DOCUMENT RESUME

ED 393 875 TM 024 749

AUTHOR	Linn, Robert L.
TITLE	Assessment-Based Reform: Challenges to Educational Measurement.
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	7 Nov 94
NOTE	22p.; The Annual William H. Angoff Memorial Lecture (1st, Princeton, NJ, November 7, 1994).
PUB TYPE	Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Accountability; Educational Assessment; *Educational Change; Equal Education; *Measurement Techniques; *Minimum Competency Testing; Multiple Choice Tests; *Performance Based Assessment; Standards; Test Results; *Test Use
IDENTIFIERS	Goals 2000; High Stakes Tests; *Opportunity to Learn; Reform Efforts

ABSTRACT

Although the use of test results to demonstrate educational shortcomings is important, tests and assessments are also expected to be an instrument of reform. In fact, they are often expected to provide the primary means of creating educational reform. The minimum competency testing movement and the expansion of the use of test results for accountability purposes have been recent waves of test-based reform. The most recent wave of reform continues to emphasize accountability, but adds emphasis on using forms of assessment that require students to perform more substantial tasks than merely selecting multiple-choice items (performance assessment). Standards that shape assessment and define acceptable levels of performance and that insist on the inclusion of all students are also being advocated. Such standards are supported by the Goals 2000 legislation, as are the most controversial standards in the Goals, those of opportunity-to-learn (OTL). OTL standards are being demanded by those who feel that it is not fair to hold students accountable for meeting performance standards unless they have been given adequate opportunity to meet them. OTL standards pose daunting measurement challenges that will be increased by high-stakes use. (Contains 1 table, 2 figures, and 37 references.) (SLD)

MEMORIAL LECTURE Series

ASSESSMENT-BASED REFORM: CHALLENGES TO EDUCATIONAL MEASUREMENT

William H. Angoff
1919 - 1993

William H. Angoff was a distinguished research scientist at ETS for more than forty years. Dr. Angoff's contributions to educational measurement include the development of a score equating system, studies of item bias and group differences, as well as work on the effects of guessing on test scores and the differential impact of curriculum on aptitude test scores. He authored some of the classic publications on psychometrics, including "Scales, Norms, and Equivalent Scores," which appeared in Robert L. Thorndike's Educational Measurement and is considered the definitive treatment of the subject. Dr. Angoff was also noted for the rare ability to make technically complex issues accessible to a broad audience.

Assessment-Based Reform:

Challenges to Educational Measurement

The first annual William H. Angoff Memorial Lecture was presented at Educational Testing Service, Princeton, New Jersey, on November 7, 1994.

Robert L. Linn
Center for Research on Evaluation,
Standards, and Student Testing
University of Colorado at Boulder

Preface

The William H. Angoff Memorial Lecture Series was established in 1994 to honor the life and work of Bill Angoff, who died in January 1993. For more than 50 years, 43 of them at ETS, Bill made major contributions to psychological and educational measurement and was deservedly recognized by the major societies in the field. As the notion of an annual lecture series took shape, the idea that these lectures should be devoted to relatively non-technical discussion of public interest issues related to educational measurement struck us all as eminently suitable. This was an aspect of our field in which Bill was keenly interested and into which he made several successful forays. I know he thought it part of our professional obligation to encourage and support reasoned public debate on important topics.

We were all very pleased indeed when Professor Robert Linn of the University of Colorado agreed to be the first speaker in the series. For a time, Bob was a colleague of Bill's at ETS, and Bill very much liked, admired, and respected him. Bob is the recipient of many awards including APA's E. L. Thorndike Award, AERA's E. F. Lindquist Award, and ETS's Distinguished Service to Measurement Award. In addition to his many technical contributions to the measurement literature, Bob has been indefatigable in serving the education community through his active participation in various committees and taskforces. Among the more noteworthy are his efforts in developing the *Standards for Education and Psychological Testing* and editing the most recent edition of the *Handbook of Educational Measurement.*

Bob is well known for his ability to balance technical, scientific, and political considerations in the analysis of complex issues—and then to communicate clearly to different audiences. Nowhere has this talent been more evident than in the current debate about performance assessments and their role in educational reform. Bob and his colleagues have played a leading and constructive role in that debate. The present paper, a slightly revised version of his lecture delivered on November 7, 1994, provides a reasoned review of the promise of well-designed performance assessments as well as the technical challenges that must be overcome if they are to have a significant, positive impact on reform. It is particularly instructive on the role of different kinds of standards in reform and the validity of issues that arise as we move beyond multiple-choice testing.

No effort such as this takes place without a great deal of work by many people. Let me first note that the Lecture Series and this publication are jointly supported by ETS and an endowment fund that was established in Bill Angoff's memory. I want to thank Bill's many friends and colleagues who contributed to the fund, and I want to particularly acknowledge a generous contribution by Eleanor Angoff on behalf of the family. Nancy Cole, ETS President, and Eleanor Horne, ETS Corporate Secretary, offered advice and encouragement. Madeline Moritz and Helen Tarr in my office provided administrative support for the lecture, and Shilpi Niyogi provided editorial support for this publication.

Thanks to all of you.

Henry Braun
Vice president for Research Management, ETS

July 1995

Preamble

It is a great privilege and honor to be here to give the first William H. Angoff Memorial Lecture. Reflection on the consistently high standards that Bill maintained for his work throughout his career makes this also a very humbling experience because it clearly is not easy to live up to those standards.

Bill was pointed out to me shortly after I arrived at ETS as a new Ph.D. in 1965 as someone to watch as a role model if I wanted to achieve success in the field of educational measurement. That observation was clearly sound, albeit not so easy to follow, advice. Bill was an exceptional scholar and a supportive colleague. One cannot think about issues of scaling or equating without thinking of Bill. In these and other areas of measurement where Bill made his major contributions, his work was always meticulous and of the highest technical quality. Equally important, it was written in an articulate and coherent fashion that made even the most complex concepts widely accessible.

The William H. Angoff Memorial Lecture Series is a fitting tribute to Bill for his long and distinguished career.

Assessment-Based Reform: Challenges to Educational Measurement

Educational tests are often rather naively expected to serve as an impartial barometer of educational quality. This expectation makes test results of particular interest and value to policymakers and politicians.

A large part of the appeal of tests to policymakers comes from their use to demonstrate shortcomings of education. The Office of Technology Assessment (OTA) report, *Testing in American Schools: Asking the Right Questions* (1992), provides a brief recounting of this history of testing in American Schools from the time that Horace Mann introduced written examinations in the mid-19th century. The OTA report summarized the view that tests could support reform by documenting the need for change as follows:

> "The idea underlying the implementation of written examinations ... was born in the minds of individuals already convinced that education was substandard in quality. This sequence—perception of failure followed by the collection of data designed to document failure (or success)—offers early evidence of what has become a tradition of school reform and a truism of student testing: tests are often administered not just to discover how well schools or kids are doing, but to obtain external confirmation—validation—of the hypothesis that they are not doing well at all" (U.S. Congress, Office of Technology Assessment, 1992, p. 106, emphasis in original).

Although the use of results to demonstrate shortcomings is important, test and assessments are expected to serve another more demanding role in reform. They are expected to be an instrument of reform. Indeed, tests are often expected to provide the primary means of creating educational reform.

Assessment has great appeal to policymakers as an agent of reform for a number of reasons.

(1) Tests and assessments are relatively inexpensive. Compared to changes that involve increases in instructional time, reduced class size, attracting more able people to teaching, hiring teacher aides, or programmatic changes involving substantial professional development for teachers, assessment is cheap.

(2) Testing and assessment can be externally mandated. It is far easier to mandate testing and assessment requirements at the state or district level than anything that involves actual change in what happens inside the classroom.

(3) Testing and assessment changes can be rapidly implemented. Importantly, new test or assessment requirements can be implemented within the term of office of elected officials.

(4) Results are visible. Test results can be reported to the press. Poor results in the beginning are desirable for policymakers who want to show they have had an effect. Based on past experience, policymakers can reasonably expect increases in scores in the first few years of a program (see, for example, Linn, Graue, & Sanders, 1990) with or without real improvement in the broader achievement constructs that tests and assessments are intended to measure. The resulting

overly rosy picture that is painted by short-term gains observed in most new testing programs gives the impression of improvement right on schedule for the next election.

A Nation at Risk (National Commission on Excellence in Education, 1983) and the reforms that were introduced in its wake during the past decade illustrate both the use of test results to demonstrate the need for reform and as an instrument of reform. Trends in scores on the Scholastic Aptitude Test (SAT), for example, played an important role in the argument in *A Nation at Risk* that education in the United States was in a state of decline. Much could be said about the inappropriateness of this use of SAT scores (see, for example, College Board, 1977), but that is another story.

Testing also played a prominent role in the reforms advocated in *A Nation at Risk*. Certification and identification of students at both extremes (those in need of remediation and those ready for advanced or accelerated work) were highlighted in the report. Nearly every educational reform that was introduced by states throughout the country following *A Nation at Risk* either mandated new testing requirements or expanded use of existing testing programs. Petrie (1987) concisely characterized this rush to testing as the primary mechanism of reform stating that: "It would not be too much of an exaggeration to say that evaluation and testing have become the engine for implementing educational policy" (p. 177).

Recent Assessment-Based Reform Experiences

Before considering the current round of assessment-based reforms, it may be useful to recall briefly some aspects of the two most recent waves of test-based reforms.

Minimum-Competency Testing

In the 1970s and early 1980s, minimum-competency testing (MCT) reforms swiftly spread from state to state. In a single decade (1973-1983), the number of states with some form of minimum-competency testing requirement went from two to 34. As the name implies, the focus was on the lower end of the achievement distribution. Minimal basic skills, while not easy to define or defend, were widely accepted as a reasonable requirement for high school graduation. The new requirements were of great importance for some students but had little relevance for most students. Gains in student achievement were observed, but they occurred mostly at the low end of the distribution. Moreover, questions were raised by some about generalizability of the observed gains.

An important concept that emerged from the MCT movement that has great relevance for the current standards and assessment-based reform efforts is that of opportunity to learn (OTL). The focus on OTL was sometimes discussed in terms of the need for minimum-competency tests to have curriculum or instructional validity (see, for example, Madaus, 1983). The match between what was tested and what students were taught was one of the key issues in determining whether students had been provided with a fair opportunity to learn the knowledge and skills required by the minimum-competency test in the *Debra P. vs. Turlington* case (474, F Supp. 244, M. D. Fla., 1979).

Pullin (1983) summarized the *Debra P* trial court's requirements for demonstrating a fair opportunity to learn the required knowledge and skills. She listed seven requirements:

(1) students must be informed of the objectives to be tested at the time of instruction, (2) instruction must be offered on the tested objectives, (3) an "orderly sequence of instruction that affords [students] an opportunity to acquire proficiency through an appropriate developmental process" (p. 17) must be provided, (4) an adequate amount of instructional time must be spent on tested skills, (5) instruction or a review must be provided just prior to test administration, (6) teaching must include a means of determining whether "objectives are being learned by individual students" (p. 17), and (7) remedial instruction opportunities must be offered. This is a formidable list of requirements when the focus of testing is limited to the relatively low-level, basic skills required to pass a minimum-competency test. If adhered to, it would appear even more daunting in the context of the more demanding reasoning and problem-solving skills with high expected standards of achievement that are part of the current reform effort.

Accountability

Overlapping with the minimum-competency testing movement and continuing past the height of that movement into the late 1980s and early 1990s was an expansion of the use of test results for accountability purposes. Accountability programs took a variety of forms, but shared the common characteristic that they increased real or perceived stakes in results for teachers and educational administrators.

Although some states and districts contracted for or developed their own tests, the accountability systems of the 1980s relied heavily on published standardized tests. Upward trends in student achievement were reported by an overwhelming majority of states and districts during the first few years of accountability testing programs. A physician, John Cannell (1987) forcefully brought to public attention what came to be known as the Lake Wobegon effect (Koretz, 1988), i.e., the incredible finding that essentially all states and most districts were reporting that their students were scoring above the national norm.

The quotes in Table 1, are a sampling of dozens that were collected as part of a follow-up study of the widespread reporting of the Lake Wobegon effect (Linn, Graue, & Sanders, 1990). The persuasiveness of such reports contrasted sharply with Cannell's

Table 1

Illustrative Quotes Regarding Reports of Standardized Achievement Test

Results by States and Districts

“.. sixth grade students statewide performed as well or better than 60 percent of their grade level peers across the nation.”

“The average achievement for x number of students ... are above the national average for the second year in a row.”

“x percent of students at all three grade levels are scoring above the national norm sample in all three skill areas.”

“The average student in district x scored equal to or higher than the average student in the national norm group.”

“... the performance of state x students on average has been consistently higher than the national average.”

“... students showed greater reading proficiency than did their peers in the national sample.”

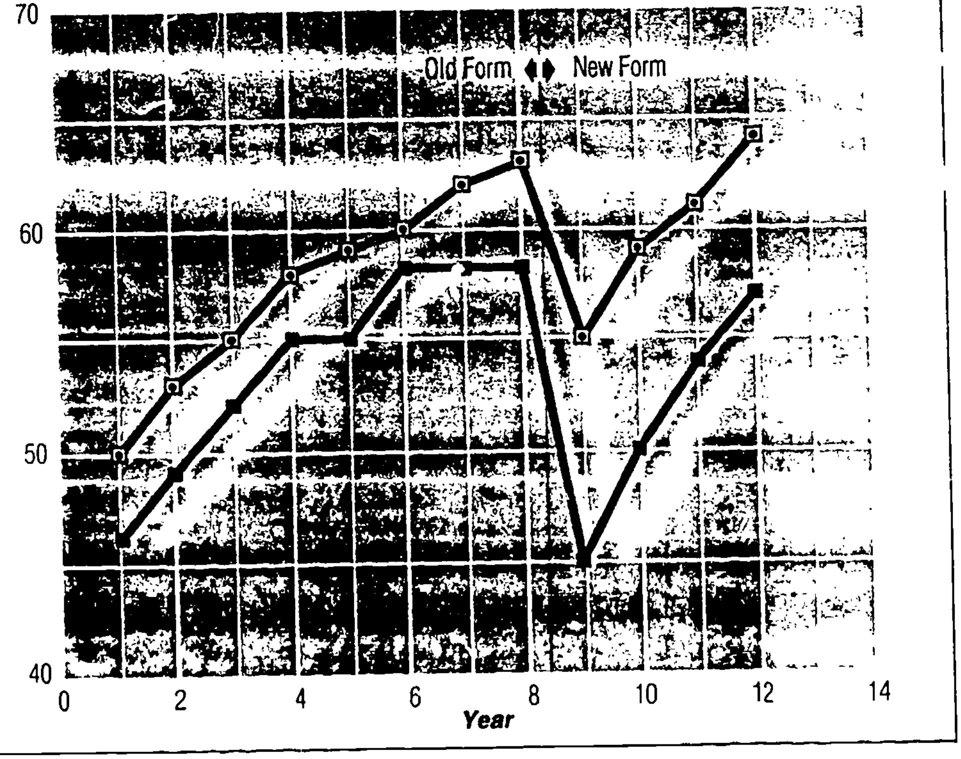
personal impressions of student performance and led him to collect information from all states and a number of districts. Based on his review of the data, Cannell (1987) concluded that "...standardized, nationally normed achievement tests give children, parents, school systems, legislatures, and the press inflated and misleading reports on achievement levels" (p. 3).

There are many reasons for the Lake Wobegon effect, most of which are less sinister than those emphasized by Cannell. Among the many reasons are the use of old norms, the repeated use of the same test form year after year, the exclusion of students from participation in accountability testing programs at a higher rate than they are excluded from norming studies, and the narrow focusing of instruction on the skills and question types used on the test (see, for example, Koretz, 1988; Linn, Graue, & Sanders, 1990; Shepard, 1990). In each of the categories, practices range from quite acceptable to quite unacceptable. For example, the focusing of instruction on the general concepts and skills included in the test may be in keeping with the belief that the test corresponds to instructionally important objectives and considered acceptable, even desirable, practice. On the other hand, the narrow teaching of the specific content sampled by the test, or coaching in specific responses to test items would be widely condemned as unacceptable practice.

Whatever the reason for the Lake Wobegon effect, it is clear that the standardized test results that were widely reported as part of accountability systems in the 1980 were giving an inflated impression of student achievement. Striking evidence of this comes from trend results for states and districts that include a shift from an old to a new test. The pattern shown in Figure 1 is similar to ones observed repeatedly where a new test replaced one that had been in use by a state or district for several years. The sawtooth appearance in Figure 1 demonstrates the lack of generalizability of the apparent gains on a test that is reused for several years.

Figure 1

Trend in Precentile Rank of State Mean



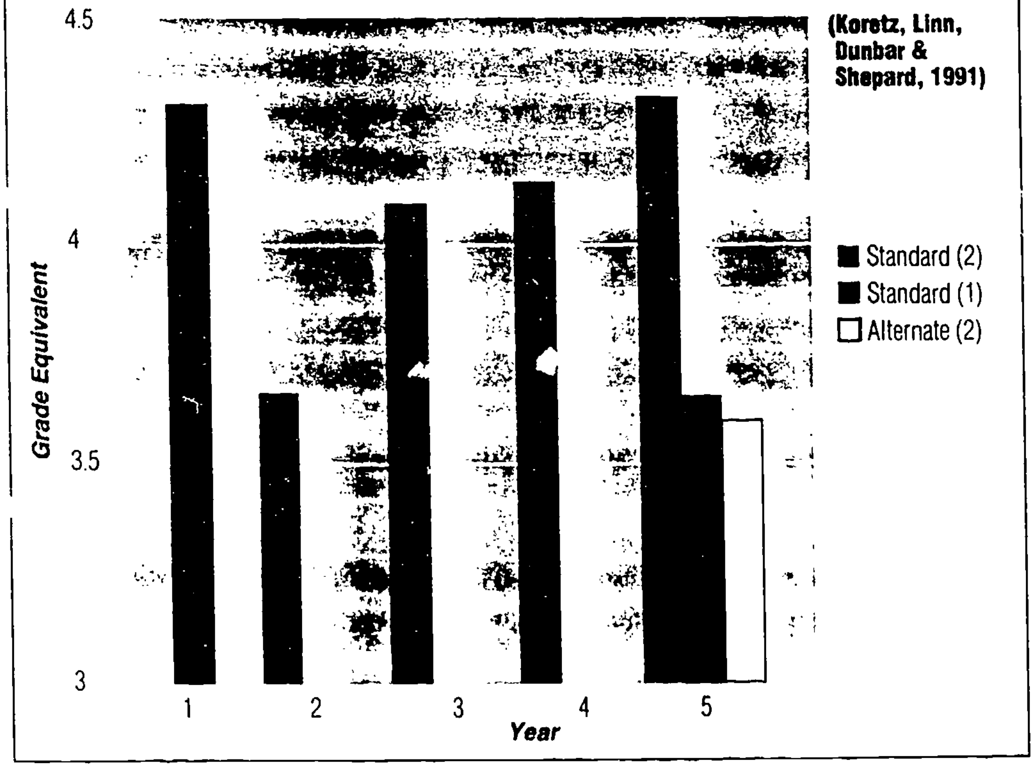
Koretz, Linn, Dunbar, and Shepard (1991) provide further evidence of the lack of generalizability of accountability test results. Figure 2 displays median third grade mathematics test results for a school district participating in that study. In the first year, when standardized test 1 was used by the district, the median grade equivalent score for the district was 4.3. A new test, standardized test 2, was first administered for accountability purposes in the second year and used in each of the following years (3, 4, and 5). The sawtooth pattern similar to that in Figure 1 is clearly evident. That is, there is a sharp drop in scores the first year a new test is administered followed by gains on administrations in subsequent years.

Standardized test 1 was administered to a sample of students in the district in the fifth year by Koretz, et al. They also administered an alternative test that was constructed for the study to cover content defined by the district curriculum and the content of standardized test 2. Data collected in other districts were used to equate the alternate test to standardized test 2. As can be seen in Figure 2, the results for both standardized test 1 (formerly the districts operational test) and the alternate test are more in line with those for the first year's administration of standardized test 2 than with the concurrent administration of that test in the fifth year.

Results such as those shown in Figures 1 and 2 were used to make the case that standardized test results in high-stakes accountability systems were yielding inflated impressions of student achievement. Strong arguments were also advanced that high-stakes accountability uses of standardized tests had undesirable effects on teaching and learning because they led to a narrowing of the curriculum and an over-emphasis on basic skills (e.g., Resnick & Resnick, 1992). One response has been to call for a change in the nature of assessments and the degree to which they are aligned with the types of learning envisioned in emerging content standards.

Figure 2

Median Grade 3 Mathematics Grade Equivalent Scores



(Koretz, Linn, Dunbar & Shepard, 1991)

Salient Characteristics of Current Reform Effort

The most recent wave of reform continues to emphasize accountability, but adds some significant new features. Perhaps the three most notable of the new features are the emphasis on (1) using forms of assessment that require students to perform more substantial tasks (e.g., construct extended essay responses, conduct experiments) rather than only select answers on multiple-choice items, (2) adopting standards that both shape the assessments and define levels of acceptable performance, and (3) the inclusion of all students.

Performance-Based Assessment

Demands for "new" approaches to assessment variously referred to as alternative assessment, authentic assessment, direct assessment, or performance-based assessment have been heard with increasing frequency in the past few years. Whatever the qualifier, "assessment" is intended to suggest a shift from fixed-response, machine-scorable tests to the use of tasks requiring students to construct responses that are scored by human judges. Each qualifier emphasizes a particular feature of the assessments. I prefer "performance-based" to the other three, because it is more descriptive of the key change and does not involve an implicit undocumented validity claim like "authentic" or "direct" (see, for example, Messick, 1994).

Although there are some signs of a slowing of the trend and even of a backlash that has pushed a few districts and states back toward more "objective" and economical machine-scorable tests (e.g. Littleton, Colorado; California, see, for example, Olson, 1995), the expansion of various kinds of performance-based assessments by districts and states has been remarkable. Writing assessments led the way. Constructed responses on mathematics assessments and performance-based science assessments followed.

Calls for the increased reliance on performance-based assessment generally rest on three premises that were articulated by Resnick and Resnick (1992). The first, is characterized by the acronym, WYTIWYG (What You Test Is What You Get). The second premise is the contrapositive of WYTIWYG, i.e., "you do not get what you do not assess." The third premise is a logical conclusion that follows from acceptance of the first two if one believes, as I do, that some form of testing or assessment will continue to be demanded for purposes of accountability. It is: "make assessments worth teaching to" (Resnick & Resnick, 1992, p. 59).

These premises are coupled with an acceptance of the argument that high-stakes testing and assessment shapes instruction and student learning. Rather than trying to change that connection, proponents of performance-based assessment argue that it is assessments that need to be modified, not only to eliminate the negative effects of teaching to the assessment but also to make that activity have the desired result of enhanced student learning.

Performance-based assessments are thought to be more compatible with modern conceptions of learning that view learners as active constructors of knowledge rather than passive receptacles of information. According to Resnick and Resnick (1992), for example, widely used machine-scorable tests reflect an outmoded model of learning that involves an accumulation of a "collection of independent pieces of knowledge" (p. 41) and skills that can be applied regardless of context. Performance-based assessments are intended to overcome these two perceived shortcomings of standardized tests that the Resnicks refer to as "decomposition" and "decontextualization".

The goal of creating assessments that are worth teaching to is appealing. There is little evidence, however, that the distortion created by previous test-based reforms can be avoided by a shift in the form of assessment. There clearly is a need to take seriously Messick's (1994) caution that, "it is not just that some aspects of multiple-choice testing may have adverse consequences for teaching and learning, but that some aspects of all testing, even performance testing, may have adverse as well as beneficial consequences" (p. 22).

Standards

The second key feature of current reform efforts is the creation of standards. Standards are central to the Clinton administration's education initiative explicated in the *Goals 2000: Educate America Act*. *Goals 2000* is reinforced by the requirements for Title I evaluation stipulated in the recently passed *Improving America's Schools Act of 1994*. Standards-based reporting is also a central part of many of the state reform efforts (e.g., Kentucky, Maryland, North Carolina, California). Indeed, states are likely to be the key actors in standards-based reforms, particularly as the result of the current Congressional plans to reduce the federal role and give more flexibility and responsibility to states.

Three types of standards were distinguished in Goals 2000: content standards, performance standards, and opportunity-to-learn standards. Distinctions among these three types of standards are apt to be critical regardless of the future role of Goals 2000 or the federal government in educational reform. Thus, each of the three types of standards deserves some elaboration.

Content Standards

Content standards are expected to specify what should be taught and students should learn. The best known model for content standards is the Curriculum and Evaluation Standards for School Mathematics developed by the National Council of Teachers of Mathematics (NCTM, 1988). Content standards in a variety of other subject areas are either under development or have appeared within the last couple of years. In addition to the content standards developed at the national level under the leadership of professional associations following the model established by NCTM, tailored versions of content standards have been or are being developed by a number of states.

Although content standards are sometimes confused with curriculum, the two are distinct. Content standards identify important concepts and skills that students are expected to learn, but they do not mandate a particular curriculum, textbook, instructional approach, or series of lessons. Content standards may serve as a guide for designing or evaluating curriculum, assessments, and instructional programs, but in each case the intent of the standards could be met in a variety of ways.

Using content standards as the guiding beacon of educational reform presupposes that a broad consensus can be achieved about what is most important for students to learn. It also assumes that the consensus can be maintained when standards are embodied in specific curriculum or assessment materials. Consensus clearly becomes more difficult to achieve as curriculum materials and assessments make standards more concrete and specific.

Recent controversy over the release of the national history standards (see, for example, Diegmueller, 1994b) demonstrates that the assumptions regarding the ability to reach a sufficiently broad-based consensus to support standards-based reform should not be taken lightly. As Cremin (1990) has noted, "... standards involve much more than determinations of what knowledge is of most worth; they also involve social and cultural differences, and they frequently serve as symbols and surrogates for those differences" (p. 9). In light of Cremin's observation, it would seem that some level of controversy over content standards is inevitable. The struggle over what gets emphasized, what gets included, and what gets excluded from the content standards, performance standards, and assessments is a struggle over educational values.

Competing values can take a variety of forms. Disagreements about what deserves emphasis within a content area such as those illustrated by the debate over the appropriate attention that should be given to traditional American heroes and successes in the history standards are to be expected. The proper role of the disciplines in defining standards can also be a source of controversy. The content that will produce a broad consensus for disciplinary specialists may not be so favorably viewed by the public. Recent public objections to mathematical assessment tasks that give more weight to effective communication than to getting the arithmetic right, for example, illustrate the potential conflict between disciplinary specialists and the public even in mathematics, a field that has been subjected to much less controversy than history (see, for example, Colvin, 1995).

In addition to concerns about public acceptance, disciplines face competition with each other. It is evident that no discipline can afford to be left out. Hence, there is a proliferation of content standards as each discipline develops and promotes its own content standards in order to compete for instructional time. It is no surprise that draft content standards for health and for physical education and coaching were recently released (Diegmueller, 1994a). The Health standards claim that "student acquisition of health knowledge and skills is as significant to economic competitiveness, quality of life and school reform as the knowledge and skills taught through any other subject" (p.2). The physical education standards promote the idea that physical education has academic standing.

Those interested in interdisciplinary work worry about disciplinary imperialism that is fostered by emphasis on discipline-based content standards. Teachers, particularly at the elementary level, where they are responsible for the full array of subjects, face a potentially overwhelming array of content standards.

Performance Standards

Performance standards, while dependent on content standards, are distinct. According to Goals 2000, "the term 'performance standard' means concrete examples and explicit definitions of what students have to know and be able to do to demonstrate that such students are proficient in the skills and knowledge framed by content standards" (Public Law 103-227, sec. 3, (a) (9)). An elaboration of this definition was provided by the Goals 3 and 4 Technical Planning Group for the National Education Goals Panel.

> Performance standards specify 'how good is good enough.' In shorthand, they indicate how adept or competent a student demonstration must be to indicate attainment of the content standards. They involve judgments of what distinguishes an adequate from an outstanding level of performance. ... Performance standards are not the skills and modes of reasoning referred to in the content standards. Rather, they indicate both the nature of the evidence (such as an essay, mathematical proof, scientific experiment, project, exam, or combination of these) required to demonstrate that content standards have been met and the quality of student performance that will be deemed acceptable (what merits a passing or an 'A' grade)" (NEGP, Goals 3 and 4 Technical Planning Group on the Review of Education Standards, 1993, p. 22).

There are at least four critical characteristics of performance standards. First, they are intended to be absolute rather 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.

The emphasis on absolute judgments in terms of fixed standards is a defining feature of performance standards. Desires for comparisons, however, continue to surface, whether implicitly in descriptions of the standards as "world-class" or explicitly in requirements that state performance be compared to national achievement or to achievement of other states through NAEP or some other means. The following description of one state's goal and the accomplishment of that goal illustrates the continued use of comparisons as a basis of interpretation even in an era of standards-based reporting. "[the state's] goal by the year 2000 is for students to be learning at least as much as the national average in every subject. ... students have met that goal in reading and language" (October, 1994). This quote also sounds a warning that although the Lake Wobegon effect may be largely forgotten, it is not necessarily gone.

The second, third, and fourth characteristics (high standards, few levels, and all students) of the present performance standards rhetoric interact. There is a huge gap between current perceptions of current student performance and what should qualify as a high standard of performance. Using the standards set by the National Assessment Governing Board for NAEP, for example, only 2% of the nation's 12th grade students achieved at the "advanced" level in mathematics in 1992. One twelfth grader in six (16%) achieved at the "proficient" level or higher while slightly more than one in three (36%) failed to reach even the "basic" level (Mullis, Dossey, Owen, & Phillips, 1993, p. 64)[1]. States such as California, Kentucky, and Maryland that have reported results in terms of a small number of performance standards have all reported relatively small percentages of students achieving either their highest standards or ones

[1] The NAEP achievement levels have been the subject of considerable controversy (see, for example, American College Testing Program, 1993; National Academy of Education, 1993). Much of the controversy surrounding the NAGB Achievement Levels stems from concerns about the clarity and accuracy of the communication about actual student achievement that is communicated by the achievement levels. The controversy does not alter the fact that there is a large gap between expected performance embodied in the achievement levels and the levels of performance achieved by the majority of students.

that might be established as expectations or requirements for all students.

The large gap between expectations and current performance of students is not particularly problematic when the stakes attached to results are low. If the stakes for individual students are increased, however, the gap will have serious implications for a substantial number of students. Furthermore, at least in the short run, failure rates for traditionally underserved minority groups would surely be substantially higher than those for their more privileged counterparts. Using the 1992 NAEP grade 12 mathematics again as an example, the percentages of students achieving at the proficient level or higher by race/ ethnicity were: White, 19%; Black, 3%; Hispanic, 6%; Asian/ Pacific Islander 31%, and American Indian 4%. The corresponding percentages achieving below the basic level were 28%, 66%, 55%, 19%, and 54%, respectively (Mullis, et al., 1993, p. 93).

The requirements for assessments and performance standards in the *Improving America's Schools Act of 1994* (IASA) for Title I programs have sometimes been referred to as the 800-pound gorilla behind *Goals 2000*. IASA requires states to have a plan to develop or adopt challenging student performance standards that "describe two levels of high performance, proficient and advanced, that determine how well children are mastering the material in the State content standards, and describe a third level of performance, partially proficient, to provide complete information about the progress of lower performing children toward achieving the proficient and advanced levels of performance" (Public Law 102-382, sec. 101 (b) (1) (D) III and IV).

The number of performance standards set for a grade and content area varies from state to state, but is generally greater than the two points dividing performance into three levels required by IASA. Kentucky, for example, has set three performance standards in each content area resulting in four levels of achievement that are labeled Distinguished, Proficient, Apprentice, and Novice (Trimble, 1994). Maryland has set four standards that yield five levels of achievement for the Maryland School Performance Assessment Program (MSPAP). Results on the California Learning Assessment System (CLAS) were reported in terms of six performance levels, with 6 denoting the highest level of performance and 1 the lowest. Since the assessments and performance standards used for Title I students must be the same as those used statewide, it would appear that states will either have to revise the number of levels or provide a way of mapping the larger number of levels used for statewide assessments into the three levels of advanced, proficient, and partially proficient required by IASA.

The use of a small number of levels interacts with the desire for high standards and the expectation that all students will meet those standards. An assessment will provide little information for a school where the majority of students fail to meet the lowest standard. Using a small number of levels also makes classification errors more serious, which are bound to occur as the result of measurement errors. More will be said about this issue in discussing measurement challenges posed by standards-based reforms.

Opportunity-To-Learn Standards

The final category of standards defined in the *Goals 2000* legislation, is also by far the most controversial. Although referred to there as delivery standards, the concept of opportunity-to-learn standards was introduced in the report of the National Council on Education Standards and Testing (NCEST, 1992) in response to concerns that it would not be fair to hold students accountable for material they had not been taught. Opportunity-to-learn standards were eventually included in *Goals 2000*, but not without substantial debate and controversy.

OTL standards were demanded by those who were concerned that it is unfair to hold students accountable for meeting performance standards unless they have been given an adequate opportunity to meet those standards. They were resisted by those who were concerned that they would dictate local practice.

> "To proponents, OTL standards represent the age-old problems of equity in education. In particular, advocates of OTL standards see them as an appropriate antidote to the potentially negative effects of high stakes testing on students who, through no fault of their own, attend schools which provide an inferior education. To opponents, OTL standards evoke all their worst fears about federal intrusion into local control of the quality and nature of education" (Porter, 1994).

The compromise achieved in *Goals 2000* stressed the voluntary nature of OTL standards. Even the minimal acknowledgment of the need to attend to opportunity to learn before holding students accountable, however, has been weakened since the enactment of the law. Based on experience with Debra P., it seems likely that if there is to be any enforcement of OTL standards it will be through legal challenges to attaching awards and sanctions to the achievement of performance standards for individual students.

If OTL standards are required in the future either as the result of legislation or judicial decisions, they will pose daunting measurement challenges. Easy to measure characteristics, such as teaching experience, degrees, or the availability of materials, bear little relationship to student achievement. On the other hand, teacher reports of instructional time spent on particular content and activities, which are more strongly related to student achievement in research settings, are not likely to withstand pressures of high-stakes use (Porter, 1995).

Measurement Challenges

Although measurement problems presented by OTL standards may pose the most daunting challenges, those posed in the more familiar territory of student assessment are also substantial. The expectations for performance assessments are high. Assessments are expected to be cognitively demanding, engaging, authentic, and closely aligned with content standards. They are also expected to contribute to reforms in education that will result in improved student learning. Each of these expectations corresponds to a validity claim. A comprehensive program of validation research is needed to evaluate these and other expectations (e.g., the generalizability of student performance on the assessment to the broader domain of student learning defined by the content standards). A comprehensive validation also requires evidence regarding plausible unintended consequences of particular uses of assessment results (e.g., increased student dropout, the Lake Wobegon effect).

Goals 2000, the *Improving America's Schools Act*, and legislation in a number of states require that assessments be "valid, reliable, and consistent with professional technical standards." Although there is some disagreement about the precise meaning of these requirements, several performance-based assessments have been found wanting when subjected to close scrutiny on technical grounds (e.g., Cronbach, Bradburn, & Horvitz, 1994; Koretz, Stecher, Klein, & McCaffrey, 1994). Shortcomings in reliability, whether expressed in conventional terms or in terms of the likelihood that decisions would be reversed if the assessment was repeated, are particularly troublesome.

Some (e.g., Delandshere & Petrosky, 1994; Moss 1994) have suggested that traditional psychometric notions of reliability are not necessary and may not even be appropriate for performance-based assessments. I agree that traditional coefficients of reliability may be of little relevance to particular uses and interpretations of assessment results. Appropriate evaluations of measurement error, however, are just as relevant for performance assessments as they are for any other type of test. Reports appraising the dependability of scores by use of the standard error of measurement or estimated probabilities of misclassification or reversal of decisions that are derived from psychometric investigations of reliability or generalizability are critical components of an overall evaluation of the technical quality of an assessment system. The use of performance standards rather than normative comparisons as the basis for interpreting assessment results makes traditional reliability coefficients largely irrelevant. The standard error of measurement, however, is quite relevant for purposes of judging the dependability of a classification of a student with regard to a given performance standard. Although not a "reliability coefficient", an estimate of the probability that a "proficient" student will be misclassified as "advanced" or as "partially proficient," using performance standards required for Title I, is clearly consistent with spirit of the technical standards for reliability.The relevance of reports of standard errors of measurement and misclassification errors is evident in standards 2.10 and 2.12 of the Standards for Educational and Psychological Testing (AERA, APA, NCME, 1985).

> 2.10. "Standard errors of measurement should be reported at critical score levels. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported for score levels at or near the cut score" (p. 22).
>
> 2.12. "For dichotomous decisions, estimates should be provided of the percentage of test

> takers who are classified in the same way on two occasions or on alternate forms of the test" (p. 23).

Small changes in wording to expand from dichotomies to several levels defined by performance standards and to replace "test" with the more fashionable "assessment," would make these two standards, if anything, more relevant and important today than they were a decade ago.

The relatively large degree of measurement error, particularly when performance-based assessments are used to make decisions about individual students, was predictable (e.g., Linn, Baker, & Dunbar, 1991). The more important questions, however, concern an overall evaluation of the validity of the uses and interpretations of assessment results that not only take into account considerations of generalizability and measurement error, but also provide evidence regarding the construct validity of the assessments and their consequences.

The Vermont portfolio assessment program in mathematics illustrates the need to consider multiple types of evidence in reaching an integrated judgment regarding validity. The results of the Vermont assessment produced greater measurement error than traditional standardized tests (Koretz, et al. 1994). However, greater measurement error needs to be weighed against other factors, such as evidence that the "program has had substantial positive effects on fourth-grade teachers' perceptions and practices in mathematics" (Stecher, Mitchell, & Koretz, 1995).

Although leading theorists (e.g., Messick, 1989) have emphasized the need to include considerations of consequences of the uses and interpretations of assessment results in an evaluation of validity, practice has lagged behind. This gap is hardly surprising given the difficulty of evaluating consequences. The need to give greater attention to consequences, while not new, is exacerbated by the fact that a key part of the rationale for performance-based assessments and performance standards depends on their presumed impact on instruction and learning. Thus, as I have argued elsewhere (Linn, 1994; Linn, Baker, and Dunbar, 1991), there is a need to give greater priority to investigations of assessment consequences.

Performance standards and assessments demand changes in emphasis in evaluations of technical quality, but fundamental principles of validity and reliability still apply. As Messick (1994) has argued,

> "performance assessments must be evaluated by the same validity criteria, both evidential and consequential, as are other assessments. Indeed, such basic assessment issues as validity, reliability, comparability, and fairness need to be uniformly addressed for all assessments because they are not just measurement principles, they are social values that have meaning and force outside of measurement wherever evaluative judgment and decisions are made" (p. 13).

Doing a better job of translating the fundamental principles of validity, reliability, comparability and fairness into practice is one of the major challenges for the measurement profession. This will require a sharp focus on (1) the stated claims for an assessment (e.g., alignment with the content standards, measurement of conceptual understanding), (2) uses of results (e.g., student certification, school-level accountability), (3) interpretations (e.g., proficient students have "mathematical power"), (4) intended consequences (e.g., increased student learning), and (5) plausible unintended consequences (e.g., narrowing of curriculum coverage).

References

American College Testing Program. (1993). Setting achievement levels on the 1992 National Assessment of Educational Progress in Mathematics, Reading and Writing: A technical report on reliability and validity. Iowa City, IA: American College Testing Program.

American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education. (1985). Standards for Educational and Psychological Testing. Washington, DC: American Psychological Association.

Cannell, J. J. (1987). Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average (2nd edition). Daniels. West Virginia: Friends of Education.

College Board. (1977). On further examination: Report of the advisory panel on the Scholastic Aptitude Test score decline. Willard Wirtz, chairman. New York: College Entrance Examination Board.

Colvin, R. L. (1995). State's reading, math reforms under review as scores fall. Los Angeles Times, March 23, pp. A1, A21.

Cremin, L. A. (1989). Popular education and its discontents. New York: Harper & Row.

Cronbach, L. J., Bradburn, N. M., & Horvitz, D. G. (1994). Sampling and statistical procedures used in the California Learning Assessment System. Report of Select Committee to the Acting State Superintendent of Public Instruction, California State Department of Education, July 25.

Debra P. v. Turlington, 644 F.2d 397, 6775 (5th Cir. 1981).

Delandshere, G. & Petrosky, A. (1994). Capturing teachers' knowledge: Performance assessment. (a)-and post-structuralist epistemology, b)-from post-structuralist perspective, c)-and post-structuralism, d)-none of the above. Educational Researcher, 23 (5), 11-18.

Diegmueller, K. (1994a). Draft standards for health, P.E. are released. Education Week, 14, no. 7, Oct. 19, p. 8.

Diegmueller, K. (1994a). Panel unveils standards for history: Release comes amid outcries of imbalance. Education Week, 14, no. 9, Nov. 2, pp. 1, 10.

Goals 2000: Educate America Act of 1994, Public Law 103-227, Sec. 1 et seq. 108 Stat. 125 (1994).

Improving America's Schools Act of 1994, Public Law 103-382, Sec. 1 et seq. 108 Stat 3542 4 (1994).

Koretz, D. (1988). Arriving at Lake Wobegon. Are standardized tests exaggerating achievement and distorting instruction? American Educator, 12, 81-5, 16-18.

Koretz, D., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991). The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL, April.

Koretz, D. Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. Educational Measurement: Issues and Practice, 13, no. 3, 13-16.

Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. Educational Researcher, 23, no. 9, 4-14.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20(8), 15-21.

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, no. 3, 5-14.

MADAUS, G. F. (ED.). (1983). The courts, validity, and minimum competency testing. Boston, Kluwer-Nijhoff Publishing.

MESSICK, S. (1994). Validity. In R. L. Linn (Ed.), Educational Measurement, 3rd ed. (pp. 13-103). New York: Macmillan.

MESSICK, S. (1994). The interplay of evidence and consequences in the validation of performance assessments Educational Researcher, 23, 13-23.

MOSS, P. A. (1994). Can there be validity without reliability? Educational Researcher, 23(2), 5-12.

MULLIS, I. V. A., DOSSEY, J. A., OWEN, E. H., & PHILLIPS, G. W. (1993). NAEP 1992 mathematics report card for the nation and the states: Data from the national and trial state assessments. Report No. 23-ST02. Washington, DC: National Center for Education Statistics.

NATIONAL ACADEMY OF EDUCATION. (1993). A report of the National Academy of Education Panel on the evaluation of the NAEP trial state assessment: An evaluation of the 1992 achievement levels. Stanford, CA: National Academy of Education, Stanford University.

NATIONAL COMMISSION ON EXCELLENCE IN EDUCATION. (1983). A nation at risk: The imperative for educational reform. Washington, DC: U. S. Government Printing Office.

NATIONAL COUNCIL OF TEACHERS OF MATHEMATICS. (1988). Curriculum and evaluation standards for school mathematics. Reston, VA: Author.

OLSON, L. (1995). The new breed of assessments getting scrutiny. Education Week, 14, no 26, March 22, pp. 1, 10-11.

PETRIE, H. G. (1987). Introduction to "evaluation and testing". Educational Policy, 1, 175-180.

PORTER, A. (1994). The uses and misuses of opportunity to learn standards. Paper presented at a Brookings Institution conference, Beyond Goals 2000: The Future of National Standards in American Education. Washington, DC: May 18.

PORTER, A. (1995). The uses and misuses of opportunity-to-learn standards. Educational Researcher, 24, no. 1, 21-27.

PULLIN, D. (1983). Debra P. v. Turlington: Judicial standards for assessing the validity of minimum competency tests. The courts, validity, and minimum competency testing. (pp. 3-19). Boston, Kluwer-Nijhoff Publishing.

RESNICK, L. B. & RESNICK, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. G. Gophered & M. C. O'CONNOR (E's.), Changing assessments: Alternative views of aptitude, achievement and instruction (pp. 37-75). Boston: Kluwer Academic Publishers.

SHEPARD, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test?" Educational Measurement: Issues and Practice, 9, no. 3, 15-22.

STECHER, B. M., MITCHELL, K. J., & KORETZ, D. (1995). Portfolio-driven reform: Vermont teachers' understanding of mathematical problem solving and related changes in classroom practice. Technical Report Grant No. R117G10027. Los Angeles, UCLA, National Center for Research on Evaluation, Standards, and Student Testing.

TRIMBLE, C. S. (1994). Ensuring educational accountability. In T. Guskey (Ed.), High stakes performance assessment: Perspectives on Kentucky's education reform (pp. 37-54). Thousand Oaks, CA: Corwin Press, Inc.

U.S. CONGRESS, OFFICE OF TECHNOLOGY ASSESSMENT. (1992). Testing in American schools: Asking the right questions. OTA-SET-519. Washington, DC: U.S. Government Printing Office.

WILLIAM H. ANGOFF
MEMORIAL
LECTURE
Series

POLICY INFORMATION CENTER
Educational Testing Service
Princeton, New Jersey 08541-0001

04202-12764•Y85M3•204901 • Printed in U S A