

CRESST POLICY BRIEF

National Center for Research on Evaluation, Standards, and Student Testing

www.cse.ucla.edu

STANDARDS-BASED ACCOUNTABILITY TEN SUGGESTIONS

Robert L. Linn

History has shown that testing is a popular instrument of accountability and reform for a number of key reasons, including:

1. Tests are relatively inexpensive.

Compared to changes that involve increases in instructional time, reduced class size, training and attracting better teachers, assessment is very low-cost.

2. Testing changes can be implemented relatively quickly.

Other school reforms may take years to implement, and it may take even longer to know if they have improved schooling.

3. Test results are visible and draw media attention.

Poor results in the first year of a new testing program are usually followed by increasing scores in subsequent years, giving the appearance that schools are improving.

4. Testing can create other changes that would be difficult to legislate.

Research has shown that state- or district-level testing and assessment requirements motivate changes in curriculum and teaching at the school and classroom levels. It is much more difficult to directly legislate changes in the classroom.

Unfortunately, when tests are used to make major decisions about schools and students, these attractive features frequently result in unexpected problems. Test results may be incomplete or misleading, resulting in poor policy decisions. Nevertheless, the policy need for rapid information about student progress and school quality ensures a continued high interest in educational testing.



Robert L. Linn is co-director of the National Center for Research on Evaluation, Standards, and Student Testing and Distinguished Professor of Education at the University of Colorado at Boulder. He is the current chairperson of the National Research Council's Board on Testing and Assessment.

UCLA Graduate School of Education & Information Studies

STANDARDS-BASED ASSESSMENT SYSTEMS

A key feature of current school reform efforts is the creation of educational standards, with the federal government encouraging states to develop challenging content and performance standards. Standards-based assessment systems have quickly become a central part of many state reform programs, led by states such as Kentucky and Maryland. Other states, including Colorado and Missouri, are in the midst of implementing their own standards-based assessments. Already we have found that these systems confront the same challenges as earlier assessment programs plus a few new ones. For example:

1. Educational standards at the national, state, and district levels are often inconsistent.

Reviews of state content standards (Education Week, 1997; Lerner, 1998; Olson, 1998; Raimi & Braden, 1998) show that state content standards range from very strong to very weak. Different raters oftentimes give different ratings to the same standards, further contributing to the problem.

2. How standards are formulated and measured makes a difference.

The choice of "what" is measured and the quality of the standards and assessments are both important. Table 1 below reports important differences in student performance in the subjects of geography, history, mathematics, and reading as measured by the National Assessment of Educational Progress, the nation's report card.

In Table 1, why are only 9% of female students reaching the proficient level in history but 43% reaching the proficient level in reading? While the differences may indeed be differences in performance, it is much more likely that they are due to how the standards were formulated or to the accuracy of the assessments in measuring their respective subjects.

An assessment only in geography would show more males (32%) than females (22%) at the proficient level while the reverse would be true of an assessment only in reading, with males 29% proficient and females 43% proficient. Further, choice of different combinations of the four tests could produce overall results that were nearly equal for males and females or results favoring one group over the other. The choice of what is measured can also alter the apparent differences in performance of racial/ethnic groups or of groups formed on the basis of other characteristics.

3. Who's included or excluded in testing can produce different results.

Driven by Title I requirements, standards-based reform emphasizes the inclusion of both special needs students and English language learners in large-scale testing programs. Testing provides important information to policymakers, educators at all levels, and to parents on how all children are doing. However, inclusion can be taken to an extreme. For example, testing students in a language they don't understand will produce inaccurately low test scores. Excluding too many students, on the other hand, will produce inflated scores. The challenges of meaningful inclusion of all students are difficult, but essential for a credible assessment system.

Table 1
Performance Differences on NAEP
by Subject and Gender*

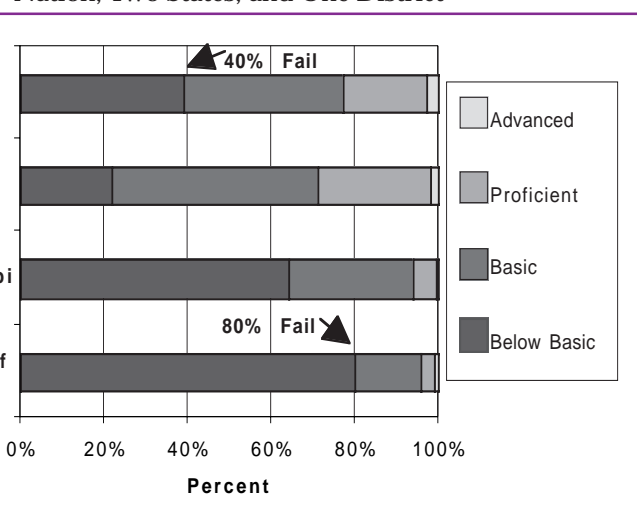
Subject	Males	Females	Difference (M-F)
Geography (1994)	32	22	10
History (1994)	12	9	3
Mathematics (1996)	18	14	4
Reading (1994)	29	43	-14

*Percentage of students at or above the National Assessment of Governing Board Proficient Level, Grade 12.

N
Nation
Iowa
Mississippi
District of Columbia

Figure 1

NAEP Grade 8 Achievement Levels for the Nation, Two States, and One District*



*Based on Reese et al., 1997.

4. Holding all students to the same high standards will result in unacceptably high retention and failure rates.

Figure 1 shows that nearly 40% of American students did not reach the basic level on the 1996 8th-grade mathematics NAEP test. Are we as a nation prepared to fail or retain as many as 40% of our students nationally or 80% in some districts? To do so would result in major political and legal challenges.

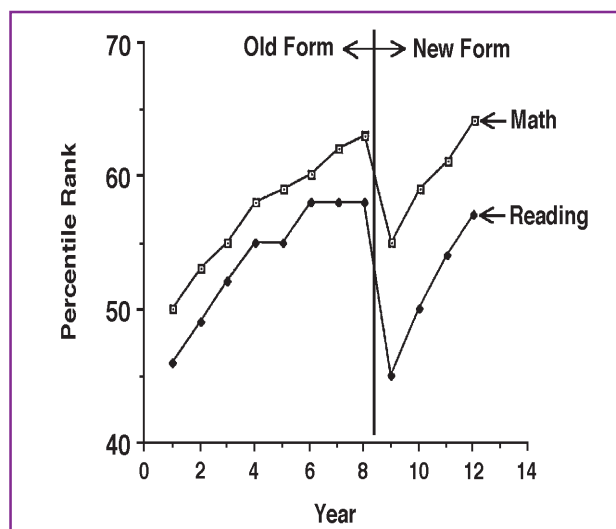
5. Gains in scores do not necessarily signal true improvements.

Research has continually shown that increases in scores on newly implemented tests reflect factors other than increased student achievement. Increases are often a result of teachers teaching to the new test or the use of old test norms (see Figure 2). Standards-based assessments do not have any better ability to correct this problem than other test formats.

6. Different methods may show different student achievement results.

Answers to important questions about student achievement may vary depending on the data analyzed or how it is analyzed and reported. For example, annual testing programs (i.e., fall-to-fall or spring-to-spring) tend to show much smaller achievement increases than testing programs that use a fall-to-spring testing cycle (Linn, Dunbar, Harnisch, & Hastings 1982). The differences may be caused by student selection, scale conversion errors, administration conditions, administration dates compared to test norming dates, practice effects, and teaching to the test.

Figure 2
Results of Changing to a New Test Form*



*Based on Linn, Graue, & Sanders, 1991.

Note that after a period of rising test scores, a new test form is introduced between years 8 and 9. Consequently, test scores drop dramatically in year 9, followed by another steady rise in years 10, 11, and 12. The increase is probably not a result of increased achievement. This is a very typical test score pattern.

This policy brief from the National Center for Research on Evaluation, Standards, and Student Testing was adapted from the CRESST Technical Report 490, *Standards-Led Assessment: Technical and Policy Issues in Measuring School and Student Progress*, 1998. The full report is available on the CRESST Web site, <http://www.cse.ucla.edu>, or by contacting Kim Hurst at CRESST, 310-206-1532 or writing to Kim at CRESST/UCLA, 301 GSE&IS, Mailbox 951522, Los Angeles, CA 90095-1522.

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U. S. Department of Education. The findings and opinions expressed in this publication do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement or the U.S. Department of Education.

TEN SUGGESTIONS FOR POLICYMAKERS

Despite these problems for standards-based assessment systems, and for most testing in general, there are a number of ways to improve the validity, credibility, and positive impact of assessment systems while minimizing their negative impact. It is recommended that policymakers:

1. Set standards that are high, but attainable. Unattainable standards lead the public to falsely believe that schools are beyond improvement. Similarly, standards that don't set a high mark will cause the public to lose faith in public schools.

2. Develop standards, then assessments. Studies on the NAEP achievement levels have clearly demonstrated the flaws in attempting to impose achievement levels or performance standards on existing assessments. Revision of existing tests, or creation of new ones, must closely measure the standards and accurately report student achievement.

3. Include all students in testing programs except those with the most severe disabilities. Use accommodated assessments for students who have not yet transitioned into English language programs or whose disabilities require it. This would help to assure accountability for all students and increase the comparability of results for different schools and districts. Report combined scores and separate subgroup scores to provide more accurate and useful information on student and school progress.

4. Useful high-stakes accountability requires new high-quality assessments each year that are comparable to those of previous years. Getting by on the cheap will likely lead to both distorted results, such as inflated scores, and distortions in education, for example, the narrow teaching to the test.

5. Don't put all of the weight on a single test when making important decisions about students and schools (i.e., retention, promotion, probation, rewards). Instead, seek multiple indicators of performance. Include performance assessments and other indicators of success such as attendance, students taking Advanced Placement courses, etc.

6. Place more emphasis on comparisons of performance from year to year than from school to school. This allows for differences in starting points while maintaining an expectation of improvement for all.

7. Set both long- and short-term school goals for all schools to reach. Short-term goals allow for differences in starting positions of different schools. Long-term goals permit expectations of the same high standards for all by including an expectation that lower achieving schools should have greater annual or biennial growth rates than current higher achieving schools. This combination will give schools a reasonable chance to show improvement, yet help guard against low expectations for schools and students.

8. Like an opinion poll, there is uncertainty in any educational testing system. That uncertainty should be reported in all test results.

9. Evaluate not only the hoped-for positive effects of standards-based assessments, but the unintended negative effects of the testing system.

10. Narrowing the achievement gap means that we must provide all children with the teachers and resources they need in order to reach our high expectations. This means improving the educational system as a whole, not just more testing or new testing systems.

References

- Education Week. (1997). Quality counts: A report card on the condition of public education in the 50 states. *A Supplement to Education Week*, Vol. 16, January, 22.
- Lerner, L. S. (1998). *State science standards: An appraisal of science standards in 36 states*. Washington, DC: Thomas B. Fordham Foundation.
- Linn, R. L., Dunbar, S. B., Harnisch, D. L., & Hastings, C. N. (1982). The validity of the Title I evaluation and reporting system. In E. R. House, S. Mathison, J. Pearsol, & H. Preskill (Eds.), *Evaluation Studies Review Annual* (Vol. 7, pp. 427-442). Beverly Hills, CA: Sage Publications.
- Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district results to national norms: The validity of the claims that "everyone is above average." *Educational Measurement: Issues and Practice*, 9(3), 5-14.
- Olson, L. (1998, April 15). An "A" or a "D": State rankings differ widely. *Education Week*, 17, 1, 18.
- Raimi, R. A., & Braden, L. S. (1998). *State mathematics standards: An appraisal of science standards in 46 states, the District of Columbia, and Japan*. Washington, DC: Thomas B. Fordham Foundation.
- Reese, C. M., Miller, K. E., Mazzeo, J., & Dossey, J. A. (1997). *NAEP 1996 mathematics report card for the nation and the states*. Washington, DC: National Center for Education Statistics.