MID-CONTINENT RESEARCH FOR EDUCATION AND LEARNING

Policy Brief

McREL

April 2000

High-Stakes Testing: Trends and Issues

by Anne Lewis

Every student deserves to be in an education system that expects all students to achieve at the highest levels possible. Ideally, the system provides the resources for all students to learn optimally and then insists on accountability for teaching and learning. These are the essential elements of the current standards-based reform movement — expect much, provide the means to meet expectations, and check on results.

When the assessments of progress lead to consequences — for students, teachers, and schools — the stakes are high. The results from testing often determine if a student is to be promoted or is able to graduate from high school. Four states and the Virgin Islands use assessments to hold teaching staff accountable for results (CCSSO, 1999). Twenty-seven states use assessments (or are planning to within the next few years) to rate all schools, or to identify the lowest performing schools statewide (Quality Counts, 1999).

Policymakers and educators must grapple with assessment dilemmas. For example, while a state may be reluctant to impose high-stakes testing, local districts might be under pressure to end social promotion or guarantee the skills and knowledge levels of high school graduates. To balance these competing priorities, it is important to seek answers to some general questions about assessment.

Is high-stakes testing ultimately an effective tool to encourage students to achieve their best and teachers to provide the instruction that will assure high levels of learning? Can a state follow a policy of encouraging or requiring high-stakes decisions on the basis of tests while satisfying local demands for deciding what students should know and how to measure their learning? What are the issues involved in high-stakes testing that policymakers, educators, and parents need to know before they insist upon actions such as eliminating social promotion?

Much of the fanfare accompanying "get tough" policies for failing students and schools has tempered in recent months as policymakers themselves begin to see the complexity of raising student achievement dramatically in a short period of time.

Repercussions of high-stakes testing

Much of the fanfare accompanying "get tough" policies for failing students and schools has tempered in recent months as policymakers themselves begin to see the complexity of raising student achievement dramatically in a short period of time. Faced with the possibility of large percentages of students being held back because of their failing scores on standardized tests, some states and school districts are rethinking their original policies. Primarily, they are either delaying the time line for consequences of test failure or lowering the cut-off scores on tests.

Many of the problems resulting from these new policies arise because state and school district officials want to be assured — as quickly as possible — that students are making academic progress. Moreover, certain federal programs require policymakers to adopt new assessment systems. Because of mandates and anxieties about student achievement, policymakers at state and local levels

often make crucial decisions about the use of tests without being fully aware of the evolving nature of assessment systems and without well-developed, cohesive tests or assessment systems in place. Education assessment systems are getting better, but they are not as good as policymakers assume — or students and schools deserve. For example, educators still do not know enough about designing assessment policies appropriate for students with disabilities or for students from limited English backgrounds.

Education assessment systems are getting better, but they are not as good as policymakers assume — or students and schools deserve.

High-stakes testing policies also have engendered a variety of complaints from parents, teachers, and civil rights groups — complaints that are increasingly being played out in the courts. And recent controversies go beyond civil rights issues. They concern using tests to make important decisions about students when the tests may lack reliability and validity, when there are errors in scoring, or when the tests are used for purposes for which they were not designed.

More than any other issue of assessment, high-stakes testing has the potential to have very dramatic effects on students as well as on parents and schools. It is the purpose of this policy brief to examine this critical issue and to provide guidance for those who must make decisions about how assessment results are used.

What is high-stakes testing?

Assessments of students take place regularly — every day in most classroom situations. Similarly, schools and districts are becoming accustomed to seeing their test scores made public through state reporting systems. High-stakes testing, however, has special characteristics. In general, the term refers to any assessment used for accountability with significant consequences. For students, that means test results that lead to very important decisions — promotion/retention, access to specific programs, or qualification for a high school diploma (and special honors diplomas). For example, in 1998, over 4,000 eighth-graders in Chicago who failed tests in reading and mathematics were required to attend six weeks of summer school — and only two-thirds of those students were then promoted to ninth grade (Johnston, 1999).

High-stakes testing, applied to schools and/or districts, determines which are to receive awards for high performance or extra investments because of low scores. In the case of low scores, schools stand to lose accreditation, be reconstituted, or even closed.

The use of test scores to hold teachers and principals accountable is rare but increasingly discussed in policy circles. In Tennessee, for example, principals receive information on the "value added" by each teacher in a school to children's academic achievement, although state law prevents the information from being used for teacher evaluations. In Colorado, Denver Public School teachers recently agreed to participate in a demonstration project which includes student test scores as part of a "pay-for-performance" program (Illescas, 1999).

In an analysis of the legal implications of high-stakes assessment, S.E. Phillips (1993) of Michigan State University listed these characteristics of high-stakes assessment:

- public scrutiny of individually identifiable results,
- a significant gain in money, property, or prestige for those with positive assessment results,
- considerable pressure on individuals or institutions to perform well or to raise scores,
- a perception that significant individual decisions are being made based on a single imperfect piece of data over which the affected entity has no input or control, and

- complex and costly security procedures designed to ensure maximum fairness for all who are assessed. (p.1)

These characteristics, Phillips added, "suggest that a high level of anxiety is associated with the assessment" (1993, p.1). Its results and decisions based on the results could deprive an individual or an institution of something valuable. For example, one of the claims cited by the plaintiffs in the recent federal court case in Texas protesting state exit exams contended that a high school diploma is "property," and, thus, students who were denied diplomas on the basis of the test lost property due to discrimination (*GI Forum et al. v. Texas Education Agency et al.*, 2000).

Using high-stakes testing bears careful scrutiny

Political leaders often view assessment much the same as the general public — students ought to be able to pass tests that measure necessary skills, and they should not advance until they can show what they know through a passing test score. Moreover, many believe that rigorous testing policies, such as high-stakes testing, encourage teachers and students to get serious about teaching and learning. Politicians point to rising test scores after adoption of policies that impose consequences on students and schools for low performance.

Even the severest critics of high-stakes testing acknowledge that assessments are necessary for a variety of purposes — public accountability, diagnosis of student strengths and weaknesses, and evidence for teachers and parents that students are learning what they should.

Would that assessment were that simple. Even the severest critics of high-stakes testing acknowledge that assessments are necessary for a variety of purposes — public accountability, diagnosis of student strengths and weaknesses, and evidence for teachers and parents that students are learning what they should. Where they disagree about assessment is when a single test is used to fulfill more than one of these purposes and when that test becomes the basis of decisions that significantly affect what happens, academically, to a student in school.

Relying on a single test for critical decisions is fraught with issues that confront policymakers and educators. Recent research and policy statements about high-stakes testing have warned against reliance on a single test for important decisions. For example, the newly revised *Standards for Educational and Psychological Testing*, jointly developed by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1999), states in Standard 13.7:

> In educational settings, a decision or characterization that will have a major impact on a student should not be made on the basis of a single test score…. (p. 146)

Charged by Congress in 1998 to recommend appropriate assessment methods, the Board on Testing and Assessment of the National Research Council recommended several principles for test use, later elaborated on by the principal researchers Jay Heubert and Robert Hauser in their 1999 book *High Stakes: Testing for Tracking, Promotion, and Graduation*:

- The important thing about a test is not its validity in general, but its validity when used for a specific purpose. Thus, tests that are valid for influencing classroom practice, "leading" the curriculum, or holding schools accountable are not appropriate for making high-stakes decisions about individual students....

- Tests are not perfect. Test questions are a sample of possible questions that could be asked in a given area. Moreover, a test score is not an exact measure of a student's knowledge or skills....

- An educational decision that will have a major impact on a test taker should not be made solely or automatically on the basis of a single test score. Other relevant information about student knowledge and skills should also be taken into account.

- Neither test scores nor any other kind of information can justify a bad decision. Research shows that students are typically hurt by simple retention and repetition of a grade in school without remedial and other instructional support services.... (p. 3)

The Improving America's Schools Act of 1994 (Title I), which sets the mode for state and district assessment systems, calls for the use of multiple measures of student achievement. However, another report from the National Research Council — *Testing, Teaching, and Learning: A Guide for States and School Districts* (Elmore & Rothman, 1999) — points out that states and districts are continuing to use one test in creating accountability. "Schools get the message that they have to raise scores on that test in order to earn rewards or avoid sanctions," said the report. "Using multiple measures could encourage schools to focus less on a single measure and more on improving achievement generally" (p. 95).

"Using multiple measures could encourage schools to focus less on a single measure and more on improving achievement generally."

Considering the strong cautions from testing experts about high-stakes testing, policymakers and educators need to ask incisive questions about three major areas of concern: psychometric issues, the impact of teaching on student behavior, and potential legal challenges.

Scoring Error Sends Students to Summer School

Among the issues of concern with high-stakes testing are simple human or mechanical errors in scoring. Last year nearly 8,700 students in New York City spent five weeks of their summer break in remedial reading and math classes due to just such an error.

Some of the items on the New York City's 1999 Citywide Tests, it turned out, were incorrectly calibrated. This skewed the students' raw scores to the national rankings and, as a result, students who had scored slightly above the passing line were calculated to have scored at or below it — and were required to attend summer school.

One week into the start of the new school year, embarrassed school officials admitted the mistake. They, along with representatives from test maker CTB/McGraw-Hill and Mayor Rudolph Giuliani, scrambled to apologize. But there was another problem. Of the 8,668 "accidental" summer students, only 5,176 had attended summer school and passed the required test allowing them to be promoted to the next grade. Of the 3,492 remaining, 1,168 had attended the program but failed the test, and the other 2,324 either had not attended the program or had not shown up for the test.

To resolve the issue, conferences were held with parents and the students who had mistakenly been assigned to summer school to give students the option of moving to the next grade.

SOURCE: Kerry A. White for *Education Week*

Assuring tests are psycho-metrically and technically sound

Researchers and measurement experts must ask themselves major questions to ensure that their testing programs are sound:

- Is the test valid for the purposes for which it will be used? Does it measure what it says it does; that is, does it sufficiently reflect the content being assessed and produce results that support whatever decisions are made on the basis of the test?
- Is the test reliable? Are test results reproducible; that is, do differences in test scores consistently reflect real differences in student knowledge or are they the result of other factors such as scoring errors, bias of the raters (as in assessment of writing samples or portfolios), or differences in how the test is administered?
- Is the test aligned to the curriculum? Does the test adequately reflect district and state content and performance standards? (The validity of TAAS, the Texas state exit exam, was challenged because of the low correlation between student grades and test scores.)
- Have the cutoff scores for the testing program been set fairly, accurately, and with sufficient research to justify them?
- Is the test fair? Does it provide accurate information on all students as to what they know? Does it provide information that leads to accurate conclusions about identifiable subgroups of students?
- Does the test fit into a total assessment system designed to provide complementary information for parents, classroom teachers, administrators, the general public, and policymakers, or is it the only source of information for all stakeholders?
- Does the testing system draw from the latest research on appropriate testing procedures for students with disabilities (who must now be assessed the same as regular students) and for students with limited English proficiency?
- Does the test provide longitudinal information for policymaking that shows achievement progress and/or problems over time, keeping the content and format consistent enough to allow comparisons?
- Does the testing program make appropriate provisions for multiple test forms and other issues of test security involving printing, distributing, collecting, and scoring tests? Are analyses conducted to monitor unusual patterns of results?

Determining the impact on teacher and student behavior

In addition to constant technical checks on a high-stakes testing system, policymakers need to establish ongoing research and evaluation of the impact on teaching and on students by asking:

- Does use of the test result in disparate effects in retentions or remedial classes for subgroups of students, such as racial minorities?
- Does the test encourage teachers to "teach to the test" and, consequently, narrow the curriculum and instruction? (Another indicator to watch is the incidence of questionable test preparation given to students, or even cheating, due to high-stakes pressure on schools.)
- Does the test create imbalance in the curriculum because of the timing of tests? (In Kentucky, for example, teachers overemphasized during a school year the subject to be tested and neglected other areas of the curriculum [Stecher, 1999].)

In addition, educators must make certain that high-stakes testing does not indirectly contribute to a higher dropout rate by discouraging students — particularly those who are at risk of failing — from even bothering to try. Schools must have plans in place to identify failing students early and to provide the instruction and support they need to be successful when they take the test.

Will the test pass legal review?

The January 2000 court decision in Texas ruling in favor of the use of the Texas Assessment of Academic Skills (TAAS) to determine high school graduation undoubtedly will have an impact on the two dozen states that now (or soon will) require students to pass a test for a diploma. In the Texas case, a civil rights group argued that schools with predominantly African-American and Latino students, particularly in poor communities, are less likely to adequately prepare students for the test and that the structure of the test adversely affects minority students. In his decision, U.S. District Judge Edward Prado acknowledged that the TAAS *does* adversely affect many minority students — some 7,500 students annually are denied their diploma after failing the test multiple times — but that the state education agency had "sufficiently demonstrated the test's educational necessity" (Gladfelter, 2000).

Various judicial decisions on the same test may come to very different conclusions. At this point there is no concrete case law to provide firm guidance for policymakers, however, there are common questions asked in the courts. The Office for Civil Rights in the U.S. Department of Education is formulating similar questions (Coleman, 1998) to help guide schools, districts, and states as they consider adopting high-stakes testing. These questions take into consideration several key issues:

- Educational justifications that support high-stakes testing (e.g., improving the quality of education; insuring that high school graduates are competitive nationally).
- Any history and/or continuing effects of racial segregation or discrimination that may unfairly influence students' ability to succeed on the test.
- Academic expectations of the new test. (Courts are leery of tests that fundamentally alter expectations surrounding performance.)
- The period of time allowed for students to learn the material being tested and for parents and students to become familiar with the new requirements.
- The process by which new test requirements are implemented (e.g., tutorial help available; multiple opportunities to take the test).
- Alignment between the test and the curriculum and instruction.

Accountability for whom?

High-stakes testing tells the public that students and schools will be accountable to its aspirations for them and its investment in them. However, accountability is everyone's job. As Elmore and Rothman (1999) conclude:

> Accountability should follow responsibility: teachers and administrators — individually and collectively — should be held accountable for their part in improving student performance. Teachers and administrators should be accountable for the progress of their students. Districts and states should be accountable for the professional development and support they provide teachers and schools to enable students to reach high standards. (p. 97)

Part of high-stakes testing policies ought to be the provision of sufficient resources for educators to adapt to higher standards and integrate well-conceived assessments into instruction and school improvement.

Thus, part of high-stakes testing policies ought to be the provision of sufficient resources for educators to adapt to higher standards and integrate well-conceived assessments into instruction and school improvement. Without sufficient knowledge about the roles of assessment in an overall plan for quality teaching and learning, teachers and administrators could make inappropriate use of high-stakes tests. The effect on students would be the opposite of what

families, communities, and policymakers want.

However, as Heubert and Hauser also point out, "when used appropriately, high-stakes tests can help promote student learning and equal opportunity in the classroom by defining standards of student achievement and by helping school officials identify areas in which students need additional or different instruction" (1998, September).

Anne Lewis is a national education policy writer.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999) Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Bond, Linda (1999, March). High-stakes assessments challenges for education policy makers. Concerning Assessment. (1–4).

Coleman, Arthur L. (1998). Excellence and equity in education: Moving from promise to reality. Prepared remarks before the National Research Council/National Academy of Sciences Board on Testing and Assessment Colloquium, Washington, DC.

Council of Chief State School Officers (1999). Annual survey of state student assessment programs (Vol. 2), 209-216. Washington, DC: Author.

Elmore, Richard & Rothman, Robert (Eds.). (1999). Testing, teaching, and learning: A guide for states and school districts. Washington, DC: National Academy Press.

Edwards, Virginia B. (Ed.). (1999). Quality Counts '99: Rewarding Results, Punishing Failure. Education Week, 18(17), 90.

GI Forum et al. v. Texas Education Agency et al., No. SA-97-CA-1278-EP (U.S. District Court, Western District of Texas, San Antonio Division). From plaintiff's complaint filed October 14, 1997. Decision rendered January 7, 2000.

Gladfelter, Hannah R. (2000, January 10). Judge upholds Texas test against bias claims. Education Daily, 1.

Heubert, J.P., & Hauser, R.M. (1999). High stakes: Testing for tracking, promotion, and graduation. Washington, DC: National Academy Press.

Heubert, J.P., & Hauser, R.M. (1998, September). High stakes: Testing for tracking, promotion, and graduation. Presentation materials from the national conference of the Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles, CA.

Illescas, Carlos (1999, September 17). DPS approves teacher contract, pay for performance plan. The Denver Post, p. B-09.

International Reading Association (1999, May). High-stakes assessments in reading: A position statement of the International Reading Association. Newark, NJ. Author.

Johnston, Robert C. (1999). Turning up the heat: Creating accountability for students. Quality Counts '99: The Challenges to Accountability, 18(17), 53–58.

Phillips, S.E. (1993). Legal implications of high-stakes assessment: What states should know. Oak Brook, IL: North Central Regional Educational Laboratory.

Stecher, Brian (1999, September). Research presented at the national conference of the Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles, CA.

White, Kerry A. (1999, September 22). Test company apologizes for NYC summer school mixup. Education Week, p. 8.

This publication is based on work sponsored wholly, or in part, by the Office of Educational Research and Improvement (OERI), U.S. Department of Education, under contract number RJ96006101. The content of this publication does not necessarily reflect the views of OERI, the department, or any other agency of the U.S. government.

Tim Waters, executive director & president
Jana Caldwell, director of communications
Shae Isaacs, editor
Jeanne Deak, graphic designer

Mid-continent Research for Education and Learning
2550 South Parker Road, Suite 500
Aurora, Colorado 80014
phone 303.337.0990
fax 303.337.3005
Visit us at our Web site: www.mcrel.org

ASSESSMENT INFORMATION FROM McREL FOR EDUCATION POLICYMAKERS!

Mid-continent Research for Education and Learning
2550 S. Parker Road, Suite 500
Aurora, CO 80014