American Educational Research Journal
Summer 1994, Vol. 31, No. 2, pp. 231–262

Assessment for Measurement or Standards: The Peril and Promise of Large-Scale Assessment Reform

Catherine Taylor
University of Washington

The current call for performance-based assessments is, in part, a consequence of inappropriate uses of norm-referenced achievement tests. Still, the use of performance-based assessment will not automatically eliminate the negative consequences of high-stakes tests, nor support hoped-for changes in schools. School reform will be supported only if new assessment systems are developed using a model that is in harmony with the goals of reform. This article reviews two models for assessment, the measurement model and the standards model, their underlying assumptions about learners, and the resulting implications for performance-based test development. It briefly reviews the current testing debate, defines terms such as authentic assessments *and* performance-based assessments, *and discusses the compromises that have led to the failed attempts to use testing to set standards for education. Finally, the article reflects on the power each assessment model can have on reform efforts.*

CATHERINE TAYLOR is an Assistant Professor in the College of Education at the University of Washington, 312 Miller Hall, DQ-12, Seattle, WA 98195. Her specializations are testing and measurement.

The call for new standards for America's schools (America 2000) has powered one of the most energetic curriculum reform movements since the mastery learning movement of the 1970s. Throughout the United States, educational agencies are engaged in the work of identifying essential exit outcomes and identifying grade-level indicators of progress toward those outcomes. At the same time, the demand for new assessment systems to support these reform efforts can be heard nationwide. Educators and policymakers are looking for assessment systems that will require students to engage in complex tasks using thinking and problem-solving skills rather than simply to demonstrate discrete knowledge and skill in applying that knowledge.

For the past 20 years, tests have been used extensively to provide school accountability information, to evaluate reform efforts, and to communicate important learning targets to schools (Jaeger, 1991). Although new thinking about the *purposes* of tests began during the 1970s, significant changes in the types of tasks on achievement tests were not made until very recently. This has been, in large part, because of underlying assumptions about what kinds of tests are needed to effectively and efficiently assess learners. Recently a new phase of exploration and experimentation with methods of large-scale testing has emerged. Many different assessment methods and formats are being implemented and the call for *authentic assessments, performance assessments*, and *portfolio assessments* is growing.

As educational agencies work toward reform and experiment with ways to determine the success of reforms, and do so under legislative and time pressures, an unarticulated tension has been growing. This tension comes from the purposes to be served by tests. Educators and legislators alike are demanding assessments that will serve two incompatible purposes: (a) determining whether students are achieving or striving toward desired standards of performance and (b) providing relative measurements of students, schools, districts, and states on scales of achievement.

We have faced this tension before and ultimately failed to resolve it. The existing assessments for these two purposes can be roughly categorized as *criterion-referenced* tests (assessment for standards) and *norm-referenced* tests (assessment for relative measurement), respectively. Over the past 20 years, the same assessment format and *often the same* tests have been used to serve both purposes. This has resulted in unfortunate consequences. The use of norm-referenced achievement tests for large-scale, criterion-referenced purposes (determining the success of programs, schools, and school districts in relation to the tested objectives) has led to norm-referenced score inflation (Koretz, 1991; Linn, Graue, & Sanders, 1990) and an erosion of standards (Darling-Hammond & Wise, 1985; Smith, 1991).

The current call for performance assessments is, in part, a consequence of inappropriate uses of norm-referenced achievement tests. Still, the use of performance assessments will not automatically eliminate the negative consequences of large-scale, high-stakes tests, nor will changes in response mode or testing format necessarily support hoped-for changes in the schools. School

reform efforts will be supported only if the new assessment systems are developed using an assessment model that is in harmony with the goals of reform.

For this article, I use the term *measurement model* to refer to the assessment model that has been the foundation of norm-referenced test development for the past 60 years. If we accept the assumptions of this model, the functions of tests are to assess general knowledge across some broadly defined area of achievement, to rank students based on their performance on the tests, and to compare students, schools, and districts on numeric scales of achievement. I use the term *standards model* to refer to what has been the conceptual foundation for criterion-referenced testing. If we accept the assumptions of this model, the function of these tests is to assess how students perform in relation to absolute standards. This model assumes that educators can define standards of performance and establish these standards as learning targets. I have chosen not to use the terms *norm-referenced* and *criterion-referenced* for two reasons:

1. Current applications and interpretations of each model are not necessarily the only ones available for the models. For example, in some circles, the term *norm* refers to a generally accepted *standard*. Therefore, it is important to use terms that are related to the deeper assumptions of each model.
2. Most educators have preconceived ideas about what *criterion-referenced tests* and *norm-referenced tests* are, based on experiences with objectives-based mastery tests and multiple-choice tests. These notions can blur the critical distinctions to be discussed here.

Performance assessments can be developed to serve the purposes of either model. In fact, elements of each model are being applied in the development of performance assessments already. There is an inherent danger in mixing these models and applying the assumptions and technology of the *measurement model* to the assessment of progress toward standards. Unless educators understand this danger, measurement specialists and policymakers will continue to demand performance assessments that adhere to the requirements needed for accurate, comparative *measurement*. This will lead to high-stakes performance tests that differ little from current standardized achievement tests.

We must use caution as we develop new ways to assess students' learning. We must choose the model that will fit the intended assessment purposes rather than hope that one assessment can actually serve incompatible purposes. Each model carries with it implications about the abilities of learners and the goals of instruction. Applying the measurement model to the development of large-scale performance assessments will ultimately undermine national efforts to improve the quality of education for all students. Before we spend even more time and money developing new assessments, we must carefully examine our own assumptions about learning and testing.

What follows is a brief review of the current testing debate, including definitions of terms such as *authentic* assessments and *performance assessments*, as well as some discussion of the compromises that have led to the failure of earlier

attempts to use testing to set standards for education. For the remainder of this article, I overview the measurement model and the standards model in their purest forms, including their underlying assumptions and the resulting implications for performance test development. Finally, I reflect on the power each model can exert on efforts to reform education in the United States.

The Current State of the Testing Debate

Much research has shown that inappropriate uses of traditional standardized achievement tests have had negative effects on schools and students (Darling-Hammond & Wise, 1985; Haladyna, Nolen, & Haas, 1991; Madaus, West, Harmon, Lomax, & Viator, 1992; Smith, 1991; Shepard, 1991a). Wiggins (1989) proposed the use of *performance assessments* as a more authentic and appropriate way to assess student learning. Wiggins (1989) defined performance assessments as evaluations of student works that are *authentic* to subject-area disciplines and that reflect the kinds of processes seen as central to each discipline (e.g., investigating a mathematical concept; writing an essay; conducting, evaluating, and generalizing from a science experiment; writing a position paper on an environmental issue). Wiggins's writings suggest that authentic performances can be identified for *all* subject areas, and these should form the foundation of new assessment programs.

Wiggins's writings sparked strong feelings among educators and policymakers. Although many were dissatisfied with current standardized achievement tests, educators at all levels had been taught to distrust other indicators of student learning (e.g., qualitative judgments of student work) and to assess learning using objectively scorable tests. Many saw standardized multiple-choice achievement tests as the only way to ensure fair and reliable large-scale testing.

Despite this tradition of distrust for qualitative judgment, performance assessments have gradually been accepted by many educators and policymakers as a promising new method for the assessment of important learner outcomes (Baron, 1991; Putka, 1989; Stiggins, 1991). Advocates for the use of performance assessments claim that when students are required to use their knowledge and skills to engage in complex performances, assessment of these performances will provide more valid information about student learning (Baron, 1990; Wiggins, 1989, 1990). Furthermore, advocates assume that large-scale performance assessment programs will influence classroom instruction in positive ways by encouraging teachers to broaden the focus of their teaching and include thinking and problem-solving processes in regular classroom activities (Resnick & Resnick, 1991).

Forty states have begun legislating for or are now developing new assessments that are to provide evidence of progress toward standards (Pipho, 1992). Because of criticisms leveled against multiple-choice tests, these states have pressed for the inclusion of performance assessments, ranging from standardized tests with short-answer items (e.g., Maryland) to portfolios of student work (e.g., New Mexico, Kentucky).

Meanwhile, psychometricians, traditional test developers, and researchers have been wary about the move toward performance assessments. They cite the need for test-score reliability and standardized testing conditions as essential elements of fair testing practices. Shavelson, Baxter, and Pine (1992) noted that student performances may vary greatly from one task to another, which leads to questions about the reliability of student level scores when scores are based on relatively few performance tasks. Suen and Davey (1990) stated that relatively short tasks covering a broader array of topics were necessary to standardize responses and improve reliability. Feinberg (1990) has raised concerns about the adequacy of content coverage and the consistency of judgments when performance assessments are used. In addition, Cole (1988) noted that large-scale assessments should be time efficient, cost effective, and centrally processed, all of which is more difficult when complex performance assessments are used. Again, despite these concerns, the enthusiasm for incorporating performance assessments into large-scale testing programs has grown.

From Absolute to Not-So-Absolute Standards

The call for assessments that reflect standards is not a new one. Criterion-referenced testing was originally defined by Glaser (1963) as assessing how students perform in relation to an absolute standard. In much the same way as contemporary proposals for the use of performance assessments, criterion-referenced tests were proposed as a way to help educators focus their teaching (Jaeger, 1991). Educational targets (learning objectives) were identified and tests were developed or selected from published tests to assess whether students were attaining the targets. Test items were intended to assess a sample of the critical knowledge and skills related to the absolute standard. The absolute standard was "a carefully defined domain of content" (Popham, 1978). Several writers attempted to distinguish criterion-referenced tests from norm-referenced tests (Bloom, Madaus, & Hastings, 1981; Ebel, 1972; Hambleton & Novick, 1972; Popham, 1978). The items on the criterion-referenced test or subtest were to measure one carefully defined behavior and content domain or learning objective (Hambleton & Novick, 1972; Popham, 1978) and the results of the tests were to be used to make decisions about the effectiveness of instruction (Hambleton & Novick, 1972). The goal of instruction was for everyone to meet the standard (Bloom et al., 1981; Hambleton & Novick, 1972). Norm-referenced tests, on the other hand, were seen as tests designed to measure separate aspects of learning spread diffusely across some domain (Ebel, 1972), allowing educators to compare students' performance on the test with that of a norm group representing the national population of students.

Gradually, the focus on absolute standards and clearly defined domains of content for criterion-referenced tests began to erode. In the late 1970s, norm-referenced test publishers began providing "objective scores" for clusters of items included in the tests. This step offered educators the illusion that they could decrease testing time by using one test to serve both norm-referenced and criterion-referenced purposes.

Norm-referenced tests are now used by many states and school districts to define the important learning targets. Test scores are more closely associated with funding, teacher salaries, and school district reputation (Madaus, 1988; Smith, 1991). To raise test scores, teachers often focus on the tested skills and concepts in isolation at the expense of skills not easily tested in multiple-choice format (Darling-Hammond & Wise, 1985; Haladyna et al., 1991; Madaus et al., 1992; Shepard, 1991a). In addition, comparisons of students, schools, and districts on scales of achievement have further entrenched the process of labeling and tracking students (Darling-Hammond, 1991). These outcomes result, in part, from the attempt to use norm-referenced tests (tests based on the measurement model) to determine whether students are achieving standards.

The Measurement Model

If we assume that the most important function of our schools is to teach students and that we are successful when students learn what we teach, we must question whether the assumptions of the measurement model apply to such goal-directed activities. The basic assumptions of the measurement model come from *trait theory* or the theory of *individual differences*.

The Assumptions of Trait Theory

"A trait is any distinguishable, relatively enduring way in which one individual varies from others" (Guilford, 1959, p. 6). Stated simply, a trait is a measurable characteristic, like height or weight. In her discussion of the learning theories held by measurement specialists, Shepard (1991b) notes that "traditional psychometrics was developed in the context of individual differences in psychology and focused on static assessment of differences rather than the assessment of changes due to learning" (p. 6). The importance of trait theory in determining methods of test development and the types of scores provided by tests deserves a careful review.

The assumptions underlying this theory are:

1. Humans consistently differ from one another on various human traits.
2. One individual's measurement on a trait can be reported relative to the distribution of other similar individuals' measurements on that trait.
3. We can develop instruments that reliably measure these individual differences on traits.

Individual differences. The history of trait theory in psychology began in the late 1700s. As a result of the research of a German astronomer, it became apparent that human beings differed in consistent and stable ways in reaction time (Lanyon & Goodstein, 1982; Rowe, 1985). Later it was realized that people differed systematically in many other measurable characteristics. This led to extensive investigations of individual differences in psychological and physiological functioning (Galton, 1871; 1889). After collecting data from 10,000 individuals, Galton (1889) employed the statistical methods of a Belgian mathematician to analyze his data. Quetelet had been the first to extend the laws

of probability to investigations of human behavior (Rowe, 1985). Galton added the use of frequency distributions to Quetelet's methods. He postulated that the distributions of intelligence and major physical attributes could be represented by the normal curve.

> The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity . . . amidst the wildest confusion. The huger the mob and the greater the apparent anarchy, the more perfect is its sway. . . . Whenever a large sample of elements is taken in hand and marshaled in order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along. (p. 66)

Relative measurement. Galton's discovery led to the belief that a measure obtained on any physiological or psychological variable for one individual can be reported relative to a distribution of measures of that variable for other people. These physiological and psychological variables were called traits, and subsequent efforts in the science of psychology were made to develop better and better measures of various physiological and psychological traits.

Later derivations of the mathematics of Galton's "beautiful form of regularity" resulted in the psychometric procedures that are now used as aids in (a) establishing the reliability of tests, (b) interpreting test scores, and (c) establishing some confidence that an examinee's score on a psychological test is close to her/his *true* (real) score (Allen & Yen, 1979). These procedures are the substance of introductory measurement courses that highlight the normal curve, "measures of central tendency" (mean, median, mode), score variability (variance and standard deviation), simple issues regarding probability, standard errors of measurement, validity and reliability indices, item difficulties (e.g., p values), item-test correlations (e.g., point-biserials), and some of the common derived scores used to compare students' scores (e.g., percentile ranks, stanines, and normal-curve equivalents). The procedures for obtaining these statistics have become the accepted psychometric methods for establishing the technical quality of testing instruments. Psychometric techniques have been used to transform examinees' responses to sets of items into "rulers" or "scales" for measuring various psychological traits.

Reliable measurement. The primary emphasis for establishing the soundness of an instrument measuring such traits has been placed on the reliability of examinee scores and the estimated size of measurement error. The procedures used for investigating measurement error and obtaining reliability coefficients are based on mathematical models that require score variability, independent test items, and lengthy tests.

First of all, without a reasonable distribution of scores, contemporary mathematical methods for obtaining indices of reliability and measurement error will not function well. "Other things being equal, the more heterogeneous the group, the higher the reliability" (Mehrens & Lehmann, 1991, p. 259). If all examinees respond in the same way to a test item or task (i.e., earn the same score), the

item or task has no measurement value. It will adversely affect mathematical indices of reliability, as well as mathematically increase the index of measurement error for the test as a whole. Second, the items on the test are assumed to be independent of one another. If an examinee's responses to one item or task affects her or his response to another item or task, measurement error is increased by an unknown amount (Yen, 1993).

Test length will also affect mathematical indices of reliability. If examinees complete relatively few items or tasks, numerical indices for reliability may be low. Typical measurement textbooks state that, under specified test conditions, an increase in test length will increase test reliability (regardless of the validity of the test).

> In general, longer tests give more reliable scores. This is true because the random positive and negative errors within the test have a better chance of canceling each other out, thus making the observed score (X) closer to the true score (T). . . . Very short tests or subtests simply give less reliable scores than they would if they were composed of more items. (Mehrens & Lehmann, 1991, p. 258)

Another aspect of scientific methodology that has been applied in order to increase test score reliability is the use of standardized testing conditions. The need for test standardization derives from a laboratory model wherein the "psychometric tester tries to standardize the state of the subject, as well as the test stimuli" (Cronbach, 1970, p. 69). To ensure scoring objectivity, the "procedure, apparatus, and scoring have been fixed so that precisely the same testing procedures can be followed at different times and places" (Cronbach, 1970, p. 27). This is done in an effort to obtain measurements of psychological traits that are consistently affected by the measurement instrument (just as a thermometer will measure temperature consistently for all patients) and to control environmental variables that could affect examinee scores. To obtain standardized testing conditions, the administration directions, practice exercises, type of testing room, and sequence of test items for a psychological test are carefully prescribed.

Validity. Although evidence for the reliability of instruments measuring psychological traits has been of *most* importance to psychometricians, issues of the validity of assessments have been addressed. For the measurement model, the most commonly used methods for obtaining evidence for the validity of the assessments are content reviews (checks to see if the content of the test seems to assess the psychological trait) and various correlational techniques: (a) correlations between item scores and test scores, (b) correlations between two instruments that are supposed to measure the same psychological trait, and (c) correlations between an instrument measuring a trait and some criterion behavior that the trait is believed to predict. In addition, validity evidence is derived through factor analysis and other discriminant validity procedures that are also based on correlational data. Again, other things being equal, the size of these validity coefficients depends upon heterogeneity of scores. If all ex-

aminees obtain the same score on a test, correlational procedures cannot be applied.

The measurement model is based on our beliefs about individual differences in psychological and physiological traits and our understandings about how to reliably measure individuals on those traits. Decisions about the length of a test, the types of items or tasks on the test, distributions of scores, standardized testing conditions, and methods for obtaining evidence for the reliability and validity of tests are driven by scientific and mathematical methods that support the measurement model.

Test Development Using the Measurement Model

The psychometric methods developed for use with physiological and psychological traits have been extended to standardized achievement testing. This was done in an effort to eliminate subjectivity and bias from testing (Feinberg, 1990), as well as to improve the efficiency of large-scale assessment programs. This extension has been so well accepted that psychometricians often consider areas of achievement to be *traits*. For example, Mehrens and Kaminski (1988) compare preparation for a standardized achievement test with practicing an eye chart prior to an eye examination. They have ignored the fact that quality of vision is not *learned*, but is a trait (a fairly stable human characteristic), whereas achievement is a goal-directed accomplishment.

To apply the measurement model to the development and scoring of achievement tests, one must assume that each area of achievement (e.g., reading comprehension, mathematical concepts, vocabulary development) represents a psychological trait and—like height or weight—is distributed differentially (and randomly) in the population. The items in each subtest within a standardized achievement test battery reflect our current definition of one of these traits. Each subtest in the battery has a numerical *scale*. The scale is the *ruler* or *yardstick* for the tested achievement area and is to measure the growth of students from 1st through 12th grade. In a sense, a nationally standardized reading comprehension subtest determines how "tall" a student has grown in reading comprehension compared to his or her peers. The absolute score on an achievement test, however, has no more value than a height of 48 inches. The test score takes on meaning when it is compared with the scores of others, just as a child's height of 48 inches has meaning only when the typical heights of children of the same age are also taken into account. When the measurement model is applied, every effort must be made to make this ruler the same for all students. Figure 1 shows the theoretical connection between trait definition, test performance, and trait scores assigned to individuals. As shown, test performance results in a score on the scale for the trait of reading comprehension. By itself the scale score has no meaning. The scale score is then compared with those of other similar students, and the resulting comparative score (percentile rank) becomes a label for the student. This student has a score for reading comprehension knowledge and skills that is higher than 95% of his or her peers.

Selecting items to obtain individual score differences. Nationally standardized achievement tests are built to optimally *differentiate* between examinees

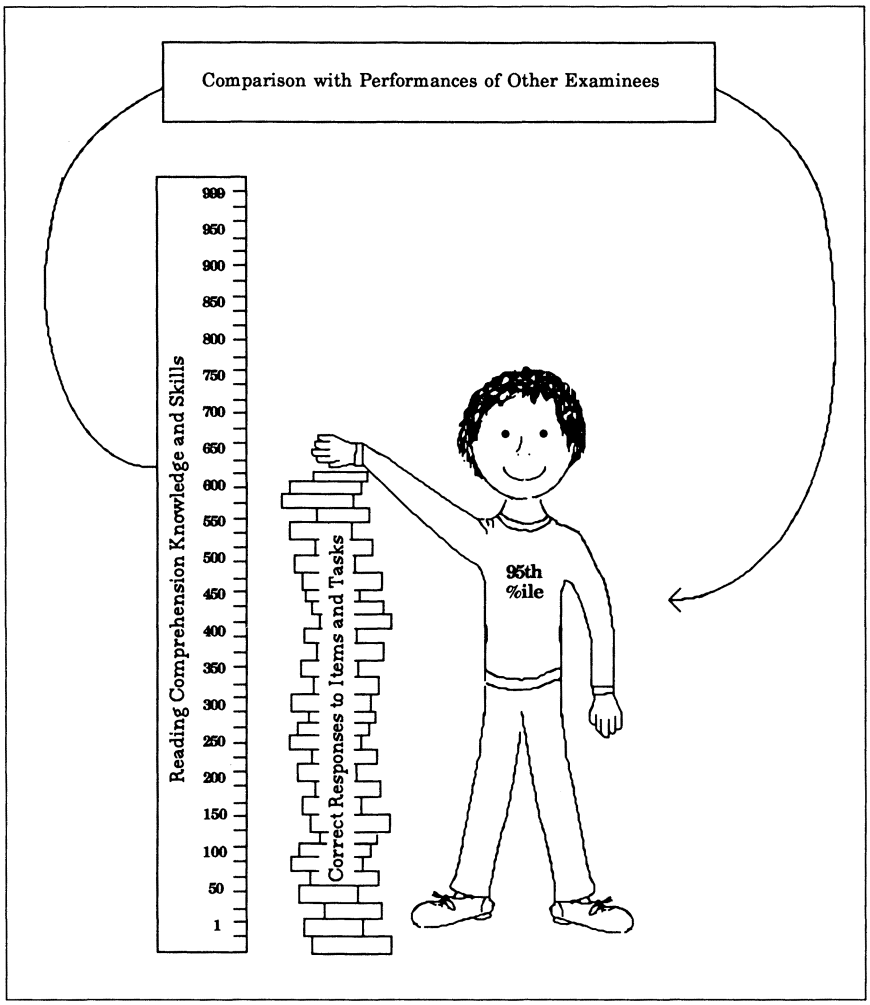

Figure 1. **Theoretical connection between trait definition, test performance, and trait score for reading comprehension**

to compare them. Items are developed and selected based on their value in discriminating between examinees with higher and lower total scores on tests. If items do not function as discriminators, they are discarded. For example, if a mathematics item is answered correctly by all fourth-grade examinees during an item tryout, it has no value in differentiating between fourth graders. The item is then discarded from further use or, if it has curricular merit, it may be used on a third-grade level mathematics test in order to establish a finer distinction between higher scoring students—even if it assesses a concept that is not usually taught until fourth grade.

Although representation of learning objectives within a content area is a consideration when selecting items for a test, the tests are not developed to provide solid information about how well students are learning various objectives. Items are selected first and foremost for their discriminating power. For example, an item might be chosen for a vocabulary test because it effectively differentiates between the 95th and 99th percentiles rather than because it represents *critical* vocabulary that should be learned if students are to be successful at higher levels of reading (an absolute standard). An example of what such an item might look like is given in Figure 2.

Items selected for nationally standardized achievement tests range from very easy to very difficult so that the test can effectively rank *all* examinees. The difficulty of an item is determined by the average performance of a large group of students (the norm group). The assumption is that students with low measures on the *trait* will respond correctly to items that were, on the average, easy for the norm group, students with moderate measures will respond correctly to easy and moderately difficult items, and students with high measures will answer all or most items correctly.

Reliability and validity. The items selected for standardized achievement tests ensure score variability; therefore, correlational procedures used for obtaining evidence for reliability and validity function well with these tests. Strategies used to increase indices of score reliability of psychological tests (e.g., standardized testing conditions and many unrelated items) are also applied in the development of these standardized achievement tests.

Performance-based assessment for measurement. If the methodology of the measurement model is used in the development of performance-based assessments, the tests will be created as follows:

1. Tasks will be developed and selected based on the degree to which they (a) differentiate between examinees at various score levels and (b) yield adequate mathematical indices of reliability and validity.
2. Performance tests will have many unrelated tasks to *mathematically* increase indices of reliability and decrease measurement error. This will be accomplished either (a) by creating many unrelated short-answer items and brief tasks or (b) by using some sort of sampling procedure (e.g., matrix sampling) so that different examinees complete different tasks.

The girl had not eaten for a long time and had become quite

_____.

Which of these words would indicate that the girl had become gaunt?

A. slender
B. sickly
*C. thin
D. weary

Figure 2. **Vocabulary item that might be selected to discriminate between the 95th and 99th percentiles on a sixth-grade vocabulary test**

3. Testing conditions will be carefully standardized.
4. Students will work independently so that scores approximate each individual's *true* measure for the trait.
5. Performance tasks will be carefully scaffolded so that all examinees who complete a task follow exactly the same directions and procedures.
6. Tested content will be unknown to students so that students cannot prepare for the tests.

Using science as an example, examinees would do many separate science tasks of varying difficulty. Some tasks might be as simple as writing a correctly formatted hypothesis statement. Other tasks might be as difficult as requiring students to make predictions based on a theory they had not yet been taught. The function of the test would be to rank the examinees. We would not need to establish standards for well-reasoned applications of scientific methodology nor would we find out whether the examinees can design, conduct, and generalize from an experiment based on their own observations and hypotheses.

By accepting the assumptions of the measurement model, performance tests will be built that reinforce the assumptions of the model. It is a self-perpetuating process. Quantitative methods used for establishing the reliability and validity of student scores will be applied based on these methods' adherence to the model, further reinforcing the model's assumptions. If the assumptions of the measurement model are *not* met (i.e., when all students achieve a standard), psychometric procedures used to evaluate the technical quality of performance assessments will yield poor or uninterpretable information. In addition, correlations between scores on the performance assessments and achievement instruments designed to rank students will be zero. The measurement model *requires* differentiation between and ranking of people rather than the establishment of clear standards or expectations for learners. Excellence is determined by whether an examinee *outranks* other examinees.

The Standards Model

The Assumptions of the Standards Model

To date, thinking about the standards model of assessment has been delayed by concerns about the adequacy of current methodology for developing and scoring performance-based assessments (Linn, Baker, & Dunbar, 1991), as well as concerns about the reliability of student-level scores on performance tasks (Shavelson, 1992; Shavelson, Baxter, & Pine, 1992). Unless we clarify our thinking, it is likely that in developing new assessments for standards, decision-makers and measurement specialists will attempt to apply the same reasoning and psychometric techniques to performance assessments as has been applied to criterion-referenced tests. Such a choice would be short-sighted.

The ideas, models, and proposed psychometrics for criterion-referenced tests developed during the 1970s are predominantly based on the use of the same testing format and, more importantly, some of the same assumptions about the measurement of human learning as are used for nationally standardized achievement tests. High-stakes criterion-referenced tests are given during a single

sitting. The attempt is made to obtain reliable *measurements* of student learning using a standardized instrument, and students are not supposed to be aware of the specific content on the test (although they are to be aware of the tested objectives). In fact, early conceptions about criterion-referenced testing came from measurement specialists who were working to create a defensible level of psychometric *soundness* for criterion-referenced tests.

In contrast to the measurement model, the standards model suggests a very different set of assumptions. These assumptions are that:

1. We can set public educational standards and strive toward them.
2. Most students can internalize and achieve the standards.
3. Very different student performances and exhibitions can and will reflect the same standards.
4. Educators can be trained to internalize the standards and be fair and consistent judges of diverse student performances.

Setting standards. There are important differences between current visions about assessment for standards and the ideas articulated by those who were writing about criterion-referenced tests during the 1970s. One important difference is that the emphasis is now on standards for what students can *do* (student performances) rather than simply for what students *know* (a defined domain of content). Once the desired outcomes of education are articulated, educators not only must define the domain of content for a discipline, but they are also challenged to identify and define the complex performances and processes that are "authentic" to that discipline (Wiggins, 1989).

A second difference is in how standards are to be established for large-scale assessments. During the 1970s, measurement specialists described a plethora of methods for setting standards for criterion-referenced tests (Angoff, 1971; Berk, 1976; Block, 1972; Darlington & Stauffer, 1966; Ebel, 1972; Emrick, 1971; Nedelsky, 1954; Millman, 1972; Zieky & Livingston, 1977). The purpose of these methods was to set passing scores or *cut-scores* on large-scale objectives-based tests. The methods ranged from procedures that required judgments about items or objectives to methods that were based on observed or hypothetical score distributions for masters and nonmasters. Each of these standard-setting methods assumed that the domain of content had already been determined and that the method was being applied to test scores or test items. Only one class of standard-setting methods involved making judgments about students' levels of competence (Zieky & Livingston, 1977). None of the methods required judges to look at student work and evaluate that work as it related to standards of quality and performance criteria for defined performances.

Contemporary writers focus on student work. They state that clear criteria for student performances (*performance criteria*) can be established. These are the specific requirements of performances, including the knowledge, concepts, skills, and processes that must be exemplified in a performance or collection of performances (Stiggins, 1988). Along with these performance criteria come *performance standards*, or levels of performance quality that are considered excellent (Wiggins, 1990). Examples of student work (*exemplars*) that repre-

sent the standards and criteria are then obtained to make these statements of expectations concrete and tangible. These desired standards, criteria, and exemplars then become part of the public domain.

> The ultimate test of whether standards are demystified is whether students and teachers can accurately assess their own work on a regular basis; to do so, they must be able to compare their performances with exemplary work (Wiggins, 1990, p. 25).

Internalizing and achieving standards. Another difference between early ideas about criterion-referenced testing and contemporary notions about assessment for standards is the degree to which assessment must be independent of the instructional process. Early writers were comfortable with the large-scale criterion-referenced tests that were independent of the instructional process. Contemporary writers see the assessments as central to the instructional process (Resnick & Resnick, 1991; Wiggins, 1989).

> The true test is so central to instruction that it is known from the start and repeatedly taken because it is both central and complex—equivalent to the game to be played or the musical piece to be performed. The true test of ability is to perform consistently well tasks whose criteria for success are known and valued. (Wiggins, 1989, p. 36)

> Assessments should be designed so that when teachers do the natural thing—that is, prepare their students to perform well—they will exercise the kinds of abilities and develop the skills and knowledge that are the real goals of educational reform. (Resnick & Resnick, 1991, p. 59)

A unique aspect of current visions of the standards model is that the examinees are examined when they are ready and students accumulate performances over time rather than in "a single high-stakes moment of possible failure" (Resnick & Tucker, 1990, p. 21).

In its purest form, assessment for standards does not require scores. Students and teachers internalize the standards and strive toward them. A student's work either does or does not achieve the standards for that type of work. Students are taught the features and qualities (criteria) for excellent work and may improve upon their work so that they attain standards. Students may need varying amounts of time and differing instructional methods, but the goal is the same for all: achievement of the standards (Bloom et al., 1981; Hambleton & Novick, 1972). Again, this is not a new idea.

> Education is a purposeful activity, and we seek to have the students learn what we teach. If we are effective in our instruction, the distribution of achievement should be very different from the normal curve. In fact, we may even insist that our educational efforts have been *unsuccessful* to the extent that the distribution of achievement approximates the normal distribution. . . . "Individual differences" in learners are facts that can be demonstrated in many ways. That students vary in many ways

> cannot be forgotten. . . . The basic task in education is to find strategies which will take individual differences into consideration but which will do so in such a way as to promote the fullest development of the individual. (Bloom et al., 1981, pp. 52–53)

Student performances and exhibitions. The specific performances that reflect the standards can differ dramatically. Music contests are an example of how the standards model is already being implemented. Each performer may select a different musical composition and use a different instrument, but expert judges evaluate these different performances and identify the ones that meet the highest standards of quality. This requires a set of performance criteria (the specific features that must be included in the piece to be performed, such as length, interpretive markings, key signatures, and time signatures), the level of quality expected (technical accuracy of the performance, performance technique, musical interpretation), and often a risk factor (a particularly difficult section played well). The number of different performances that could qualify as excellent and as meeting the criteria is infinite.

Professional judgment. The standards model requires the use of professional judges of student performances. Educators who are knowledgeable in the subject matter are trained to internalize the standards and to be familiar with the performance criteria.

> Using performance assessments as part of public accountability programs would require that students' performances be evaluated by panels of judges other than the students' own teachers. These judges must, of course, be trained to apply agreed upon criteria for performance. . . . Strategies for training judges, assessing interjudge reliabilities, and maintaining reliabilities through periodic sessions in which judges review and discuss each others' ratings have been developed by various groups that have long used open-ended performance assessments in education. (Resnick & Resnick, 1991, p. 70)

Judges of student performances must be knowledgeable of the structure and content of the discipline for a given performance. "It is impossible, for example, for a teacher to assess a student's level of writing proficiency if that teacher does not clearly understand the attributes of good writing" (Stiggins, 1992, p. 36). This is true for every discipline. Judges who do not truly understand mathematics will not be able to judge the adequacy of a student's strategies for problem-solving if the student is innovative and has used a combination of mathematical strategies and concepts that are not commonly used or combined. For example, if a third-grade student uses a novel technique for regrouping during multiple-digit whole-number addition, unless the judge understands the underlying structure of regrouping in mathematics, he or she may be forced to focus on the *accuracy* of the student's final solution rather than on the *depth of understanding* of place value evidenced in the problem-solving *process*. If student performances require combining processes across several disciplines (e.g., physics and mathematics), the judges must be thoroughly versed in each

discipline. Otherwise, assessment will be narrowed to the technical accuracy of student work.

Reliability and validity. The quantitative methods now used to gather evidence for the reliability and validity of assessments within the framework of the measurement model will not function for the standards model. If we work to ensure that students achieve standards and students are not assessed until they are prepared to do so, correlational procedures used for obtaining evidence for validity and reliability of assessments will yield poor or uninterpretable information. Some measurement specialists claim that the move to performance-based assessments represents a paradigm shift (e.g., Linn et al., 1991; Trevisan, 1991) and may require new ways to gather evidence for the reliability of assessments because the function of assessment is no longer the identification an individual's true score at a single point in time. Instead:

1. The standards model allows for the assessment of student work that takes an extended period of time to complete, as well as the collection of student work samples across time. Growth during the period in which work is gathered is not of concern. The important question is whether the student has achieved the standards.
2. The standards may be achieved early in a student's career for mathematics and late in a student's career for writing. Students may have many opportunities to demonstrate their achievements, yet the desired standards remain the same.
3. The standards model allows for performances that include collaboration as well as cycles of feedback and revision. Collaboration and revision cycles are typical aspects of adult work; therefore, authentic performances may involve either.

The reliability of the student's assessment, then, is a matter of the degree to which the body of evidence for the student in any given area gives a clear message about whether or not the student has achieved the standards. Indeed, given that the focus of the standards model is not on the static assessment of individuals using controlled methodologies, new definitions and methods for investigating reliability are a likely consequence of the adoption of the standards model.

As with other shifts in thinking that arise from adopting the standards model, issues related to standardization must also be addressed. Standardization procedures are often used to enhance the reliability or dependability of test scores. Despite Wiggins's (1990) call for standards rather than standardization (p. 25), the standards model does not mean that standardization is eliminated. Identifying important educational outcomes, defining performances that reflect the outcomes, clarifying performance criteria and performance standards, and obtaining examples of student performances that show excellence are all ways of standardizing the assessment process. For the standards model, however, there is a recognition that learning and performance are contextual and not simply a function of the individual's ability (Resnick & Tucker, 1990; Wiggins, 1989). Therefore, although the standards themselves may be "standardized," the stu-

dent performances that are held up to those standards can differ from individual to individual and from group to group.

The primary focus of the measurement model has been on reliability; however, validity is the primary emphasis for those who write about the standards model. Wiggins (1989) states that authentic performances that are evaluated based on clear standards are "true tests" (p. 706). It is not sufficient, however, to simply make judgments about the degree to which performances are more direct assessments of desired student learnings. Methods must be developed to investigate such validity issues as whether students who achieve standards are truly prepared for work later in life, as proponents often claim. In addition, if unique performances are to be compared to the same standards, the validity questions that must be carefully investigated are (a) whether performances that reflect differences over time and across groups and settings can attain proposed standards (Messick, 1990) and (b) whether judges can evaluate performances that differ substantially in content or structure but are intended to reflect the same performance criteria and performance standards.

Test Development Using the Standards Model

Large-scale assessments currently serve two important assessment needs: They provide accountability information about schools and districts, and they establish a consistent standard of measurement for students. Unless both of these assessment needs are met through the standards model, efforts to replace assessments based on the measurement model will fail. For this reason, the most critical aspect of the work for the standards model is that of identifying the essential performances in given disciplines, establishing standards and criteria for those performances, obtaining examples of performances that reflect those standards and criteria, and communicating all of this to the public. This process requires multiple iterations of information gathering and decision-making. In fact, for the current standardized achievement tests, much of this process is done by teams of content editors and project directors who work for testing companies. The standards model gives responsibility for these stages of information gathering and decision-making processes to educators and educational stakeholders (parents, legislators, community leaders, and students). This helps to ensure that standards are public.

Setting standards. The overall assessment process for the standards model is very different from assessment methods based on the measurement model. First of all, before standards can be set, educators and educational stakeholders must articulate their values and expectations in words and examples that can be understood by students as well as teachers. The relevant parties must work together to identify important and tangible outcomes of education. This process is already being done in many states, school districts, and national professional organizations (e.g., California Curriculum Frameworks, Connecticut Common Core of Learning, and the curriculum standards documents prepared by groups such as the National Council of Teachers of Mathematics, the National Academy of Science, and the National Council of Teachers of English).

The next step in the development of standards is to establish "benchmarks" or descriptions of the types of student performances at different developmental levels that are central to each discipline and that reflect the desired outcomes. For example, if mathematical problem solving is considered a desired outcome, primary children may be expected to investigate fairly simple mathematical concepts whereas high school students may be expected to investigate several complex mathematical concepts and the interrelations between the concepts, using computer technology to develop graphic models of the mathematical ideas.

Along with benchmark performances, educators must define performance criteria (the important features, knowledge, skills, and thinking processes) for performances at each developmental level. Identifying performance criteria is one of the most difficult tasks in the process of establishing standards for two basic reasons: (a) educators may disagree about what criteria are essential for excellent work and (b) educators may disagree about what is possible at a given developmental level. For example, whereas some educators believe that the most important performance criteria in writing are those related to rhetorical style and organization, others insist that language conventions (grammar, punctuation, capitalization, spelling) and use of language (e.g., varied sentence structures, broad vocabulary) must carry equal weight. In debating these issues, the context of the writing cannot be overlooked. If students are given time to prepare portfolios that show examples of work for each important rhetorical style, work that has been reviewed and revised based on criteria for organization, language conventions, and language usage, then a broader set of performance criteria is possible. If, however, the writing must be completed in a performance examination wherein less than an hour is devoted to a written piece, a less stringent set of criteria is more appropriate. In addition, for schools where writing workshops are part of the instructional process beginning in first grade, expectations for writing may be higher than in schools that have extensively used grammar, punctuation, and capitalization worksheets for writing instruction. Despite the inevitable impact of the context on the decision-making process, these debates must take place.

Once performance criteria are determined for each developmental level, the standards are then established by supplementing the performance criteria with examples of student work that reflect the criteria and represent the desired quality of work for each developmental level. Standards are not based on average performances. Instead, standards represent expectations for *performances of quality.* The process of setting standards takes time and requires explicit discussions about general expectations for students. It is the first stage of standardization.

An important consideration in setting standards is that the criteria for performances must be relevant to the outcomes to be assessed and that the exemplary performances in each area extend beyond written work. Clearly, writing must be demonstrated through writing; however, mathematical problem solving can be demonstrated through a variety of mediums that may include writing, computer modeling, three-dimensional models with videotaped oral presentations, and so forth. The critical task then is to establish criteria that are directly related to the various learnings students are to demonstrate. Student per-

formances can then be developed based on individual strengths rather than formulated in a way that is biased in favor of one group or another.

The process of setting standards is a difficult one. Disseminating standards in a form that teachers and students can understand and internalize is more difficult. It requires making the standards and criteria public and assisting teachers in understanding them (Wilson, 1992). This process has been undertaken in California, where several documents provide teachers with samples of student work to exemplify standards in writing and mathematical problem solving (California Assessment Program, 1989, 1990, 1991). California teachers also particpate in scoring sessions designed to help them internalize the standards and criteria.

Student performances and exhibitions. Resnick and Tucker (1990) have listed three types of performance assessments: performance tasks, performance examinations, and portfolios of students' work. Performance tasks may include work such as science labs, mathematics investigations, critical essays, oral presentations, or any other performance that is appropriate to the given discipline or interdisciplinary endeavor. Performance examinations (similar to doctoral comprehensives) may include a broad array of task types ranging from written work to complex problem-solving tasks. The major distinction between performance tasks and performance examinations is that performance tasks are completed in the context of instruction and performance examinations are administered in much the same way as any other high-stakes examination. Portfolios include collections of related works such as written works (e.g., expository, narrative, persuasive), mathematics projects and investigations representing a range of mathematical content and processes (e.g., geometry, measurement, statistics, and problem solving), science projects and lab summaries, as well as other related collections. The function of portfolios is to gather a broad range of evidence together to show that all standards within a given domain have been met.

Each assessment type can be used to reflect different aspects of standards. For example, if depth and breadth of subject-matter *knowledge* is considered a critical learning target, performance examinations may be used to reflect the students' *breadth* of knowledge. Performance tasks may be used to reflect *processes* seen as central to a subject discipline. These tasks may take several days or weeks to complete and may include feedback and revision stages if such processes are important in the discipline. Finally, student portfolios can be used to gather several examples of student work that show a breadth as well as depth of understanding within and across subject areas. Student work would be selected for inclusion based on the degree to which the work reflects the standards and performance criteria for the learning outcomes (Arter & Spandel, 1992; Paulson, Paulson, & Meyer, 1991; Valencia & Calfee, 1991, Valencia, 1990).

Professional judgment. Although assessment for standards does not require scores, a great deal of information about student learning is lost when there is no vehicle for evaluating performances that do *not* meet the standards. When students have not achieved the standards, they may need guidance about how to improve. Teachers may need help in evaluating students who fall below the

standards. These needs necessitate the development of scoring rules (rubrics). When these scoring rules are developed, a second level of standardization occurs.

Complex performances can be evaluated using *holistic* scoring rules or rubrics (Wilson, 1992). These procedures require evaluation of the overall effectiveness of the work from one or more perspectives. For example, a position paper may be scored for strength of the case made for the position (a social science performance standard). If, on the other hand, the position paper is used to reflect a social science standard, a writing standard, and a word-processing technology standard, the paper may be scored once for strength of position, a second time for overall writing quality, and a third time for effectiveness of computer use in creating an appropriate text layout and in developing graphic and tabular support for the paper. Another method for scoring performances requires analysis of separate specific dimensions of a work. This method is called *analytic* scoring (Spandell & Stiggins, 1990). Separate scores are given for each of the important dimensions of the work (e.g., writing conventions, rhetorical style, organization, language usage for a written work).

The use of such holistic and analytic scoring techniques does not automatically mean that the standards model is being applied, however. Standardized tasks of differing difficulty and discriminative power can be developed, administered, and scored. The scores can be "scaled" and reported in terms of relative performance in the same way as are nationally standardized achievement tests. The use of holistic and analytic scoring procedures reflects the standards model only when clear, external standards of excellence exist against which all performances are judged.

The process of creating scoring rules for complex performances begins with the standard (performance criteria and examples of excellent student work) for each targeted performance. To further develop analytic or holistic scoring rules, a fixed number of points in the range of possible performances is selected (e.g., the Maryland State writing program uses a 4-point scoring range and the California Assessment Program writing assessment uses a 6-point scoring range). This range must be one that allows for unambiguous distinctions in levels of performance (W. Yen, personal communication, July 20, 1990) and one that can be used to define the score range for excellent to poor performances. The range is not norm-referenced, in that excellent and poor performances are not established simply by describing the range of *typical* performances for a grade level. Using typical performances as the basis of standard setting could result in an erosion of standards and "teaching to the minimums" (Ebel, 1973). The description of excellent performance reflects the standard, and the description of poor performance reflects an attempt to complete a given task with many features that are poorly executed.

Once the range of performance is established, examples of student work are obtained for each point in the range. Scoring guides assist judges in scoring student work and assist teachers and students in achieving the standards. These guides include the performance criteria as well as descriptions and examples of performances for each point in the range. Figure 3 shows an example of

Performance: Essay on Character Development in Literature

Performance Criteria

- character is identified
- at least three aspects of the character's development during the course of the story are described
- appropriate support for each character aspect is given using excerpts from the story
- character's contribution to the story's plot is described
- at least three excerpts from the story are given as support for writer's ideas about the character's contribution to the story
- text references used for support are appropriate

Scoring Rubric

4 points	Eassay is complete, thorough, and insightful in describing the character's development and contribution to the story. Adequate support is given to encourage us to consider the writer's point of view. All excerpts from the text enhance our understanding of the writer's view of the character.
3 points	Essay is complete in describing the character's development and contribution to the story. Adequate support is given to encourage us to consider the writer's point of view. Most excerpts from the text enhance our understanding of the writer's view of the character.
2 points	Essay is complete in its description of either the character's development *or* the character's contribution to the story. Some support is given to help us consider the writer's point of view. Most excerpts from the text enhance our understanding of the writer's view of the character for the element described.
1 point	Essay is mostly complete in its description of either the character's development or the character's contribution to the story. Support is given for the writer's point of view but it is not always convincing. Few excerpts from the text enhance our understanding of the writer's view of the character for the element described.
0 points	The written essay was not completed, is significantly lacking in performance of all criteria, or is off task.

Figure 3. **Sample scoring rubric: Targeted performance, performance criteria, and a description of performances at different score points**

a scoring rubric for a reading task designed to assess students' understanding of character development and the role of characters in a plot. To complete a scoring guide, the rubric would be supplemented with diverse examples of student work for each score point.

Internalizing and achieving standards. The next stage in the development of assessments for standards is the most difficult to implement. As will be discussed in the final section of this article, this stage requires a high degree of professionalism among teachers. This stage also requires public awareness of what is to be assessed and the standards of expected performance.

During this stage, teachers work with students to design performances that reflect the performance criteria. Students then work on these performances to bring them up to the standards. In keeping with the idea that students are not assessed until they are prepared, students engage in performance examinations when, in the judgment of the teachers and the students, they are prepared to perform well on the examination. Students are given ample practice with the types of performance tasks included in the examination as well as opportunities to work with the tested concepts.

Students work on projects until they and their teachers believe that representative tasks have achieved the standards and exhibit the performance criteria. At that time, the students' works are submitted for formal evaluation. As with other assessments for standards, portfolios are submitted for formal evaluation when, in the judgment of teachers and the student, they include the specified works, all of which reflect the standards and criteria. For any of these performance assessment types, all targeted standards must be public, and both teachers and students must know that students are expected to achieve all of the standards.

Students (with the guidance of teachers) may then select unique combinations of skills and concepts in a task and still develop a performance that meets or exceeds the standards. Teachers must understand how to guide students in developing projects that meet standards and criteria and that capitalize on students' strengths. Therefore, classroom teachers who are helping students work on their performance tasks or projects must be prepared to analyze carefully student work in light of performance criteria and the performance standards.

Reliability and validity. To obtain evidence for the validity of assessments based on standards, several types of research must be conducted. Ongoing research must be conducted, including reviews of collections of students' work for representativeness of the thinking processes, knowledge, and skills they are designed to assess and individual interviews with examinees to investigate the thinking that underlies their performances. These studies will provide evidence about whether the student performances actually reflect the outcomes they were intended to reflect. Ongoing investigations of the consequences of the interpretation and use of assessments must be conducted (Messick, 1990). In addition, background variables that affect performance must be investigated; and success rates for individuals in different groups, across time, and in different contexts must be gathered to determine whether the standards or the judgment process has resulted in biased assessment conditions for any groups of examinees (Messick, 1990).

When the standards model is applied, many student performances are collected over time. Reliability can be established by determining whether the collection of student performances is internally consistent. Quantitative methods proposed for criterion-referenced assessments, such as decision-consistency formulas (Cohen, 1960) wherein scores for a group of examinees on two performances or across two judges are compared for consistency, can be applied to performance examinations. These methods investigate whether judgments made about students' works are consistent across judges or across different perfor-

mances for the same individuals. Decision consistency models are limited, however, when all students develop unique portfolios or projects. One recommended method for enhancing the reliability of judgments about examinees' portfolios is the use of interviews wherein examinees discuss components of their portfolios with judges (Wineburg, 1993). Finally, Moss, Beck, Ebbs, Matson, Muchmore, Steele, and Taylor (1992) have recommended that qualitative research methods be used to establish credibility for the evaluations of portfolios of student work.

Application of the standards model to performance-based assessment can also lead to a self-perpetuating system. When standards and criteria for essential performances are public, teachers can focus instruction on achievement of the standards and criteria, thereby helping to insure students' success. The standards model does not use numbers as proxies for observed phenomena. Therefore, if the standards and criteria set are minimal, if the types of performances are inconsequential, and if the performances are not authentic to the discipline, then the performances, the standards, and the criteria are public and available for scrutiny and comment. Excellence can been seen rather than inferred. Whereas some states and school districts have done a fairly good job of communicating these standards and criteria to educators, the need for accountability will not be met until greater efforts are made to communicate these standards and criteria to all stakeholders. Once this is accomplished, stakeholders must then be informed on a regular basis about how well schools are helping students attain the standards (e.g., what proportion of students have attained given benchmark standards each year). Table 1 provides a hypothetical method for reporting success in attaining standards.

Table 1

Percentage of Students at Each Grade Level Achieving Benchmark Writing Standards

	Standards			
Grade	Primary	Intermediate	Middle-School	Graduation
3	75			
4	87			
5	95	46		
6	100	78		
7		88		
8		100	38	
9			75	
10			86	46
11			100	79
12				94
Post Grade 12				100

Implications for Educational Reform

In discussing the potential impacts of the measurement and standards models on educational reform efforts, I have not addressed the implications of each model for individual high-stakes assessments. It is beyond the scope of this article to discuss in detail how each model fits within the current need for assessments that can, for example, be used to select individuals for special programs or for entry into prestigious universities and/or to determine whether individual students have attained graduation standards. Such a discussion is needed. However, in this article I will focus on the implications of the use of each model in large-scale assessment programs that are intended to evaluate the success of educational programs. Inevitably, such large-scale programs affect individual students. These effects are systemic. Even if our individual assessments were completely unbiased and perfectly valid, even if the inferences we made about individual students were perfectly reliable, large-scale assessment programs that adversely affected systems would adversely affect the students within those systems. For this reason I have chosen to focus on the implications of each assessment model on the systems as a whole.

Because of the differences in how each model of assessment is implemented, policymakers must make a choice regarding the model to be put in place for large-scale assessment programs. Making an informed choice requires all those who will participate in the decision-making process to (a) be knowledgeable about the assumptions underlying each model, (b) reflect on and consider the purposes to be served by the assessments (measurement of the current status of students or movement toward explicit standards of excellence), and (c) consider the potential influences of each model on school restructuring efforts.

The first step in this process is to make explicit to all decisionmakers the choices to be made and the assumptions that underlie each choice. Unless the underlying assumptions of the models are put in language all educators and policymakers can understand, the tensions between assessment for standards and assessment for measurement can result in a failure to achieve the changes expected from the use of performance-based assessments. Once we educators and policymakers understand the assumptions and the tensions, a choice must be made. Do we continue creating instruments that are designed to rank and compare students, or do we want assessment systems that give us clear ideas about whether students are achieving complex learning targets? Can we accept the former model knowing that our current methods for standardized assessment have had many deleterious consequences? On the other hand, are we willing to articulate our standards and to argue for our hopes and expectations?

As a nation we do not all agree on the purposes of schools. Do we believe that schools are supposed to sort students to find the brightest and the best, or do we believe that our democracy will be stronger if we foster the creativity and capacity of every individual? A true choice between models requires much public debate over these very questions. In choosing between the standards model and the measurement model, we will have made an implicit statement about what we believe to be the purpose of schools. Current rhetoric

suggests that the focus of the educational reform movement is to raise standards for *all* students and to ensure that *all* students succeed (America 2000). In what follows, I argue that the standards model is the assessment model that will truly support these educational reform efforts.

The influences of each assessment model on our ways of thinking about learners and about our tasks as educators cannot be ignored. Ultimately, if real change is to take place in the quality of students as they leave our schools, we have to make real changes in our beliefs about learners. We must begin to believe that *most* students are quite capable of learning and achieving; that the dramatic differences we see in student performance are the result of conditions unrelated to students' capacity to learn. It may be difficult, however, to change beliefs about the innate abilities of students. The daily language of educators, policymakers, parents, and students is comparative. Tests that rank students have become a powerful vehicle for making those comparisons. Although psychometricians do not claim that student performances on standardized achievement tests are solely a function of variations in innate ability, nor that a single test score is adequate to describe the true achievement level of the examinee, the real consequence of the use of these tests is that they have affected what we believe about children's capacities to learn. The concerted application of the standards model and the suspension of the use of tests designed to rank students will be necessary to make a significant impact on these beliefs.

Change to the standards model will require significant and long-term changes in schools. Schools will be expected to ensure the success of learners in attaining the standards. Can the existing educational systems break free of underlying assumptions about differential abilities in learners and make the changes that are needed? They can with real and *sustained* support. Although this will not be a simple task, if we accept the assumptions of the standards model, we are very likely to create or maintain an educational system that reinforces that model. If we set clear and public standards for student work and help schools and communities provide learning environments that help students meet those standards, we are very likely to succeed. We will have to provide support to those schools and districts that are willing to make the changes necessary for students to meet standards. We will have to challenge schools and districts that resist change.

Yet schools alone cannot insure that students meet standards. To help students achieve standards of excellence, we have to address the impact of political, social, and economic conditions on children. We cannot ignore the dramatic differences between children as they enter school. We will have to face the fact that not all students are equally open to learning. We will have to accept that not all students are equally prepared to learn; that some are learning despite dramatic obstacles while others are nurtured by supportive environments. We will have to address the differential conditions that affect learning. If we choose the standards model, we must be willing to go to whatever length is necessary to make our schools and communities places that foster learning. We will have to challenge any institutional, economic, and legal structures that prevent students and teachers from attaining standards.

At the same time that we address the political, social, and economic conditions that affect students and schools, the standards model will require significant changes in the currently accepted organizational structure of schools. To ensure that all students achieve standards, student-teacher ratios will have to be smaller. Teachers and students will need fewer disruptions and time constraints in their activities so that students can take the time necessary to think through complex problems. For example, in a recent visit to a third-grade classroom, I watched as an experienced teacher spent from 10 to 15 minutes on each subject area (multiplication and division fact drill, group reading, vocabulary, silent reading, handwriting, and social studies), moving the children from one subject focus to another with few behavioral disruptions. In addition, she took her students to the library, recess, an assembly, lunch, lunch recess, and the computer lab, all in the space of 4 hours. I wondered, as I watched this parade of events, whether the students were learning anything. From the puzzled faces, I suspect that many got lost at some point. No effort was made to reach closure with any of the concepts she was teaching. There was no time to ensure that every student was on the same page or had located the same place on the worksheet. Clearly, with the student-teacher ratio and the constant interruptions, the teacher did not have time to attend to their individual needs and *cover* the skills she was expected to teach.

Rather than insisting that every teacher teach the *volume* of information they are now expected to cover, we will have to support teachers as they focus on effective teaching of fewer, more central concepts; as well as teaching students how to *access* and effectively *use* information themselves. School administrations will have to create institutional structures that support teachers, and teachers will have to adopt instructional practices that foster learning for each individual learner.

If schools are effective in communicating standards and criteria, students can become more self-directed in their learning. They may work toward attainment of different standards, choosing areas of greater interest first and then broadening to areas of lesser interest. Individual students may attain the standards at different times. Teachers will have to be cognizant of standards for higher and lower developmental levels so that students who have attained standards for a given developmental level can remain with their age mates.

The standards model suggests that essential performances are those that are central to a discipline. If teachers are to build instruction from essential performances or if teachers are to be able to judge whether students are prepared for performance examinations or to submit their work for formal evaluation, they will have to be good judges of student performances within the assessed disciplines. This requires (a) subject-matter knowledge, (b) an understanding of the processes that are central to different disciplines, and (c) pedagogical strategies that help students approach each discipline in appropriate ways. Teachers will have to become professionals within their disciplines. School administrations will have to support teachers in ongoing professional development within the disciplines they teach (Wilson, 1992).

Because expertise in several disciplines is difficult, elementary teachers may

have to work in teams rather than as distinct managers of single classrooms. In addition, if some of the desired outcomes of education require students to cross disciplinary boundaries to develop their work, teachers at all levels will have to collaborate and to develop ways to help students integrate their learnings across disciplines. In either case, schools will have to be structured in such a way that gives teachers time for collaboration with one another.

Teachers will also have to be competent assessors of student work. Research suggests that most teachers have received almost no training in classroom assessment during professional preparation (Schafer & Lissitz, 1987). Along with support for ongoing professional development in the subject areas, teachers will need professional training in appropriate classroom assessment practices. Teachers will have to become skilled at varying tasks to meet the individual needs of students so that reading and writing are not the only vehicles for assessment. The decision to use the standards model means that all schools will have to become places that foster learning. This will take time. To implement an assessment system based on the standards model will also take time.

In summary, if we accept the standards model, we will be forced to make difficult changes. We will have to make clear and public statements about the expectations and hopes we have for students. We will have to ensure that students achieve the standards. If we no longer accept differential student performance as the inevitable outcome of differences in *innate* student abilities, we will be forced to address the situations that prevent students from attaining standards, whether these be economic and social conditions or poor classroom instructional practices.

We will also have to find ways to respond to the very real needs that our communities have for accountability. This will require making significant efforts to improve our level of communication with parents and communities about the standards of work expected of students and about how well we are helping students attain those standards. We will no longer be able to let numbers communicate for us (numbers that parents and educators alike have long misinterpreted; numbers based on test content that is closed to public view). We will have to place responsibility for defining the important targets of learning in the hands of professional educators. The two greatest dangers we face in implementing the standards models are that we will not be willing to take the time needed, demanding instant results and failing to support the process of change; or we will be tempted to return to comparative discussions and blame the learners when we are not immediately successful.

The decision to use the measurement model for performance-based assessment will essentially result in the same self-defeating practices that are now common. First, we will continue to build tests based on a theory of individual differences. We will rank students, even if real differences in achievement are small. In all likelihood, we will place the blame for below-average test scores on the "defective abilities" of the students. We may change our scoring methods and our numbering systems, but students will be labeled and tracked, and schools and districts will be stigmatized. The consequences of labeling and tracking for students will remain unchanged.

Second, these tests will be used to define the critical content and concepts. If the overall test is based on a poor definition of some domain and if the definition does not reflect performances that are appropriate for a given discipline, performance tasks will be developed based on the erroneous definition. Tasks will be selected based on their relation to overall scores on a test that is measuring an invalid domain. On the other hand, if substantial changes occur in our understandings about a discipline between the time a test is developed and when it becomes obsolete, schools will be reluctant to change their modes of instruction for fear that this will lead to poor test performances.

Third, testing will remain in a shroud of secrecy. Tasks will be limited to those that are independent of the learning context. Many small, unrelated tasks will serve as proxies for the authentic performances required in real life. Teachers and students will be asked to prepare for exams that are "secure" so that differential coaching does not lead to score differences that are more a reflection of the effectiveness of instruction than of the students' "true" abilities. The higher the stakes, the more likely that students and teachers will use questionable strategies to ensure higher ranks on the tests (Haladyna et al., 1991; Smith, 1991). Many of the same instructional and institutional practices that have come about because of the stakes placed on standardized achievement tests will continue to occur. Teachers may teach the tested performance tasks rather than focus on the central understandings of various disciplines (Shepard, 1991a). Students will continue to take courses designed to raise test scores. Teachers will continue to use curriculum materials that are closely aligned with the content and format of the tests.

For the measurement model, few structural changes will be required in schools. Tests will be external to the ongoing work of schools, and instructional practices that focus on discrete tasks within fixed time periods will continue to be used as ways to enhance test performances. Existing structures can remain intact. The current practice of assessing each subject area in isolation to obtain *pure* measures of *traits* may continue to dominate assessment practices, which can limit methods such as "writing across the curriculum" and "interdisciplinary teaching."

Finally, by choosing the measurement model, we will create assessment systems that simply reinforce what we already know—that test performance is highly correlated with social and economic conditions. We need not face the difficulties of structuring schools in ways that help students achieve standards. We need not face the social and economic changes that are necessary for success. We will continue to have educational systems that accept "failure" (below-average performance) for some, "mediocre" (average) performance for some, and "success" (above-average performance) for a few. The measurement model will provide no standards of *quality* toward which we want our students to strive.

Today we have to decide what tools we will use to support our schools in their efforts to create learning environments that foster important learning. The implementation of performance-based assessment systems based on the measurement model will predominantly support schools whose students already

perform well on traditional achievement tests, as well as fostering practices designed to yield higher scores on high-stakes tests. This being the case, for the measurement model, we do not need expensive, new performance-based tests.

On the other hand, implementation of performance-based assessment systems based on clear and public standards can support all schools in reaching the goals recently espoused by educators and policymakers throughout the United States. Despite the inevitable challenges of change and the public debate that must occur, despite the inevitable shifts in and debates about how we define learning and excellence, if the standards model of assessment is implemented and sustained, it is the model most likely to support real changes in our schools.

References

Allen, M., & Yen, W. (1979). *Introduction to measurement theory.* New York: McGraw-Hill.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In. R. L. Thorndike (Ed.), *Educational measurement.* Washington, DC: American Council on Education.

Arter, J. A., & Spandel, V. (1992). Using portfolios of student work in instruction and assessment. *Educational Measurement: Issues and Practice, 11*(1), 36–44.

Baron, J. B. (1991). Strategies for the development of effective performance exercises. *Applied Measurement in Education, 4*(4), 305–318.

Berk, R. A. (1976). Determination of optimal cutting scores in criterion referenced measurement. *Journal of Experimental Education, 45*, 4–9.

Block, J. H. (1972). Student learning and the setting of mastery performance standards. *Educational Horizons, 50*, 183–190.

Bloom, B. S., Madaus, G. F., & Hastings, J. T. (1981). *Evaluation to improve learning.* New York: McGraw-Hill.

California Assessment Program. (1989). *A question of thinking: A first look at students' performance on open-ended questions in mathematics.* Sacramento, CA: California State Department of Education.

California Assessment Program. (1990). *Writing assessment handbook: Grade eight.* Sacramento, CA: California State Department of Education.

California Assessment Program. (1991). *A sampler of mathematics assessment.* Sacramento, CA: California State Department of Education.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Cole, N. S. (1988). A realist's appraisal of the prospects for unifying instruction and assessment. In *Assessment in the service of learning: Proceedings of the 1987 ETS invitational conference.* Princeton, NJ: Educational Testing Service.

Cronbach, L. J. (1970). *Essentials of psychological testing.* New York: Harper & Row, Publishers.

Darling-Hammond, L. (1991). The implications of testing policy for quality and equality. *Phi Delta Kappan, 73*(3), 220–225.

Darling-Hammond, L., & Wise, A. E. (1985). Beyond standardization: State standards and school improvement. *Elementary School Journal, 85*, 315–336.

Darlington, R. B., & Stauffer, G. F. (1966). A method for choosing a cutting-point on a test. *Journal of Applied Psychology, 50*, 229–231.

Ebel, R. (1972) *Essentials of educational measurement.* Englewood Cliffs, NJ: Prentice Hall.

Ebel, R. (1973). Evaluation and educational objectives. *Journal of Educational Measurement, 10*, 273–279.

Emrick, J. A. (1971). An evaluation model for mastery testing. *Journal of Educational Measurement, 8*, 321–326.

Feinberg, L. (1990). Multiple-choice and its critics: Are the alternatives any better? *Commentaries from the College Board*, 3–15.

Galton, F. (1871). *Hereditary genius: An inquiry into its laws and consequences.* New York: D. Appleton and Co.

Galton, F. (1889). *Natural inheritance.* New York: Macmillan.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*, 519–521.

Guilford, J. P. (1959). *Personality.* New York: McGraw-Hill.

Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher, 20*(5), 2–7.

Hambleton, R. K., & Novick, M. R. (1972). *Toward an integration of theory and method for criterion-referenced tests* (Research Report No. 53). Iowa City, IA: Research and Development Division, American College Testing Program.

Jaeger, R. M. (1991). Legislative perspectives on statewide testing. *Phi Delta Kappan, 73*(3), 239–242.

Koretz, D. M. (1991, April). *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.

Lanyon, R. I., & Goodstein, L. D. (1982). *Personality assessment.* New York: John Wiley and Sons.

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 test results to national norms: Interpretations of scoring "above the national average"* (CSE Technical Report No. 308, Grant OERI G-86-0003). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Madaus, G. F. (1988). The influence of testing on curriculum. In L. N. Tanner (Ed.), *Critical issues in curriculum: Eighty-seventh yearbook of the national Society for the Study of Education* (pp. 83–121). Chicago, IL: University of Chicago Press.

Madaus, G. F., West, M. M., Harmon, M. C., Lomax, R. G., & Viator, K. A. (1992). *The influence of testing on teaching math and science in grades 4–12.* Unpublished report. Chestnut Hill, MA: Boston College, National Science Foundation, Center for the Study of Testing, Evaluation, and Educational Policy.

Maryland School Performance Assessment Program. Baltimore, MD: Maryland State Department of Education.

Mehrens, W. A., & Kaminski, J. (1988). Using commercial test preparation materials for improving standardized test scores: Fruitful, fruitless, or fraudulent? *Educational Measurement: Issues and Practice, 8*(1), 14–22.

Mehrens, W. A., & Lehman, I. J. (1991). *Measurement and evaluation in education and psychology.* San Francisco: Holt, Rinehart and Winston.

Messick, S. (1990). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: American Council on Education.

Millman, J. (1972). Passing scores and test length for domain referenced measures. *Review of Educational Research, 43*, 205–216.

Moss, P.A., Beck, J.S., Matson, B., Muchmore, J., Steele, D., & Taylor, C. (1992). Portfolios, accountability, and an interpretive approach to validity. *Educational Measurement: Issues and Practice, 11*(3), 12–21.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14*, 13–19.

Paulson, F. L., Paulson, P. R., & Meyer, C. A. (1991). What makes a portfolio a portfolio? *Educational Leadership*, February, 60–63.

Pipho, C. (1992, April). *The impact of a national test at the state level.* Paper presented at the Annual Meeting of the American Educational Research Association, San

Francisco.
Popham, W. J. (1978). *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice Hall.
Putka, G. (1989, November 16). New kid in school: Alternative exams. *The Wall Street Journal*, p. B1.
Resnick, L. B., & Resnick, D. P. (1991). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction.* Boston: Kluwer.
Resnick, L. B., & Tucker, M. (1990). *Setting a new standard: Toward an examination system for the United States.* Unpublished manuscript. Pittsburgh, PA: Learning Research and Development Center; and Washington, DC: National Center on Education and the Economy.
Rowe, H. A. H. (1985). *Problem solving and intelligence.* Hillsdale, NJ: Lawrence Erlbaum and Associates, Publishers.
Schafer, W. D., & Lissitz, R. W. (1987). Measurement training for school personnel: Recommendations and reality. *Journal of Teacher Education, 38*(3), 57– 63.
Shavelson, R. (1992, April). *Evaluating the stability of hands-on science assessments.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.
Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22–27.
Shepard, L. (1991a). Will national tests improve student learning? *Phi Delta Kappan, 73*(3), 232–238.
Shepard, L. (1991b). Psychometricians' beliefs about learning influence testing. *Educational Researcher, 20*(7), 2–16.
Smith, M. L. (1991). Meanings of test preparation. *American Educational Research Journal, 28*(3), 521–542.
Spandell, V., & Stiggins, R. J. (1990). *Linking assessment and writing instruction.* New York: Longman.
Stiggins, R. J. (1988). The design and development of performance assessments. *Educational Measurement: Issues and Practice, 6*(3), 33–42.
Stiggins, R. J. (1991). Facing the challenges of a new era of educational assessment. *Applied Measurement in Education, 4*(4), 263–273.
Stiggins, R. J. (1992). High quality classroom assessment: What does it really mean? *Educational Measurement: Issues and Practice, 10*(2), 35–39.
Suen, H. K., & Davey, B. (1990, April). *Potential theoretical and practical pitfalls and cautions of the performance assessment design.* Paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA.
Trevisan, M. S. (1991, April). *Reliability of performance assessments: Let's make sure we account for the errors.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.
U.S. Department of Education. (1991). *America 2000.* Washington, DC: Author.
Valencia, S. W. (1990). A portfolio approach to classroom reading assessment: The whys, whats, and hows. *The Reading Teacher*, January, 338–340.
Valencia, S. W., & Calfee, R. (1991). The development and use of literacy portfolios for students, classes, and teachers. *Applied Measurement in Education, 4*(4), 333–345.
Wiggins, G. (1989). Teaching to the (authentic) test. *Educational Leadership*, April, 41–47.
Wiggins, G. (1990, January 24). "Standards" should mean "qualities," not quantities. *Education Week*, pp. 25, 36.
Wilson, M. (1992). Educational leverage from a political necessity: Implications of new perspectives on student assessment for Chapter 1 evaluation. *Educational Evaluation and Policy Analysis, 14*(2), 123–144.
Wineburg, S. (1993). *Collaboration in teacher assessment.* (Study #9: National Board

of Professional Teaching Standards), Philadelphia, PA: National Board of Professional Teaching Standards.

Yen, W. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187–213.

Zieky, M. J., & Livingston, S. A. (1977). *Manual for setting standards on basic skills assessment tests.* Princeton, NJ: Educational Testing Service.

Manuscript received March 18, 1993
Revision received October 5, 1993
Accepted November 19, 1993