Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 7, Number 11, January, 2001 ISSN=1531-7714

Assessments and Accountability (Condensed version)

Robert L. Linn
Center for Research on Evaluation, Standards, and Student Testing

Adapted, with permission of Robert L. Linn and the American Educational Research Association, from Linn, R. L. (2000). Assessments and accountability. Educational Researcher, 29 (2), 4-16.

Assessment and accountability have played prominent roles in many of the education reform efforts during the past 50 years. In the 1950s, under the influence of James B. Conant's work on comprehensive high schools, testing was used to select students for higher education and to identify students for gifted programs. By the mid-1960s test results were used as one measure to evaluate the effectiveness of Title I and other federal programs. In the 1970s and early 1980s, the minimum competency testing movement spread rapidly; 34 states instituted some sort of testing of basic skills as a graduation requirement. Overlapping the minimum competency testing movement and continuing into the late 1980s and early 1990s was the expansion of the use of standardized test results for accountability purposes.

Assessment is appealing to policymakers for several reasons: it is relatively inexpensive compared to making program changes, it can be externally mandated, it can be implemented rapidly, and it offers visible results. This Digest discusses significant features of present-day assessment programs and offers recommendations to increase positive effects and minimize negative ones.

What Are the Characteristics of Current Reform Efforts?

Although a number of other important features might be considered in any discussion of assessment and education reform (e.g., the emphasis on performance-based approaches to assessment, the concept of tests worth teaching to, and the politically controversial and technically challenging issue of opportunity to learn), I focus on the following three:

- An emphasis on the development and use of ambitious content standards as the basis of assessment and accountability.
- The dual emphasis on setting demanding performance standards and on the inclusion of all students.
- The attachment of high-stakes accountability mechanisms for schools, teachers, and sometimes, students.

Content standards. The federal government has encouraged states to develop content and performance standards that are demanding. Standards-based reform is also a central part of many of the state reform efforts, including ones such as Kentucky and Maryland that have been using standards-based assessments for several years and ones such as Colorado and Missouri that have more recently introduced standards-based assessment systems. A great deal has been written about the strengths and weaknesses of content standards (e.g., *Education Week,* 1997; Lerner, 1998; Olson, 1998; Raimi & Braden, 1998).

It is worth acknowledging that content standards vary a good deal in specificity and in emphasis. Content standards can, and should, if they are to be more than window dressing, influence both the choice of constructs to be measured and the ways in which they are eventually measured.

Performance standards. Performance standards are supposed to specify how good is good enough. There are at least four critical characteristics of performance standards. First, they are intended to be absolute rather than normative. Second, they are expected to be set at high, world-class levels. Third, a relatively small number of levels (e.g., advanced, proficient) are typically identified. Finally, they are expected to apply to all, or essentially all, students, rather than a selected subset such as college-bound students seeking advanced placement.

Should the intent be to aspire not just to high standards for all students, but to the same high standards for all students and on the same time schedule for all students (e.g., meet reading standards in English at the end of Grade 4)? Coffman (1993) sums up the problems of holding common high standards for all students as follows: "Holding common standards for all pupils can only encourage a narrowing of educational experiences for most pupils, doom many to failure, and limit the development of many worthy talents" (p. 8). Although this statement runs counter to the current zeitgeist and may not even be considered politically correct, it seems to me a sensible conclusion that is consistent with both evidence and common sense. Having high standards is not the same as having common standards for all, especially when they are tied to a lock step of age or grade level.

High-stakes accountability. The use of student performance on tests in accountability systems is not new. Examples of payment for results such as the flurry of performance contracting in the 1960s can be found cropping up and fading away over many decades. What is somewhat different about the current emphasis on performance-based accountability is its pervasiveness. As Elmore, Abelmann, and Fuhrman note, "What is new is an increasing emphasis on student performance as the touchstone for state governance" (1996, p. 65).

Student achievement is being used not only to single out schools that require special assistance, but also to provide cash incentives for improvements in performance. Yet several fundamental questions remain about the student assessments, the accountability model, and the validity, impact, and credibility of the system.

As noted earlier, for example, the choice of constructs matters. Content areas (and subareas within those content areas) that are assessed for a high-stakes accountability receive emphasis while those that are left out languish. Meyer (1996) has argued that "in a high-stakes accountability system, teachers and administrators are likely to exploit all avenues to improve measured performance. For example, teachers may 'teach narrowly to the test.' For tests that are relatively immune to this type of corruption, teaching to the test could induce teachers and administrators to adopt new curriculums and teaching techniques much more rapidly than they otherwise would" (p. 140).

It is unclear, however, that there is either the know-how or the will to develop assessments that are sufficiently "immune to this type of corruption." It is expensive to introduce a new, albeit well-equated, form of a test on each new administration. And if ambitious performance-based tasks are added to the mix, still greater increases in costs will result.

A second area of concern regarding high-stakes assessments relates to what data the basic model should employ. Some possibilities include current status, comparisons of cross-sectional cohorts of students at different grades in the same year, comparisons of cross-sectional cohorts in a fixed grade from one year to the next, longitudinal comparisons of school aggregate scores without requiring matched individual data, and longitudinal comparisons based only on matched student records. Should simple change scores be used or some form of regression-based adjustment? And, if regression-based adjustments are used, what variables should be included as predictors? In particular, should measures of socioeconomic status be used in the adjustments?

Elmore, Abelmann, and Furhman (1996) present both sides of this issue, noting that on the one hand, schools can fairly be held accountable only for those factors they can control, but on the other, controlling for student background or prior achievement institutionalizes low expectations for poor, minority, low-achieving students (pp. 93-94). Kentucky's interesting approach to this dilemma has been to set a common goal for all schools by the end of 20 years, thus establishing faster biennial growth targets for initially low-scoring schools than initially high-scoring schools (Guskey, 1994).

The biggest question of all is whether the assessment-based accountability models that are now being used or being considered by states and districts have been shown to improve education. Unfortunately, it is difficult to get a clear-cut answer to this simple question. Certainly, there is evidence that performance on the measures used in accountability systems increases over time, but that can also be linked to the use of old norms, the repeated use of test forms year after year, the exclusion of students from participating in accountability testing programs, and the narrow focusing of instruction on the skills and question types used on the test (see Koretz, 1988; Linn et al., 1990; Shepard, 1990). Comparative data are needed to evaluate the apparent gains. The National Assessment of Educational Progress provides one source of such data. Comparisons of state NAEP and state assessment results sometimes suggest similar trends; for example, increases in numbers of students scoring at or above basic or proficient levels on NAEP may track with improved state test scores over time. In other cases, the trends for a state's own assessment and NAEP will suggest contradictory conclusions about the changes in student achievement. Divergence of trends does not prove that NAEP is right and the state assessment is misleading, but it does raise important questions about the generalizability of gains reported on a state's own assessment, and hence, about the validity of claims regarding student achievement.

How Can Assessments Be Used More Wisely?

Assessment systems that are useful monitors lose much of their dependability and credibility for that purpose when high stakes are attached to them. The unintended negative effects of the high-stakes accountability uses often outweigh the intended positive effects. It is worth arguing for more modest claims about uses that can validly be made of our best assessments and warning against the over-reliance on them that is so prevalent and popular. To enhance the validity, credibility, and positive impact of assessment and accountability systems while minimizing their negative effects, policymakers should:

1. Provide safeguards against selective exclusion of students from assessments.
2. Make the case that high-stakes accountability requires new high-quality assessments each year that are equated to those of previous years.
3. Don't put all of the weight on a single test. Instead, seek multiple indicators. The choice of construct matters and the use of multiple indicators increases the validity of inferences based upon observed gains in achievement.
4. 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.
5. Consider both value added and status in the system. Value added provides schools that start out far from the mark a reasonable chance to show improvement while status guards against institutionalizing low expectations for those same students and schools.
6. Recognize, evaluate, and report the degree of uncertainty in the reported results.
7. Put in place a system for evaluating both the intended positive effects and the more likely unintended negative effects of the system.

References

Coffman, W. E. (1993). A king over Egypt, which knew not Joseph. *Educational Measurement: Issues and Practice*, 12(2), 5-8.

Education Week (1997, January 22). Quality counts: A report card on the condition of public education in the 50 states. *A Supplement to Education Week*, vol. 16.

Elmore, R. F., Abelmann, C. H., & Fuhrman, S. H. (1996). The new accountability in state education reform: From process to performance. In H. F. Ladd (Ed..), *Holding schools accountable: Performance-based reform in education* (pp. 65-98). Washington, DC: The Brookings Institution.

Guskey, T. R. (Ed.) (1994). *High stakes performance assessment: Perspectives on Kentucky's reform*. Thousand Oaks, CA: Corwin Press.

Koretz, D. (1988). Arriving at Lake Wobegon: Are standardized tests exaggerating achievement and distorting instruction? *American Educator 12*(2), 8-15, 46-52.

Lerner, L. S. (1998). *State science standards: An appraisal of science standards in 36 states*. Washington, D.C. Thomas B. Fordham Foundation.

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.

Meyer, R. H. (1996). Comments on chapters two, three, and four. In H. F. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education* (pp. 137-145). Washington, DC: The Brookings Institution.

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.: The Thomas B. Fordham Foundation.

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test? *Educational Measurement: Issues and Practice*, *9*(3), 15-22.

Descriptors: Academic Achievement; Academic Standards; * Accountability; Educational Change; Educational History; Elementary Secondary Education; * Minimum Competency Testing; Standardized Tests; * Student Evaluation; Reform Efforts

Citation: Linn, Robert L. (2001). Assessments and accountability (condensed version). *Practical Assessment, Research & Evaluation*, 7(11). Available online: http://PAREonline.net/getvn.asp?v=7&n=11.