To help policymakers, reporters, and others better understand both the strengths and limitations of testing, the Center on Education Policy has reviewed studies and advice from testing experts, state departments of education, and other sources. This issue of TestTalk summarizes what we’ve learned. Our goal is to encourage those who set policy and shape opinion to develop realistic and balanced expectations about what tests can and can’t do.

When state tests and other large-scale tests are well designed and properly used, they can tell us a lot about what students know and can do. For some purposes, they have advantages over less standardized forms of measurement, like grades and teacher-made tests. For example, state test results can:

- Provide information that is more standardized and consistent from school to school or district to district than the results of measures based on an individual teacher’s judgment.

- Be used to compare achievement across different classrooms, schools, or districts, or between various racial, ethnic, income, and other subgroups of students.

- Provide valuable summary information about student performance by subject, skill, and knowledge area.

- Be collected, analyzed, and reported efficiently and at relatively low cost.
But even good tests have limitations—something state and federal policymakers don’t always consider when they design education accountability systems.

Many state accountability systems treat test scores like precise calculations, when really they are more like estimates. Test scores can fluctuate for reasons that have nothing to do with student learning or the quality of teaching. For example, a student’s test score on any given day could be affected by:

- The particular sample of questions (called “items”) included on that version of the test, out of all possible questions that could be asked about a broad subject like 8th grade mathematics.
- The student’s physical condition or state of mind.
- Distractions outside the classroom, like construction noise or a barking dog.
- Other factors, such as a few lucky guesses or errors in scoring.

Test developers try to account for these kinds of score variations by computing a “standard error” of measurement for every test — a set of numbers intended to give people an idea of the accuracy of that test. Similar to the margin of error in an opinion poll, the standard error shows how much a student’s score could be expected to vary, either up or down, if the test were given again and again (assuming no learning took place in between).

Scores can also be affected by “teaching to the test” — instruction that familiarizes students with the content and format of a specific test but doesn’t necessarily improve their understanding of the broader subject being tested. (For more on this issue, see Teaching to the Test: The Good, the Bad, and Who’s Responsible, Issue 1 of TestTalk by the Center on Education Policy.)

Like individual scores, average test scores for a classroom, grade, or school can also fluctuate due to factors unrelated to learning or the quality of teaching — a fact not fully recognized by the new federal No Child Left Behind Act. This law requires schools to track yearly changes in average test scores for most grade levels and for racial, ethnic, income, and other subgroups of students. Schools and districts that fail to improve average scores for all tested grades and all major subgroups must undergo restructuring or other interventions.

Changes in the population of students being tested can produce significant year-to-year variations in average test scores. For example, if this year’s group of 5th graders includes more students with disabilities, English language learners, or low-income children than last year’s group of 5th graders, the average test scores for the 5th grade could drop, simply because of these demographic differences. Small groups, such as an elementary school grade
cohort or the even smaller racial/ethnic subgroups within that grade, are especially sensitive to changes in group composition, because each student’s score has a greater impact on the average test score.

One study estimated that more than 70% of the year-to-year fluctuations in average test scores for a given school or grade can be attributed to variations in student population and other external factors, rather than to changes in learning (Kane & Staiger, 2001). Several studies have found that a school that makes large test score gains in one year is more likely than not to do worse the next year.

In short, annual changes in average test scores for a subgroup, grade, or school can be an undependable gauge of the teaching and learning in that school. An annual rise in average test scores doesn't necessarily mean a school is succeeding, just as a drop in scores doesn't always mean it's failing.

This natural volatility of average scores will make it difficult for schools to post the consistent test score gains demanded by the No Child Left Behind Act. If this score instability results in large numbers of schools being misidentified as low-performing, it could put pressure on states to make their tests easier, lower their cutoff scores, or weaken their standards — the opposite of what the law intends. It could also damage public support for state testing programs.

**Actions for state and national leaders**

No measure of achievement is perfect. State and national policymakers must weigh the benefits and limitations of available tests against the benefits and limitations of other measures. Proposed alternatives to tests should be at least as accurate, consistent, fair, and cost efficient.

In many cases tests will be the most viable option, usually in combination with other measures. With the choice to test, however, comes the responsibility to ensure that tests are used appropriately, with respect for their limitations. If you are a state or national policymaker, here are some steps you can take to help ensure appropriate test use and build public understanding of the benefits and limitations of testing:

- **Self-education.** Educate yourself about the different kinds of tests available and which are most appropriate for which purposes. Read materials developed to go along with your state’s tests, such as the score reports for students and teachers or the guides for parents. Look at practice booklets and released items from past tests to get a feel for the difficulty, fairness, and soundness of the questions. Examine the proficiency levels and cutoff scores for the test, to understand the types of knowledge and skills expected of students at different grades.
- **Score reporting.** Make sure your state reports test scores and other test data in a way that helps people understand such concepts as standard error and discourages them from taking a “horse race” attitude toward test results. For example, use bands to display the range of standard error for individual student scores or graphics to show when differences between average scores for schools are or are not significant. Report test information that will help students and teachers determine which curriculum standards students have mastered and which need more work.

- **Appropriate use.** Pay attention to what testing experts say about how and when to use a particular test. Use tests only for the purposes for which they were designed. For example, a test designed mainly to rank or compare students with each other should not be used to determine whether students have learned the material in state standards for their grade. Don’t expect a single test to do everything, such as track student progress, hold people accountable, and diagnose student needs.

- **Alignment.** Make sure that state tests used for accountability are closely aligned with the state’s curriculum standards for the subject being tested, and that the content and skills being tested are closely aligned with what students are actually being taught.

- **Multiple measures.** Don’t use a single test to make high-stakes decisions about such issues as student promotion or graduation or school quality. Instead, use multiple tests or a test in combination with other measures of student achievement or school performance.

- **Sound accountability systems.** Take a good look at your state’s accountability system to make sure it doesn’t penalize students or educators for test score fluctuations caused by external factors. Base determinations of whether schools are making adequate yearly progress on multiple years of data (the No Child Left Behind Act allows states to combine up to 3 years of data to make these determinations). Explore ways to address or compensate for instability in school test scores, such as tracking the same students from year to year or using rolling averages of 2 or more years of test results.

- **Public education and communication.** Support efforts to educate teachers, school administrators, parents, and the public about the strengths and limitations of the major tests used in your state. Be careful not to oversell testing as a cure-all for problems of low achievement or failing schools.
Understanding the Strengths and Limitations of Tests

State and national policymakers are relying on large-scale tests more than ever to make high-stakes decisions about students and schools. (By large-scale tests, we mean the state tests and other standardized tests that students take once or twice a year, not the classroom quizzes or end-of-unit tests that teachers give regularly.)

States are using these tests to motivate students to study harder and hold teachers and administrators accountable for raising student achievement. Some states are also requiring students to pass tests before they can graduate or be promoted to the next grade. Throughout the country, parents, reporters, and other citizens are scrutinizing test scores to make judgments about the quality of schools in their community.

This dependence on tests will increase as states implement the No Child Left Behind Act. Under this new federal law, states must test students every year in grades 3 through 8 and once during high school, and use the resulting scores to help determine whether schools and school districts are doing an adequate job of educating students. In this environment, it's critical that policymakers, reporters, parents, and others understand both the strengths and limitations of tests.

Large-scale tests can do some things quite well. Critics of testing often downplay these advantages or fail to offer practical alternatives for evaluating the performance of large groups of students.

But tests also have limitations, as scientists who design and study tests are among the first to point out. Although they support testing as a valuable tool when properly used, expert groups, such as the Board on Testing and Assessment of the National Research Council, are often candid about the limits of testing and the dangers of test misuse (National Research Council, 1999). For example, these experts caution people not to use tests for purposes for which they were not designed, yet policymakers sometimes ignore this advice and expect tests to do more than they were intended to do.
What Tests Can Do Well

When large-scale tests are well designed and properly used, they can tell us a lot about what students know and can do. The benefits of these tests include the following:

Consistency

Tests can provide information that is more standardized and consistent from school to school or district to district than measures based on individual teachers’ judgments.

Report card grades, teacher-made tests, written evaluations, and other forms of classroom assessment can provide rich and useful information. But the meaning of a “B” grade or the value of a written evaluation can differ considerably from teacher to teacher or school to school. For many important decisions, policymakers and educators want information that is more uniform and dependable than what they can get from measures that rely on the standards of individual teachers. This is especially critical when results are being used to measure and compare student performance across classroom, schools, and districts or at different points in time.

A typical large-scale test goes through multiple rounds of research, development, and pilot testing to make sure it is reliable — in other words, that the scores it yields are consistent and stable across multiple test administrations. A large-scale test is also expected to be valid, which means that it measures what it claims to measure and leads users to draw accurate and meaningful conclusions about what students know and can do. Test designers also strive to make sure tests are fair in various ways; for example, the questions should not be systematically biased against a particular group of students.

Large-scale tests differ in how well they meet these technical criteria, but tests used for high-stakes decisions, as most state tests are, are expected to meet high technical standards. In general, tests that have been developed according to professional measurement standards will be much more consistent across different settings than most locally-developed assessments or measures.

Of course, other forms of assessment have their own advantages. At the classroom level, good teachers use a wide variety of methods to obtain a richer picture of students’ competencies and give students feedback. Teachers may observe younger children as they read aloud, watch how students approach a long-term project, or ask students to explain their thinking. Research has found that students can learn a lot from classroom assessments that give them immediate feedback about what they are doing right and wrong (National Research Council 2001) — something most large-scale tests don’t offer. Most teachers need better training in classroom assessment, however, to be able to do this well.

At the school, district, or state levels, alternative forms of assessment can measure certain kinds of complex knowledge or skills that aren’t effectively tapped by most large-scale tests. Examples include analyses of portfolios of
student work, observations of students as they do a science experiment, or formal exhibitions or performances, like a science fair or debate.

In other words, alternative and classroom assessments have their strengths, just as large-scale tests have their strengths. Often the best assessment choice is not a test or another measure, but a test and another assessment or measure.

**Comparative information**

Tests produce information that is comparable across different localities and can be used for various kinds of group comparisons.

In a decentralized education system like ours, curricula and instruction can vary considerably from district to district. Although measures like grades and teacher-made tests can be valuable within the classroom, they aren’t consistent enough to be useful for comparisons outside the classroom. A clear benefit of large-scale tests is their ability to produce comparable information across jurisdictions.

Test scores can be readily combined, or “aggregated,” and used to analyze and compare achievement across classrooms, grade cohorts, schools, or districts. These comparisons can help educators and policymakers decide where they need to target resources and technical support.

Test scores can also be broken down, or “disaggregated,” to compare the achievement of various subgroups of students — for example, by racial, ethnic, or income subgroups, disability status, or language minority status. Disaggregated test data can highlight achievement gaps between subgroups and spur states and districts to address them.

**Summary information**

Tests can provide valuable summary information about the performance of individual students or groups by subject, topics, and skills.

Although teachers generally keep close track of their students’ progress, it is not always easy to determine what a student knows or doesn’t know, or can or can’t do. Tests can help flesh out the picture by providing summary information about student’s knowledge and skills in a subject. If a teacher sees that many students in the class are having trouble summarizing the point of a reading passage, then the teacher has a better idea of where to focus instruction during the coming weeks. If a student knows that he or she got all the division problems right on a state math test, but missed all the questions involving statistics, then the student can devote more study time to statistics. This type of test information is most useful, however, when there is close alignment among the tests, the state standards, and the curriculum actually taught in classrooms.

States report different kinds of test information, but generally a score report will show the student’s scores in the major subjects tested, such as English/language...
arts and math, along with subscores for certain learning objectives within a subject, such as understanding word meanings in language arts. Most state score reports also show which performance category a student falls into, such as Advanced, Proficient, Basic, or Minimal (the number and names of the categories vary by state).

Some states report additional information showing whether the student has met the performance standards for his or her grade. Texas, for example, reports scores to parents and students in the form of a Texas Learning Index, which shows how far a student’s score falls above or below the passing standard in reading and math. The report also shows how many questions a student answered correctly for 12 learning objectives in math, such as use of multiplication to solve problems, and 6 objectives in reading, such as the ability to make inferences and generalizations (Texas Education Agency, 2002).

A key issue is whether state tests provide enough useful and timely information to help teachers identify student needs and adjust instruction accordingly. Some states provide little of this type of information, while others are relatively specific. In Indiana, for example, teachers receive a report showing each student’s score in English/language arts and math, each student’s performance category, a class average score, and the number of students above and below the state’s academic standard in each subject. The teacher report also indicates how many students have and have not mastered specific learning objectives within each subject. For example, a teacher can see which students need to work on sentence construction in writing and which need to focus on geometry in math (Indiana Department of Education, 2002).

In most states, there’s a significant lag time between testing and score reporting, so the feedback students and teachers receive is somewhat dated. Some states are piloting computerized testing, whereby students take the test on a computer and receive score reports almost immediately. This option is not yet ready for use with high-stakes testing, but it offers promise for the future.

**Efficiency and cost**

Tests are efficient and generally cost effective.

Tests can produce extensive information about students with a limited amount of testing time and for relatively lower costs than many other measures. Most kinds of test questions can be scored by machine, which cuts down on costs. Computerization of test data makes it possible to efficiently categorize, analyze, and report data in various ways to meet the needs of different audiences.

Many states are moving toward more open-ended test items, such as asking students to come up with their own answer or write a short essay. Although these kinds of items must be scored by hand (using standardized scoring criteria), they make it possible to conduct large-scale assessments of skills that aren’t well measured by multiple-choice items.
The Limitations of Tests

Because large-scale tests have a basis in science and report their results in numbers, many people think they are absolutely accurate. But as testing experts often note, tests are a blunt instrument. A test score is more like an estimate than an exact reading. Both individual and average test scores can fluctuate for reasons unrelated to learning or the quality of teaching.

Standard error of measurement

Every large-scale test has a “standard error of measurement,” similar to the margin of error in an opinion poll, which is intended to give people an idea of the test’s accuracy.

Test designers realize that tests aren’t precision instruments. When constructing a test, they try to take into account the score variations that might occur because of factors unrelated to learning, such as the sample of questions chosen (discussed below), a few lucky guesses, the student’s physical condition or state of mind, distractions during testing time, or scoring errors. To make allowances for these and other external factors, test developers compute a standard error, which shows how much a student’s test score would be likely to vary, either up or down, if the student took the test again and again (assuming no learning took place in between). Standard error can also be viewed as an estimate of how much a student’s score on a given day will vary from his or her “true” score — in other words, the hypothetical score the student would receive if the test were perfectly accurate. The standard error is different for different tests.

Consider an example from the English/language arts test of the California High School Exit Exam, which has a score scale of 250 to 450 (California Department of Education, 2002). If a student earns a score of 410 on this test, the test developers are 95% confident that the student’s true score lies between 434 and 386 — a standard error of plus or minus 24 points. The test developers are only 68% confident that the student’s true score falls within the narrower range between 422 and 398 — a standard error of plus or minus 12 points. If that same student scored 400 on a retest a short time later, it wouldn’t necessarily mean that his or her achievement had dropped, because the 10-point difference from the first score would be within the range of standard error at either confidence level.

How imprecise are test scores? The answer varies for different tests. But according to one study of a nationally normed test, two students with the same “true” achievement at the 45th percentile are likely to score within 10 percentile points of each other less than half the time (Rogosa, 1999).

When test scores are reported as a single number, people can easily forget about standard error. To rectify this, Wisconsin, North Carolina, and other states display the standard error in their score reports as a band or line within which a student’s true score may fall.
Many state policies, however, call for a clear cutoff score, regardless of standard error. Examples include the score necessary to pass a state exit exam required for graduation, or the cutoff between the Proficient and Basic performance categories on a state accountability test. If the passing score is 350 on the California English/language arts exit exam, for instance, a student who scores 352 will pass, while a student who scores 348 will fail, even though there’s no meaningful difference between their levels of achievement. In these situations, states should be sure the cutoff score is tied to justifiable standards for what students should learn in a particular grade. They should also provide multiple chances for retesting and additional ways to measure competency. (For more information about exit exams, see the Center’s report, State High School Exit Exams: A Baseline Report.)

The concept of measurement error also applies to average test scores for a classroom, school, or larger unit. For example, if a state uses average test scores to identify low-performing schools or rate schools based on performance, one school could be labeled as low-performing while another is deemed adequate, even though there’s no meaningful difference between their average scores.

The National Assessment of Educational Progress (NAEP) attempts to report its state-by-state average scores in ways that will discourage people from drawing false conclusions based on small differences in scores among states. Some NAEP charts use shaded bands to show which score differences between states are significant and which are not (NAEP, 2000).

**Sampling variation**

A test is a sample of all possible questions that could be asked about a subject. A test is also a sample of a student’s behavior at a single point in time.

Sampling variation is a key reason why test scores may fluctuate. The questions on a test are a sample of all the knowledge and skills in the subject being tested. No matter how well-designed, a test that lasts a few hours, as state tests typically do, can’t possibly address all the important topics, concepts, or skills in broad subject like math. For instance, the Massachusetts state test in 8th grade math includes 48 test items; the North Carolina end-of-course test in 8th grade math has 80 items. Neither test can cover more than a small portion of everything the state expects 8th graders to learn about math.

Test developers try to ensure that the items selected for a test cover a representative sample of important knowledge and skills in the subject being tested, and that different versions of the same test (developed for security reasons) are parallel in content and level of difficulty. If a test is well-designed, its items will be a reasonable sample of important knowledge and skills, and teachers can use its scores to make sound inferences about how well a student understands the subject. Nevertheless, there will always be students who would have scored higher if a particular test version had included a different sample of questions that happened to hit on topics they knew well. When the number of questions
is small, users of test scores must be careful not to conjecture too much about a test taker’s understanding of the broader subject.

Some researchers have raised concerns about how samples of items are typically chosen for norm-referenced tests (Popham, 1999) — tests designed to compare and rank a student’s performance along a curve. In these kinds of tests, a certain percentage of students will always score high, a certain percentage will always score low, and the majority will score in the middle. In order to spread out scores along a curve, test designers must choose questions that do a good job of differentiating among students, especially those at the middle range of achievement. Questions that most students can answer correctly — or that almost no one can answer correctly — don’t accomplish this goal, so these questions are often discarded from the item sample after pilot testing. The questions that do the best job of differentiating are those that will be answered correctly by about half of all test takers. This is why most questions on a norm-referenced test are of medium difficulty. But by skimping on questions that can be answered correctly by most students, the test may omit important knowledge and skills — the very content we want students to learn. Perhaps the reason why most students answered a pilot question correctly was because it covered material their teachers had emphasized.

When tests are constructed this way, it works against the goal of raising all students’ scores to a proficient level, because questions that a high percentage of students get right may be dropped from the test. Similar concerns have been raised about procedures for selecting items for state tests that are essentially criterion-referenced (designed to rate student performance against absolute standards) but also have some norm-referenced features (Haney, 2002).

A test is also a one-time sample of a student’s behavior. On any given day, a variety of unusual factors — from a headache, to a jackhammer or barking dog outside the school, to an argument with a parent that morning — could negatively affect the student’s performance. If the test had been given on another day, maybe these external factors wouldn’t have been present.

Score inflation

Teaching to the test can raise scores without students actually learning more.

In today’s climate of high-stakes testing, it’s not uncommon for teachers to drill students in practice questions that look a lot like real test questions or to emphasize the specific topics and skills that are most likely to appear on an important test. Teachers also coach students in test-taking skills, such as how to narrow down answers in a multiple-choice question. These practices — commonly called “teaching to the test” — help familiarize students with the content and format of a high-stakes test, which helps them do better on the test. But the students haven’t necessarily learned more about the broader subject being tested or acquired knowledge and skills they can apply in a non-testing situation. If the students took a different test on the same subject with a completely different sample of questions, they probably wouldn’t do as well.
Several studies have asserted that state tests with high stakes are particularly susceptible to score “inflation” — in other words, the scores on the high-stakes tests go up faster than scores on other assessments given at approximately the same time in the same subjects (Amrein & Berliner, 2002; Klein et al., 2000; Koretz & Barron, 1998; Stecher & Hamilton, 2002). Researcher Robert Linn (2000) found that average scores on state accountability tests tend to rise, sometimes dramatically, every year for the first three or four years after a new test is introduced, then level off or even decline. He surmised that it takes teachers and students a few years to develop a feel for the format and content of a new accountability test, but once they do, any gains attributable to growing familiarity with the test, instead of increased learning, disappear.

A related type of score inflation is the Lake Wobegon effect, named after broadcaster Garrison Keillor’s fictional town where “all the children are above average.” In the late 1980s, researcher J. J. Cannell observed that every state that reported statewide test scores claimed that its scores were above the national average, even though some states were doing poorly according to other indicators of achievement (U.S. Office of Technology Assessment, 1992). This was happening mostly because states were using norm-referenced tests with outdated norms. Before a norm-referenced test is released, it is administered to a nationally representative sample of students to determine the “norms” — the distribution of scores on a curve for the nation. Sometimes the same norms are used for several years, so that a current group of students is being compared with the norming group from several years earlier. If, during the intervening years, students become more familiar with the test content, or if their average achievement rises, then well over half the students could score above the 50th percentile.

Changes in the test-taking population

Yearly changes in student population can cause fluctuations in the average test scores of a class or a school.

Under the No Child Left Behind Act, states make judgments about a school’s annual progress by comparing the average test scores of students in the grade level being tested with the average scores of students who were in that grade the previous year — in other words, by comparing this year’s group of 5th graders with last year’s group of 5th graders. But the law fails to consider the possibility that average test scores for a grade cohort can fluctuate due to changes in the mix of students or other factors unrelated to learning or the quality of teaching.

As teachers already know, each year’s class is like a “random draw,” as researchers Thomas Kane and Douglas Staiger describe it (2001), containing a unique mix of students in terms of economic, linguistic, and racial/ethnic backgrounds, achievement levels, personalities, and behavior, to list just a few characteristics. Countless factors can change the composition of the student body from year to year. A school could experience an influx of immigrants, bringing in more English language learners. A class could have a higher num-
ber of students with severe disabilities whose competencies may not be well measured by large-scale tests. Or a school could be making a greater effort to comply with federal requirements to include students with disabilities in testing — students who were often excluded in previous years. The loss of a major manufacturer, a rise in the dropout rate, an exodus of students from neighborhood schools to charter schools, the construction of an upscale housing development in the neighborhood — any of these could significantly change the nature of the test-taking group.

Changes in the test-taking population matter because studies have shown a correlation between some student background characteristics and group test scores. (Of course, these broad correlations don’t justify drawing stereotypical conclusions about any individual student; high- and low-achieving students come from all backgrounds.) Most testing systems are based on the assumption that tests measure what students have learned in school. Tests do measure learning, but they also measure “what students bring to school,” as researcher W. James Popham has noted (1999). Family income is often among the most reliable predictors of how a student will score on a given test. This is not surprising, since students from higher-income families have greater access than poor children to out-of-school experiences that promote learning.

If the number of test takers is large, these differences in background characteristics tend to average out. But in a relatively small grade cohort (fewer than 100 students, according to research by Haney, 2002), annual changes in class composition can produce wider fluctuations in average scores, because each student’s score has a greater impact on the average. According to researchers Kane and Staiger (2001), the Massachusetts school that made the greatest leap in state test scores in 1998-99 had only 26 students taking the state test in 10th grade, a small enough group that dramatic yearly swings were almost a given. With the average elementary school containing only 68 students per grade, this kind of score instability is not unusual. It also applies to schools with high mobility or very diverse enrollments.

Test score averages can be especially unreliable for racial, ethnic, income, and other subgroups, which tend to be even smaller in size than a grade cohort (Elmore, 2002). This is an area where tension exists between the limitations of large-scale tests and the well-intended goals of the No Child Left Behind Act. The law requires schools to demonstrate progress for all subgroups or to be targeted for interventions. Although states can exempt very small subgroups, the typical subgroup is still small enough that yearly changes in composition could produce significant score swings and possibly cause some schools to be misclassified as low-performing.

**Volatility of average scores**

Volatility in average test scores due to external factors could make it difficult for schools to show the continuous progress demanded by the No Child Left Behind Act.
As already mentioned, aggregate scores for a classroom, grade, or school tend to fluctuate considerably from year to year due to changes in student population, sampling variations, and other external factors. Extraordinary school circumstances, such as a teacher strike or an unusually disruptive group of students, can also cause a blip in average scores (Linn, Baker & Betebenner, 2002).

Furthermore, when the difference is calculated between one year’s average score and the next year’s, the net gain or loss is a less reliable calculation than either year’s scores, because it is subject to standard errors of measurement in both the baseline and follow-up years (Linn, Baker & Betebenner, 2002). The No Child Left Behind Act permits states to aggregate up to three years of data when they determine whether schools are making adequate progress, but this will not totally eliminate the problem of unstable scores.

Analyses in several states have not only confirmed this instability in average school scores, but have also found that a school that scores especially high in one year is likely to show a decline the next year. For example:

- Kane and Staiger (2001) estimated that more than 70% of the year-to-year variations in average test scores for a given school or grade can be attributed to external factors, rather than educational factors. This helps to explain why there are few repeaters when states base rewards on test scores. Between 1997 and 2001, North Carolina gave out 101 awards to schools for high or improved achievement. These awards were won by 90 different schools; only 9 schools won twice, and only 1 school won three times.

- Linn, Baker, and Betebenner (2002) looked at four consecutive years of 4th grade test results in Colorado to see how many schools made significant gains from year to year. (Based on the timetable for improvement in the No Child Left Behind Act, the study defined a significant gain as a 1-point increase or better in the percentage of students scoring at or above the proficient level on the state test). Many schools that showed significant gains in one year failed to improve the next year. Only 1 school in 20 met this improvement target for three years in a row.

- According to Walt Haney (2002), most Massachusetts schools that showed a score gain of at least 10 points between 1999 and 2000 posted declines for the next year. After analyzing the scores of all Massachusetts schools during two consecutive 2-year testing periods, Haney concluded that a school that showed a test score gain for one testing period was more likely than not to do worse in the next period, and vice versa.

Even with the provisions of the Act allowing states to look at multiple years of data, volatility in average scores is still likely to result in some schools being unfairly identified as low-performing.
Conclusion

Tests are an indispensable tool in the measurement toolbox. Good tests can provide consistent, comparable, and useful information about student achievement not easily obtained through other means. But tests aren’t perfect. Several factors unrelated to learning can cause test scores to fluctuate at the individual or aggregate levels. Consequently, test scores don’t always mean what many people think they mean.

Some limitations of existing tests may eventually be addressed by new forms of assessment being developed by cognitive scientists and others (National Research Council, 2001). For example, most existing tests don’t tell us much about why a student missed a question. But a new generation of computerized “intelligent tutors” can track the steps students take to solve a problem, analyze correct and incorrect patterns of reasoning, and help the student identify more promising strategies. “Progress maps” can show teachers where a student’s current level of understanding fits along the pathway from novice thinking to expertise in a field like algebra, and can help teachers devise learning tasks that will move the student forward. Other new modes of assessment can shed light on how students organize knowledge in their minds, give students immediate feedback as they take a test, and provide better diagnostic information for teachers. These tools are not yet ready for wide implementation, but they demonstrate what it is possible.

In the meantime, state and national policymakers have a responsibility to design testing and accountability systems that draw on the strengths that tests offer, but also address their limitations. How test results are reported is a crucial part of responsible test use. Effective reporting helps people understand what scores mean but discourages them from jumping to wrong conclusions about student achievement or school performance.

When testing was a low-stakes enterprise, issues of appropriate test use were often neglected. The pressure of new federal requirements could have a positive impact on testing practice by encouraging educators and policymakers to pay more attention to expert advice about appropriate test use. States can build a good accountability system now, by combining well-designed tests with other measures. As state and national leaders gain more experience with high-stakes testing, deepen their understanding of what tests can and can’t do well, and bring new testing technologies into wider use, they can make their systems even better in the future.
Credits

This publication was researched and written by Nancy Kober, a freelance writer and consultant to CEP, with assistance from Naomi Chudowsky, a testing expert and consultant to CEP, and Diane Stark Rentner, CEP’s deputy director.

The Center on Education Policy thanks our funders who supported the writing and publication of this report and the other activities of CEP through their generosity.

Jack Jennings
Director

Sources


Haney, W. (2002). Ensuring failure: How a state’s achievement test may be designed to do just that. Education Week, 10 July 2002, 56, 58.


© 2002 Center on Education Policy

Center on Education Policy
1001 Connecticut Avenue NW, Suite 522
Washington, D.C. 20036

phone 202-822-8065
fax 202-822-6008
email cep-dc@cep-dc.org
web www.cep-dc.org