Carolina tests are more similar to the NAEP tests.²⁷ The possibility of serious grade inflation on high-stakes test scores reinforces the need for a comparison test such as the NAEP, against which we can measure any improvement in high-stakes test scores.

²⁵ See the article by Stephen Klein, Laura Hamilton, Daniel McCafferty, and Brian Stecher. Robert Linn, Eva Baker, and Damian Betebenner also point out that the percent of students meeting proficiency levels in the TAAS tests increased much faster than the percent of students meeting proficiency levels on the Texas NAEP tests.

²⁶ See the article by Martin Carnoy, Susanna Loeb, and Tiffany Smith.

²⁷ See Thomas Kane and Douglas Staiger, 2002a.

SUMMARY:

SOME CONSIDERATIONS ON THE USE OF TEST SCORES TO EVALUATE SCHOOL QUALITY

Test scores are primarily a measure of the achievement levels of individual students, but they are increasingly being used to measure the quality of schools. This new role for testing is a response to the performance of U.S. students relative to students from other industrialized countries and to the large percent-

age of U.S. students who do not meet proficiency levels on standardized tests. The new testing programs are designed to hold schools as well as students accountable. Test scores and changes in test scores are one of the few quantitative measures of school quality we have, but special precautions need to be taken when test scores are used to evaluate schools rather than students.

Perhaps the most popularly accepted notion in judging the quality of schools is that *all* students should achieve a minimum level of competency based on some standardized test in order to graduate or be promoted. But there is some measurement error in the score on every test, and some students who are above the minimum in achievement will not receive the minimum score on a single administration of the test. Therefore, if all students are required to score above the minimum on a single administration of the test, states will be tempted to lower the cutoff score for proficiency to account for the measurement error. Students should have more than one chance to achieve the minimum score


on these tests, and this should be true of tests that have serious consequences for the school as well as those that have serious consequences for the students.

Average school scores are less susceptible to measurement error than individual student scores. But the average score may not measure the school's contribution to the students' achievement for several reasons. Each cohort of students in a school will differ in their abilities, and the family characteristics and innate abilities of students will differ from school to school. Moreover, peer effects can magnify these differences. Therefore, if we want to use average scores to judge the quality of schools, we must look at scores over several years and compare

scores for schools that have students from similar backgrounds.

Theoretically, changes in test scores should be a better measure of the school's contribution to student achievement than average scores. But there is a lot of random variation in the changes in scores for a given grade or for a given class from one year to the next. Longer-term trends in test scores can eliminate some of this random variation in the changes in scores. But not all of the long-term improvement in class scores or individual scores can be attributed to the school. Family characteristics and peer effects influence how quickly students advance in their education. So every easily accessible measure of student achievement has some drawback as a

measure of school quality.

Despite the shortcomings of standardized test scores as a measure of school performance there is no generally recognized substitute; test scores simply have to be used with caution. Improvements in high-stakes test scores should be checked against improvements in other tests such as the state-level NAEP tests. Other measures of student achievement, such as course grades and performance on longer-term projects, can be incorporated into the evaluation of school quality. Finally, other criteria, such as graduation rates and the percent of students attending college are important in evaluating how well our schools perform. 

Appendix: Achievement Levels of U.S. Students

S

ince the 1960s, a number of countries have administered math and science tests so that student achievement can be compared across countries. U.S. students have tended to score in the middle of the pack or lower in these international comparisons.^a In the latest Third International Mathematics and Science Study (TIMSS) conducted in 1999, eighth-grade students in the U.S. ranked 19th in math among the 38 countries participating. In science, they ranked 18th out of 38 (Table A1). A number of explanations have been offered for the poor ranking of the U.S. in the TIMSS tests relative to nations like Japan, Korea, the Netherlands, and Australia, but there are no simple explanations for the differences in performance across countries.^b Nonetheless, the rankings suggest considerable room for improvement in the American education system.

The trend and dispersion in student achievement within the U.S. are illustrated by the scores on the

two types of tests given as part of the National Assessment on Educational Progress (NAEP) — the national trend tests and the state tests. The scores from the long-term trend NAEP tests offer the best assessment of student achievement over time, since the tests have changed very little since they were first administered. The average math and science scores on these tests show a pattern of deterioration in the 1970s, improvement in the 1980s, and a leveling off in the 1990s. The math scores have shown the most consistent improvement (Figures A1 and A2). Reading scores have shown little sustained improvement since the tests were first administered (Figure A3). For all age groups (9, 13, and 17) the latest reading scores are not significantly higher than they were in 1980.^c

The state-level NAEP tests, which were first administered in the early 1990s, differ from the tests that capture the national trend because they are adjusted over time to reflect changing curricula and they are given in specified grades, not at given age levels. The National

^a See Eric Hanushek's 1998 article.

^b See the article by Deborah Nelson.

^c See the report from the U.S. Department of Education.

TABLE A1**1999 TIMSS SCORES**

| 8th Grade Math | | 8th Grade Science | |
|--------------------|-----|-------------------|--------------------|
| Singapore | 604 | 569 | Chinese Taipei |
| Korea | 587 | 568 | Singapore |
| Chinese Taipei | 585 | 552 | Hungary |
| Hong Kong | 582 | 550 | Japan |
| Japan | 579 | 549 | Korea |
| Belgium | 558 | 545 | Netherlands |
| Netherlands | 540 | 540 | Australia |
| Slovak Republic | 534 | 539 | Czech Republic |
| Hungary | 532 | 538 | England |
| Canada | 531 | 535 | Belgium |
| Slovenia | 530 | 535 | Finland |
| Russian Federation | 526 | 535 | Slovak Republic |
| Australia | 525 | 533 | Canada |
| Czech Republic | 520 | 533 | Slovenia |
| Finland | 520 | 530 | Hong Kong |
| Malaysia | 519 | 529 | Russian Federation |
| Bulgaria | 511 | 518 | Bulgaria |
| Latvia | 505 | 515 | United States |
| United States | 502 | 510 | New Zealand |
| England | 496 | 503 | Latvia |
| New Zealand | 491 | 493 | Italy |
| Lithuania | 482 | 492 | Malaysia |
| Italy | 479 | 488 | Lithuania |
| Cyprus | 476 | 482 | Thailand |
| Romania | 472 | 472 | Romania |
| Moldova | 469 | 468 | Israel |
| Thailand | 467 | 460 | Cyprus |
| Israel | 466 | 459 | Moldova |
| Tunisia | 448 | 458 | Macedonia |
| Macedonia | 447 | 450 | Jordan |
| Turkey | 429 | 448 | Iran |
| Jordan | 428 | 435 | Indonesia |
| Iran | 422 | 433 | Turkey |
| Indonesia | 403 | 430 | Tunisia |
| Chile | 392 | 420 | Chile |
| Philippines | 345 | 345 | Philippines |
| Morocco | 337 | 323 | Morocco |
| South Africa | 275 | 243 | South Africa |

Source: U.S. Department of Education, National Center for Education Statistics. *Pursuing Excellence: Comparisons of International Eighth Grade Mathematics and Science Achievement from a U.S. Perspective, 1995 and 1999.*

TABLE A2**Percent of Students Scoring Below Basic Level (NAEP)**

| | Math | | Reading | |
|-------------------------|-----------|-----------|-----------|-----------|
| | 4th Grade | 8th Grade | 4th Grade | 8th Grade |
| Nation (public schools) | 24 | 33 | 38 | 28 |
| Alabama | 35 | 47 | 48 | 35 |
| Alaska | 25 | 30 | 42 | 33 |
| Arizona | 30 | 39 | 46 | 34 |
| Arkansas | 29 | 42 | 40 | 30 |
| California | 33 | 44 | 50 | 39 |
| Colorado | 23 | 26 | 31 | 22 |
| Connecticut | 18 | 27 | 26 | 23 |
| Delaware | 19 | 32 | 29 | 23 |
| District of Columbia | 64 | 71 | 69 | 53 |
| Florida | 24 | 38 | 37 | 32 |
| Georgia | 28 | 41 | 41 | 31 |
| Hawaii | 32 | 44 | 47 | 39 |
| Idaho | 20 | 27 | 36 | 24 |
| Illinois | 27 | 34 | 39 | 23 |
| Indiana | 18 | 26 | 34 | 23 |
| Iowa | 17 | 24 | 30 | 21 |
| Kansas | 15 | 24 | 34 | 23 |
| Kentucky | 28 | 35 | 36 | 22 |
| Louisiana | 33 | 43 | 51 | 36 |
| Maine | 17 | 25 | 30 | 21 |
| Maryland | 27 | 33 | 38 | 29 |
| Massachusetts | 16 | 24 | 27 | 19 |
| Michigan | 23 | 32 | 36 | 25 |
| Minnesota | 16 | 18 | 31 | 22 |
| Mississippi | 38 | 53 | 51 | 35 |
| Missouri | 21 | 29 | 32 | 21 |
| Montana | 19 | 21 | 31 | 18 |
| Nebraska | 20 | 26 | 34 | 23 |
| Nevada | 31 | 41 | 48 | 37 |
| New Hampshire | 13 | 21 | 25 | 19 |
| New Jersey | 20 | 28 | 30 | 21 |
| New Mexico | 37 | 48 | 53 | 38 |
| New York | 21 | 30 | 33 | 25 |
| North Carolina | 15 | 28 | 34 | 28 |
| North Dakota | 17 | 19 | 31 | 19 |
| Ohio | 19 | 26 | 31 | 22 |
| Oklahoma | 26 | 35 | 40 | 26 |
| Oregon | 21 | 30 | 37 | 25 |
| Pennsylvania | 22 | 31 | 35 | 24 |
| Rhode Island | 28 | 37 | 38 | 29 |
| South Carolina | 21 | 32 | 41 | 31 |
| South Dakota | 18 | 22 | 31 | 18 |
| Tennessee | 30 | 41 | 43 | 31 |
| Texas | 18 | 31 | 41 | 29 |
| Utah | 21 | 28 | 34 | 24 |
| Vermont | 15 | 23 | 27 | 19 |
| Virginia | 17 | 28 | 31 | 21 |
| Washington | 19 | 28 | 33 | 24 |
| West Virginia | 25 | 37 | 35 | 28 |
| Wisconsin | 21 | 25 | 32 | 23 |
| Wyoming | 13 | 23 | 31 | 21 |

Source: <http://nces.ed.gov/nationsreportcard/>
 The three states in the Third Federal Reserve District are shaded.

Assessment Governing Board, which oversees the test, has adopted three achievement levels for reporting the results — basic, proficient, and advanced.^d The basic level “denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade” (www.nagb.org/about/achievement.html).

The No Child Left Behind Act required all states to participate in these tests for fourth- and eighth-grade students by the 2002-03 school year. The results were not encouraging (Table A2). Nationwide, 24 percent of fourth-grade public-school students and 33 percent of eighth graders scored below the basic level in math. Even in the best performing states, 13 percent of fourth graders and 18 percent of eighth graders scored below the basic

level. In the three worst performing states, more than one-third of the fourth graders scored below the basic level and in 10 states more than 40 percent of the eighth graders scored below basic.^e On the reading tests 38 percent of fourth graders and 28 percent of eighth graders nationwide scored below the basic level. Even in the best performing states, 25 percent of fourth graders and 18 percent of eighth graders scored below basic. In the 13 worst performing states, more than 40 percent of fourth graders scored below basic in reading, and in seven states, more than one-third of the eighth graders scored below basic in reading. These results suggest that the need for improvement in student achievement is not limited to a few states or school districts.

^d These are not related to the proficiency levels to be determined by each state according to the No Child Left Behind Act.

^e These numbers exclude the District of Columbia where more than 50 percent of the fourth- and eighth-grade students scored below basic on the math and reading tests.

FIGURE A1

Average Math Scores (NAEP)



Source: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics. *NAEP 1999, Trends in Academic Progress: Three Decades of Student Performance.*

FIGURE A2

Average Science Scores (NAEP)



Source: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics. NAEP 1999, *Trends in Academic Progress: Three Decades of Student Performance*.

FIGURE A3

Average Reading Scores (NAEP)



Source: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics. NAEP 1999, *Trends in Academic Progress: Three Decades of Student Performance*.

REFERENCES

- Angrist, Joshua D., and Kevin Lang. "How Important Are Classroom Peer Effects? Evidence from Boston's METCO Program," NBER Working Paper 9263 (October 2002).
- Baker, Eva L., and Robert L. Linn. "Validity Issues for Accountability Systems," CSE Technical Report 585, Center for the Study of Evaluation, National Center for Research on Education (December 2002).
- Brown, Byron W., and Daniel H. Saks. "The Production and Distribution of Cognitive Skills Within Schools," *Journal of Political Economy*, 83 (1975), pp. 571-93.
- Carnoy, Martin, Susanna Loeb, and Tiffany L. Smith. "Do Higher State Test Scores in Texas Make for Better High School Outcomes?" Research Report Series RR-047, Consortium for Policy Research and Education, November 2001.
- Center on Education Policy. *State High School Exit Exams Put to the Test*. Washington, DC, August 2003.
- Chay, Kenneth Y., Patrick J. McEwan, and Miguel Urquiloa. "The Central Role of Noise in Evaluating Interventions That Use Test Scores to Rank Schools," NBER Working Paper 10118, November 2003.
- Coleman, James S., et al. *Equality of Educational Opportunity*. Washington, DC: U.S. Department of Health, Education, and Welfare, 1966.
- Figlio, David N. "Aggregation and Accountability," presented at Will No Child Truly Be Left Behind? The Challenge of Making This Law Work, a conference sponsored by the Thomas B. Fordham Foundation, February 13, 2002.
- Figlio, David N., and Marianne E. Page. "Can School Choice and School Accountability Successfully Coexist?" in Caroline M. Hoxby (ed.): *The Economics of School Choice*. Chicago: University of Chicago Press, 2003.
- Hamilton, Laura S., and Daniel M. Koretz. "Tests and Their Use in Test-Based Accountability Systems," in Laura Hamilton, Brian M. Stecher, and Stephen P. Klein (eds.): *Making Sense of Test-Based Accountability in Education*. Rand Corporation, 2002, pp.13-49.
- Haney, Walt. "The Myth of the Texas Miracle in Education," *Education Policy Archives*, 8, 41 (August 2000).
- Hanushek, Eric A. "Conceptual and Empirical Issues in the Estimation of Educational Production Functions," *Journal of Human Resources*, 14 (1979), pp. 351-88.
- Hanushek, Eric A. "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature*, 24 (1986), pp. 1141-77.
- Hanushek, Eric A. "Assessing the Effects of School Resources on Student Performance: An Update," *Educational Evaluation and Policy Analysis*, 19 (1997), pp. 141-64.
- Hanushek, Eric A. "Conclusions and Controversies about the Effectiveness of School Resources," *Economic Policy Review*, Federal Reserve Bank of New York, March 1998, pp. 11-27.
- Hanushek, Eric A., John F. Kain, Jacob M. Markman, and Steven G. Rivkin. "Does Peer Ability Affect Student Achievement?" NBER Working Paper 8502 (October 2001).
- Hanushek, Eric A., and Lori L. Taylor. "Alternative Assessments of the Performance of Schools: Measurement of State Variations in Achievement," *Journal of Human Resources*, 25 (1990), pp. 179-201.
- Henderson, Vernon, Peter Mieszkowski, and Yvon Sauvageau. "Peer Group Effects and Educational Production Functions," *Journal of Public Economics*, 10 (1978), pp. 97-106.
- Hoxby, Caroline. "Peer Effects in the Classroom: Learning from Gender and Race Variation," NBER Working Paper 7867 (August 2000).
- Jaeger, Richard M. "The Final Hurdle: Minimum Competency Achievement Testing," in Gilbert R. Austin and Herbert Garber (eds.): *The Rise and Fall of National Test Scores*. NY: Academic Press, 1982, pp. 223-46.
- Kane, Thomas J., and Douglas O. Staiger. "Improving School Accountability Measures," NBER Working Paper 8156 (March 2001).
- Kane, Thomas J., and Douglas O. Staiger. "The Promise and Pitfalls of Using Imprecise School Accountability Measures," *Journal of Economic Perspectives*, 16 (2002a), pp. 91-114.
- Kane, Thomas J., and Douglas O. Staiger. "Volatility in School Test Scores: Implications for Test-Based Accountability Systems," in Dianne Ravitch (ed.): *Brookings Papers on Education Policy*, 2002b, Washington, DC: Brookings Institution, pp. 235-83.

REFERENCES

- Klein, Stephen P., Laura S. Hamilton, Daniel F. McCaffery, and Brian Stecher. "What Do Test Scores in Texas Tell Us?" *Education Policy Analysis Archives*, 8, 49 (October 26, 2000).
- Koretz, Daniel M. "Limitations in the Use of Achievement Tests as Measures of Educators' Productivity," *Journal of Human Resources*, 38 (2002), pp.752-77.
- Koretz, Daniel M. "Using Student Assessments for Educational Accountability," in Eric A. Hanushek and Dale W. Jorgenson (eds.): *Improving America's Schools*. Washington, DC: National Academy Press, 1996, pp. 171-95.
- Koretz, Daniel M., and Sheila I. Barron. "The Validity of Gains in Scores on the Kentucky Instructional Results Information System (KIRIS)," Santa Monica: Rand Corporation, 1998.
- Koretz, Daniel M., Robert L. Linn, Stephen B. Dunbar, and Lorrie A. Shepard. "The Effects of High-Stakes Testing on Achievement: Preliminary Findings About Generalization Across Tests," presented at *Effects of High-Stakes Educational Testing on Instruction and Achievement*, symposium presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Chicago, April 5, 1991.
- Lazear, Edward P. "Educational Production," *Quarterly Journal of Economics*, 116 (2001), pp. 777-803.
- Le, Vi-Nhuan, and Stephen P. Klein. "Technical Criteria for Evaluating Tests," in Laura S. Hamilton, Brian M. Stecher, and Stephen P. Klein (eds.): *Making Sense of Test-Based Accountability in Education*. Rand Corporation, 2002, pp. 51-77.
- Linn, Robert L., Eva L. Baker, and Damian W. Betebenner. "Accountability Systems: Implications of Requirements of the No Child Left Behind Act of 2001," *Educational Researcher* 31, 6 (2002), pp. 3-16.
- Meyer, Robert H. "Value-Added Indicators of School Performance," in Eric A. Hanushek and Dale W. Jorgenson (eds.): *Improving America's Schools*. Washington, DC: National Academy Press, 1996, pp. 197-223.
- National Commission on Excellence in Education. *A Nation at Risk: The Imperative for Educational Reform*. Washington, DC: Government Printing Office, 1983.
- Nelson, Deborah I. "What Explains Differences in International Performance? TIMSS Researchers Continue to Look for Answers," *Policy Briefs*, Consortium for Policy Research in Education, September 2003.
- Pennsylvania Department of Education. *The Pennsylvania System of School Assessment: Handbook for Report Interpretation 2002 PSSA Mathematics and Reading Assessment*. Harrisburg, PA: November 2002.
- Rothstein, Jesse M. "Good Principals or Good Peers? Parental Valuation of School Characteristics, Tiebout Equilibrium, and the Incentive Effects of Competition Among Jurisdictions," Working Paper 3, Princeton University, Education Research Section, October 2003.
- Stecher, Brian M. "Consequences of Large-Scale, High-Stakes Testing on School and Classroom Practice," in Laura S. Hamilton, Brian M. Stecher, and Stephen P. Klein (eds.): *Making Sense of Test-Based Accountability in Education*. Rand Corporation, 2002, pp. 79-100.
- Summers, Anita A., and Barbara I. Wolfe. "Do Schools Make a Difference?" *American Economic Review*, 67 (1977), pp. 639-52.
- U.S. Congress, Office of Technology Assessment, *Testing in American Schools: Asking the Right Questions*. (OTA-SET-519) Washington, DC: U.S. Government Printing Office, February 1992.
- U.S. Department of Education. National Center for Education Statistics. *Pursuing Excellence: Comparisons of International Eighth Grade Mathematics and Science Achievement from a U.S. Perspective, 1995 and 1999*. NCES 2001-028, by Patrick Gonzales, Christopher Calsyn, Leslie Ioce-lynn, Kitty Mak, David Kastberg, Sousesan Arafeh, Trevor Williams, and Winnie Tsen. Washington, DC: U.S. Government Printing Office, 2000.
- U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics. *NAEP 1999, Trends in Academic Progress: Three Decades of Student Performance*. NCES 2000-469, by J.R. Campbell, C.M. Hornbo, and J. Mazzeo. Washington, DC: 2000.
- Zimmer, Ron W., and Eugenia F. Toma. "Peer Effects in Private and Public Schools Across Countries," *Journal of Policy Analysis and Management*, 19 (2000), pp. 75-92.