

## CHAPTER 6

# **Standardized Tests in schools: A Primer**

## Contents

Highlights .....	165
How Do Schools Test? .....	165
Creating a Standardized Test: Concern for Consistency and Accuracy .....	171
What is a Standardized Test? .....	173
Reliability of Test Scores .....	175
Validity Evidence for Tests .....	176
How are Achievement Tests Used? .....	179
Consumers of Achievement Tests .....	181
Test Misuse .....	184
Changing Needs and Uses for Standardized Tests .....	185
What Should the Yardstick Be? .....	186
How Much is Enough? Setting Standards .....	187
What Should the Tests Look Like? .....	188
Multiple Choice: A Renewable Technology? .....	191
Redesigning Tests: Function Before Form .....	194
Conclusions .....	197

## Boxes

<i>Box</i>	<i>Page</i>
6-A. Achievement and Aptitude Tests: What is the Difference? .....	168
6-B. Types of Standardized Achievement Tests .....	170
6-C. How a Standardized Norm-Referenced Achievement Test is Developed .....	172
6-D. Large-Scale Testing Programs: Constraints on the Design of Tests .....	174
6-E. Test Score Reliability: How Accurate is the Estimate? .....	176
6-F. Helping the Student Understand Expectations: The Need for Clear Criteria .....	188
6-G. Setting and Maintaining Standards .....	190

## Figures

<i>Figure</i>	<i>Page</i>
6-1. Tests Used With Children .....	166
6-2. Thorndike's Scale for Measuring Handwriting .....	173
6-3. Testing Requirements: Three District Examples .....	183
6-4. Sample Multiple-Choice Items Designed To Measure Complex Thinking Skills ..	193
6-5. Sample Multiple-Choice Item With Alternative Answers Representing Common Student Misconceptions .....	194

## Tables

<i>Table</i>	<i>Page</i>
6-1. Three Major Functions of Educational Tests .....	180
6-2. Consumers and Uses of Standardized Test Information .....	181
6-3. Functions of Tests: What Designs Are Needed? .....	195

## Standardized Tests in Schools: A Primer

---

### Highlights

A test is an objective and standardized method for estimating behavior, based on a sample of that behavior. A standardized test is one that uses uniform procedures for administration and scoring in order to assure that results from different people are comparable. Any kind of test—from multiple choice to essays to oral examinations—can be standardized if uniform scoring and administration are used.

Achievement tests are the most widely used tests in schools. Achievement tests are designed to assess what a student knows and can do as a result of schooling. Among standardized achievement tests, multiple-choice formats predominate because they are efficient, easily administered, broad in their coverage, and can be machine scored.

Advances in test design and technology have made American standardized achievement tests remarkably sophisticated, reliable, and precise. However, misuse of tests and misconceptions about what test scores mean are common.

Tests are often used for purposes for which they have not been designed. Tests must be designed and validated for a specific function and use of a test should be limited to only those functions. Once tests are in the public domain, misuse or misinterpretation of test results is not easy to control or change.

Because test scores are estimates and can vary for reasons that have nothing to do with student achievement, the results of a single test should never be used as the sole criterion for making important decisions about individuals. A test must meet high standards of reliability and validity before it is used for any “high-stakes” decisions.

The kind of information policymakers and school authorities need to monitor school systems is very different from the kind teachers need to guide instruction. Relatively few standardized tests fulfill the classroom needs of teachers.

Existing standardized norm-referenced tests primarily test basic skills. This is because they are “generic” tests designed to be used in schools throughout the Nation, and basic skills are most common to all curricula.

Current disaffection with existing standardized achievement tests rests largely on three features of these tests: 1) most are norm-referenced and thus compare students to one another, 2) most are multiple choice, and 3) their content does not adequately represent local curricula, especially thinking and reasoning skills. This disaffection is driving efforts among educators and test developers to broaden the format of standardized tests. They seek to design tests more closely matched to local curricula, and to design tests that best serve the various functions of educational testing.

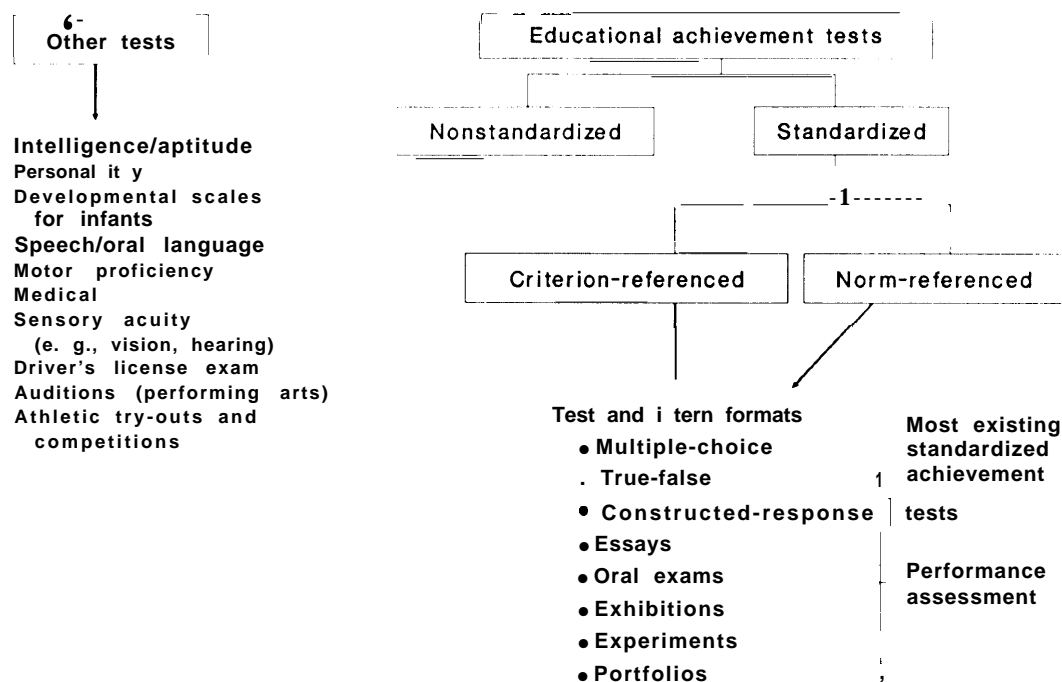
Changing the format of tests will not, by itself, ensure that tests are better measures of desired goals nor will it eliminate problems of bias, reliability, and validity. In part because of these technical and administrative concerns, test developers are exploring ways to improve multiple-choice formats to measure complex thinking skills better. As new tests are designed, new safeguards will be needed to ensure they are not misused.

### How Do Schools Test?

Nearly every type of available test designed for use with children is used in schools. Tests of personality, intelligence, aptitude, speech, sensory acuity, and perceptual motor skill, all of which have

applications in nonschool settings as well, are used by trained personnel such as guidance counselors, speech-language specialists, and school psychologists. Certain tests, however, have been designed specifically for use in educational settings. These

Figure 6-1—Tests Used With Children



SOURCE: Office of Technology Assessment, 1992; adapted from F.L. Finch, "Toward a Definition for Educational Performance Assessment," paper presented at the ERIC/PDK Symposium, August 1990.

tests, commonly referred to as achievement tests, are designed to assess student learning in school subject areas. They are also the most frequently used tests in elementary and secondary school settings; with few exceptions all students take achievement tests at multiple points in their educational careers. Educational achievement tests are the primary focus of this report.

Figure 6-1 shows the distinction between educational achievement tests and the other kinds of tests. Achievement tests are designed to assess what a student knows and can do in a specific subject area as a result of instruction or schooling. Achievement test results are designed to indicate a student's degree of success in past learning activity. Achievement tests are sometimes contrasted with aptitude tests, which are designed to predict what a person can be expected to accomplish with training (see box 6-A).

Achievement tests include a wide range of types of tests, from those designed by individual teachers

to those designed by commercial test publishing companies. Examples of the kinds of tests teachers design and use include a weekly spelling test, a final essay examination in history, or a laboratory examination in biology. At the other end of the achievement test spectrum are tests designed outside the school system itself and administered only once or twice a year; examples of this include the familiar multiple-choice, paper-and-pencil tests that might cover reading, language arts, mathematics, and social studies (see box 6-B).

The first important distinction when talking about achievement tests is between standardized and nonstandardized tests (see figure 6-1 again).<sup>1</sup> A standardized test uses uniform procedures for administering and scoring. This assures that scores obtained by different people are comparable to one another. Because of this, tests that are not standardized have limited practical usefulness outside of the classroom. Most teacher-developed tests or "back-of-the-book" tests found in textbooks would be consid-

<sup>1</sup>Fredrick L. Finch, The Riverside Publishing Co., "Toward a Definition for Educational Performance Assessment," paper presented at the ERIC/PDK Symposium, 1990.



Photo credit: Dennis Galloway

Standardized achievement tests are often administered to many students at the same sitting. Standardization means that tests are administered and scored under the same conditions for all students and ensures that results are comparable across classrooms and schools.

**Box 6-A—Achievement and Aptitude Tests: What is the Difference?**

Attempts to measure learning as a result of schooling (achievement) and attempts to measure aptitude (including intelligence) each have different, yet intertwined, histories (see ch. 4). Intelligence testing, with its strong psychometric and scientific emphasis, has influenced the design of achievement tests in this country. Achievement tests are generally distinguished from aptitude tests in the degree to which they are explicitly tied to a course of schooling. In the absence of common national educational goals, the need for achievement tests that can be taken by any student has resulted in tests more remote from specific curricula than tests developed close to the classroom. The degree of difference can be subtle and the test's title is not always a reliable guide.

A test producer's claims for an achievement test or an aptitude test do not mean that it will function as such in all circumstances with all pupils.<sup>1</sup>

There clearly is overlap between a pupil's measured ability and achievement, and perhaps the final answer to the question of whether any test assesses a pupil's achievement or a more general underlying trait such as verbal ability rests with the local user, who knows the student and the curriculum he or she has followed.<sup>2</sup>

The farther removed a test is from the specific educational curricula that has been delivered to the test taker, the more that test is likely to resemble a measure of aptitude instead of achievement for that student.

Whenever tests are going to be used for policy decisions about the effectiveness of education, it is important to assure that those tests are measuring achievement, not ability; inferences about school effectiveness must be directly tied to what the school actually delivers in the classroom—not to what children already bring to the classroom. Accordingly, tests designated for accountability should be shown to be sensitive to the effects of school-related instruction.<sup>3</sup>

To understand better the distinctions currently made between achievement and aptitude tests, it is helpful to turn to one of the “pillars of assessment development,”<sup>4</sup> Anne Anastasi:

Surpassing all other types of standardized tests in sheer number, achievement tests are designed to measure the effects of a specific program of instruction or training. It has been customary to contrast achievement tests with

<sup>1</sup>Eric Gardner, “Some Aspects of the Use and Misuse of Standardized Aptitude and Achievement Tests,” *Ability Testing: Uses, Consequences, and Controversies*, part 2, Alexandra. Wigdor and Wendell R. Garner (eds.) (Washington, DC: National Academy Press, 1982), p. 325.

<sup>2</sup>Peter W. Airasian, “Review of Iowa Tests of Basic Skills, Forms 7 and 8,” vol. I, James V. Mitchell, Jr. (ed.) (Lincoln, NE: The University of Nebraska Press, 1985), p. 720.

<sup>3</sup>No achievement test, though, will measure onZy school-related learning. For any child, learning takes place daily and as a result of all his or her cumulative experiences. “No test reveals how or why the individual reached that level.” Anne Anastasi, *Psychological York*, MacMillan Publishing Co, 1988), p. 413.

<sup>4</sup>Carol Schneider Lidz, “Historical Perspectives,” *Dynamic Potential* C.S. Lidz York, Guilford,

ered nonstandardized. Although these tests may be useful to the individual teacher, scores obtained by students on these tests would not be comparable--across classrooms, schools, or different points in time--because the administration and scoring are not standardized.

Thus, contrary to popular understanding, “standardized” does not mean norm-referenced nor does it mean multiple choice. As the tree diagram in figure 6-1 illustrates, standardized tests can take many different forms. All achievement tests intended for widespread use in decisions comparing children, schools, and districts should be standardized. Lack of standardization severely limits the inferences and

conclusions that can be made on the basis of test results. A test can be more or less standardized (there is no absolute criterion or yardstick to denote when a test has “achieved” standardization); as a result, teacher-developed tests can incorporate features of standardization that will permit inferences to be made with more confidence.

Most existing standardized tests can be divided into two primary types based on the reference point for score comparison: norm-referenced and criterion-referenced.

Norm-referenced tests help compare one student's performance with the performances of a large group of students. Norm-referenced tests are de-

aptitude tests, the latter including general intelligence tests, multiple aptitude batteries, and special aptitude tests. From one point of view, the difference between achievement and aptitude testing is a difference in the degree of uniformity of relevant antecedent experience. Thus achievement tests measure the effects of relatively standardized sets of experiences, such as a course in elementary French, trigonometry, or computer programming. In contrast, aptitude test performance reflects the cumulative influence of a multiplicity of experiences in daily living. We might say that aptitude tests measure the effects of learning under relatively uncontrolled and unknown conditions, while achievement tests measure the effects of learning that occurred under partially known and controlled conditions.

A second distinction between aptitude and achievement tests pertains to their respective uses. Aptitude tests serve to predict subsequent performance. They are employed to estimate the extent to which the individual will profit from a specified course of training, or to forecast the quality of his or her achievement in a new situation. Achievement tests, on the other hand, generally represent a terminal evaluation of the individual's status on the completion of training. The emphasis on such tests is on what the individual can do at the time.<sup>5</sup>

Although in the early days of psychological testing aptitude tests were thought to measure 'innate capacity' (unrelated to schooling, experience, or background), while achievement tests were thought to measure learning, this is now considered a misconception.<sup>6</sup> Any test score will reflect a combination of school learning, prior experience, ability, individual characteristics (e.g., motivation), and opportunities to learn outside of school. Aptitude and achievement tests differ primarily in the extent to which the test content is directly affected by school experiences.

In the 1970s, aptitude tests, particularly IQ tests, came under increasing scrutiny and criticism. A highly political debate, set off by Arthur Jensen's controversial analysis of the heritability of racial differences in intelligence, thrust IQ tests into the limelight. Similarly, the late 1960s and early 1970s saw several significant court challenges to the use of IQ tests in ability tracking. Probably because of these controversies, as well as increased understanding of the limitations of intelligence tests, many large school systems have moved away from using aptitude tests as components of their basic testing programs.<sup>7</sup> These tests are still widely marketed, however, and their use in combination with achievement tests is often promoted.

Achievement and aptitude tests differ, but the distinctions between the two in terms of design and use are often blurred. For policy purposes, the essential point is this: even though a test maybe defined as an achievement test, the more it moves away from items tied to specific curriculum content and toward items that assess broader concepts and skills, the more the test will function as an aptitude test. Should a national test be constructed in the absence of national standards or curriculum, it is therefore likely to be essentially an aptitude test. Such a test will not effectively reflect the results of schooling.

<sup>5</sup>Anastasi, *op. cit.*, footnote 3, pp. 41-414.

<sup>6</sup>*Ibid.*

<sup>7</sup>C. Dimengo, *Basic Testing Programs Used in Major School Systems Throughout the United States in the School Year 1977-78* (Akron, OH: Akron Public Schools Division of Personnel and Administration, 1978).

signed to make fine distinctions between students' performances and accurately pinpoint where a student stands in relation to a large group of students.<sup>2</sup>

These tests are designed to rank students along a continuum.

Because of the complexities involved in obtaining nationally representative norms, norm-referenced tests (NRTs) are usually developed by commercial test-publishing companies who administer the test to

large numbers of school children representative of the Nation's student population (see box 6-C). The score of each student who takes that test can be compared to the performance of other children in the standardization sample. Typically a single NRT is used by many schools and districts throughout the country.<sup>3</sup>

*Criterion-referenced tests* (CRTs) are focused on ". . . what test takers can do and what they know, not

<sup>2</sup>Lawrence Rudner, Jane Close Conoley, and Barbara S. Plake (eds.), *Understanding Achievement Tests* (Washington DC: ERIC Clearinghouse on Tests, Measurement, and Evaluation, 1989), p. 10.

<sup>3</sup>Many publishers offer district-level norms as well. Several publishers now create custom-developed norm-referenced tests that are based on local curricular objectives, yet come with national norms. These norms, however, are only valid under certain circumstances. See *ibid.*

### Box 6-B—Types of Standardized Achievement Tests

Currently available standardized achievement tests are likely to be one of four types.<sup>1</sup> The best known and most widely used type is the broad general survey achievement battery. These tests are used across the entire age range from kindergarten through adult, but are most widely used in elementary school. They provide scores in the major academic areas such as reading, language, mathematics, and sometimes science and social studies. They are usually commercially developed, norm-referenced, multiple-choice tests. Examples include the Comprehensive Test of Basic Skills, the Metropolitan Achievement Test, and the Iowa Tests of Basic Skills (ITBS). In addition, many test publishers now offer essay tests of writing that can accompany a survey achievement test.

In the 1989-90 school year, commercially published, off-the-shelf, achievement battery tests were a mandated feature of testing programs in about two-thirds of the States and the District of Columbia (see figure 6-B1). Five of those States required districts to select a commercial achievement test from a list of approved tests, while 27 specified a particular test to be administered. In addition, many districts require a norm-referenced test (NRT), even if the State does not. A survey of all districts in Pennsylvania, which does not mandate use of an NRT, found that 91 percent of the districts used a commercial off-the-shelf NRT<sup>2</sup>

The second type of test is the test of minimum competency in basic skills. These tests are usually criterion-referenced and are used for certifying attainment and/or awarding a high school diploma. They are most often used in secondary school and are usually developed by the State or district.<sup>3</sup>

Far less frequently available as commercially published, standardized tests, the third category includes achievement tests in separate content areas. The best known examples of these are the Advanced Placement examinations administered by the College Board, used to test mastery of specific subjects such as history or biology at the end of high school for the purpose of obtaining college credit.

The final type of achievement test is the diagnostic battery. These tests differ from the survey achievement battery primarily in their specificity and depth; diagnostic tests have a more narrowly defined focus and concentrate on specific content knowledge and skills. They are generally designed to describe an individual's strengths and weaknesses within a subject matter area and to suggest reasons for difficulties. Most published diagnostic tests cover either reading or mathematics. Many of the diagnostic achievement tests need to be individually administered by a trained examiner and are used in special education screening and diagnosis.

<sup>1</sup>This discussion of the four types of achievement tests is drawn from Anne Anastasi, *Psychological Testing* (New York, NY: Macmillan Publishing Co., 1988).

(New York, NY: Macmillan Publishing Co., 1988).

<sup>2</sup>Ross S. Blust and Richard L. Kohr, Pennsylvania Department of Education, "Pennsylvania School District Testing Programs," ERIC Document ED 269 449, TM 840-300, January 1984.

<sup>3</sup>See ch. 2 for a discussion of uses of minimum competency tests.

how they compare to others. CRTs usually report how a student is doing relative to specified educational goals or objectives. For example, a CRT score might describe which arithmetic operations a student can perform or the level of reading difficulty he or she can comprehend. Some of the earliest criterion-referenced scales were attempts to judge a student's mastery of school-related skills such as penmanship. Figure 6-2 illustrates one such scale, developed in 1910 by E.L. Thorndike to measure handwriting. The figure shows some of the sample

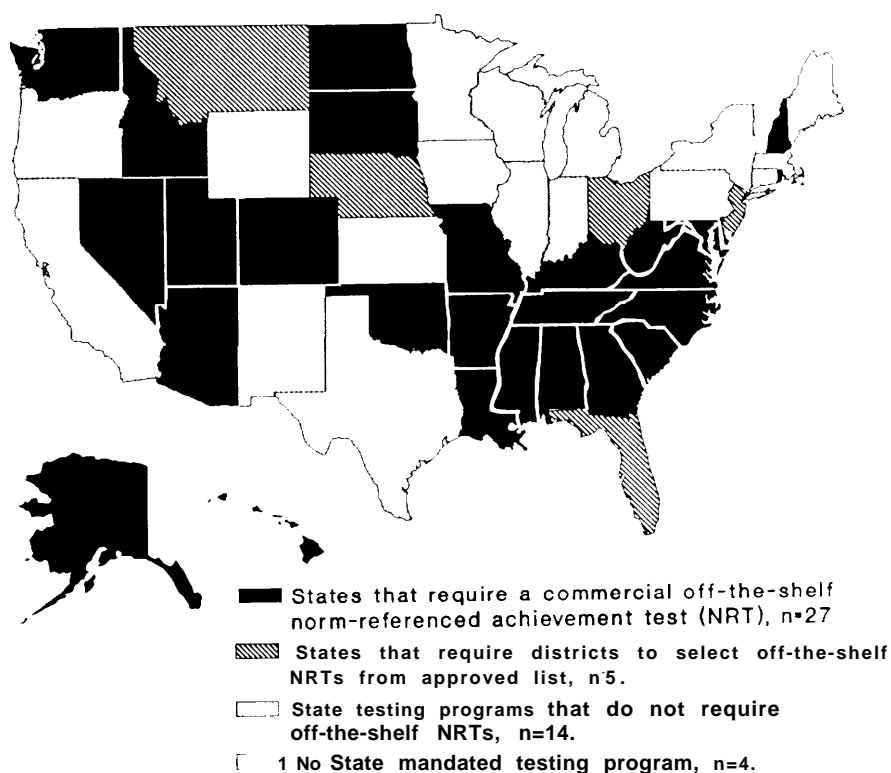
specimens against which a student's handwriting could be judged and scored.

Most certification examinations are criterion-referenced. The skills one needs to know to be certified as a pilot, for example, are clearly spelled out and criteria by which mastery is achieved are described. Aspiring pilots then know which skills to work on. Eventually a pilot will be certified to fly not because she or he can perform these skills better than most classmates, but because knowledge and mastery of all important skills have been demonstrated.

<sup>4</sup>Anne Anastasi, *Psychological Testing* (New York, NY: MacMillan Publishing CO., 1988), p. 102. The term 'criterion-referenced test' is being used here in its broadest sense and includes other terms such as content-, domain-, and objective-referenced tests.



**Figure 6-B1--State Requirements: Commercial Norm-Referenced Achievement Tests, 1990**



NOTE: Kentucky and Arizona are currently changing their norm-referenced test (NRT) requirements (see ch. 7). Although Iowa has no State testing requirements, 95 percent of its districts administer a commercial NRT.

SOURCE: Office of Technology Assessment, 1992.

Such tests will usually have designated “cutoff” scores or proficiency levels above which a student must score to pass the test.

Another component of a standardized achievement test that warrants careful scrutiny is the format of the test, the kind of items or tasks used to demonstrate student skills and knowledge. The final level in figure 6-1 depicts the range of testing formats. Almost all group-administered standardized achievement tests are now made up of multiple-choice items<sup>5</sup> (see box 6-D). Currently, educators and test developers are examining ways to use a broader range of formats in standardized achieve-

ment tests. Most of these tasks, which range from essays to portfolios to oral examinations, are labelled ‘performance assessment’ and are described in the next chapter.

### **Creating a Standardized Test: Concern for Consistency and Accuracy**

The construction of a good test is an attempt to make a set of systematic observations in an accurate and equitable manner. In the time period since Binet’s pioneering efforts in the empirical design of

<sup>5</sup>A number of commercially developed achievement tests have added optional direct sample writing tasks.

**Box 6-C—How a Standardized Norm-Referenced Achievement Test is Developed<sup>1</sup>****Step 1—Specify general purpose of the test****Step 2—Develop test specifications or blueprint**

- **Identify the content that the test** will cover: for achievement tests this means specifying both the subject matter and the behavioral objectives.
- Conduct a curriculum analysis by reviewing current texts, curricular guidelines, and research and by consulting experts in the subject areas and skills selected. Through this process a consensus definition of important content and skills is established, ensuring that the content is valid.

**Step 3—Write items**

- Often done by teams of professional item writers and subject matter experts.
- Many more items are written than will appear on the test.
- Items are reviewed for racial, ethnic, and sex bias by outside teams of professionals.

**Step 4—Pretest items**

- Preliminary versions of the items are tried out on large, representative samples of children. These samples must include children of all ages, geographic regions, ethnic groups, and so forth with whom the test will eventually be used.

**Step 5—Analyze items**

- **Statistical** information collected for each item includes measures of item difficulty, item discrimination, a g e differences in easiness, and analysis of incorrect responses.

**Step 6—Locate standardization sample and conduct testing**

- **To obtain a nationally representative** sample, publishers select students according to a number of relevant characteristics, including those for individual pupils (e.g., age and sex), school systems (e.g., public, parochial, or private) and communities (e.g., geographical regions or urban-rural-suburban).
- Most publishers administer two forms of a test at two different times of the year (fall and spring) during Standardization.

**Step 7—Analyze standardization data, produce norms, analyze reliability and validity evidence**

- Alternate forms are statistically equated **to one another**.
- **Special norms (e.g.,** for urban or rural schools) are often prepared as well.

**Step 8—Publish test and test manuals**

- Score reporting materials and guidelines are designed.

<sup>1</sup>Adapted from Anthony J. Nitko, *Educational* 1983), pp. 468-476.

(New York, N.Y.: Harcourt Brace Jovanovich,

**tests,**<sup>6</sup> considerable research effort has been expended to develop theories of measurement and statistical procedures for test construction. The science of test design, called psychometrics, has contributed important principles of test design and use. However, a test can be designed by anyone with a theory or a view to promote--witness the large number of "tests" of personality type, social IQ, attitude preference, health habits, and so forth that appear in popular magazines. Few mechanisms currently exist for monitoring the quality, accuracy, or credibility of tests. (See ch. 2 for further discussion of the issues of standards for tests, mechanisms

for monitoring test use, and protections for test takers.)

How good is a test? Does it do the things it promises? What inferences and conclusions can be drawn from the scores? Does the test really work? These are difficult questions to answer and should not be determined by impressions, judgment, or appearances. Empirical information about the performance of large numbers of students on any given test is needed to evaluate its effectiveness and merits. This section addresses the principal methods used to evaluate the technical quality of tests. It

<sup>6</sup>See ch. 4.

**Figure 6-2—Thorndike's Scale for Measuring Handwriting**

<p>Quality 18</p> <p><i>showed that the rise and fall of the tides the attraction of the moon and sun upon</i></p>
<p>Quality 17</p> <p><i>Then the carelessly dressed gentleman stepped lightly into Warren's carriage and held out a small card, John vanished be</i></p>
<p>Quality 14</p> <p><i>Then the carelessly dressed gentleman stepped lightly into Warren's carriage and held out a small card, I</i></p>
<p>Quality 9</p> <p><i>Then the carelessly dressed gentleman stepped lightly into Warren's carriage and held out a small card, John vanished behind the</i></p>
<p>Quality 5</p> <p><i>bus her and the carriage raved along down the driveway. se andie</i></p>
<p>Quality 4</p> <p><i>seated on the cub like my dancer and</i></p>

**NOTE:** A series of handwriting specimens were scaled on a numerical "quality" scale. To use the scale, a student's sample of writing is matched to the quality of one of the specimens and assigned the given numerical value. This figure shows only some of the specimens.

**SOURCE:** Anthony J. Nitko, *Educational Tests and Measurement.* An Introduction (New York NY: Harcourt Brace Jovanovich, 1983), p. 450.

begins by dissecting the basic definition of a test and then examines concepts of reliability and validity.

### What is a Standardized Test?

This type of test is an objective and standardized method for estimating behavior based on obtaining a sample of that behavior.<sup>7</sup> There are four key elements of this definition.

### Sample of Behavior

Not all of an individual's behavior relevant to a given topic can be observed. Just as a biochemist must take representative samples of the water supply to assess its overall quality, a test obtains samples of behavior in order to estimate something about an individual's overall proficiency or skill level with respect to that behavior. Thus, to estimate a student skill at arithmetic computations, a test might provide a number of problems of varying complexity drawn from each of the areas of addition, subtraction, multiplication, and division. The samples chosen must be sufficiently broad to represent the skill being tested. For example, performance on five long division problems would not provide an adequate estimate of overall computational skill. Similarly, a behind-the-wheel driving test that consists only of parking skills (parallel parking, backing into a space) would hardly constitute a valid indicator of a driver's overall competence.

### Estimation

Precisely because much of human behavior is variable and because a person's knowledge and thinking cannot be directly observed, scores obtained on any educational test should always be viewed as estimates of an individual's competence. In general, the accuracy of estimates generated by tests will be enhanced when technical procedures are used to design, field test, and modify tests during development.

### Standardization

Standardization refers to the use of a uniform procedure for administering and scoring the test. Controlling the conditions under which a test is given and scored is necessary to ensure comparability of scores across test takers. Each student is given identical instructions, materials, practice items, and amount of time to complete the test. This procedure can reduce the effects of extraneous variables on a student's score. Similarly, procedures for scoring need to be uniform for all students.

### Objectivity

Objectivity in test construction is achieved by eliminating, or reducing as much as possible, the amount of subjective judgment involved in develop-

<sup>7</sup>The word "behavior" is used here in its broadest sense and includes more specific constructs such as knowledge, skills, traits, and abilities. This discussion of the components of the definition of a test is drawn from Anastasi, op. cit., footnote 4.

### Box 6-D—Large-Scale Testing Programs: Constraints on the Design of Tests

**The** demand for standardized tests of achievement is driven by the need to collect comparable achievement data about large numbers of students, schools, and districts. Tests are required that can be given to a large number of students simultaneously and in many school districts. Because of this, and more so than for most other kinds of tests, the technology of standardized achievement testing reflects the practical considerations of economy, efficiency, and limits on the amount of time that can be devoted to test taking. The need for efficiency and economy has affected the design of standardized achievement testing in at least three important ways, each of which requires some tradeoffs in the information obtained.

**Group administration**—Most standardized achievement tests are group administered; large numbers of students take the test at the same sitting with no guidance by an examiner. Many other types of standardized tests (e.g., personality, speech, and visual-motor skills) are individually administered by trained examiners who can ensure systematic administration and scoring of results. While far more labor intensive and time consuming, individual examiners can make observations of the student that provide a rich source of supplementary information. Individually administered tests can also be tailored to the level of knowledge demonstrated by the child and thus can cover a number of content areas in some detail without becoming too long or frustrating for the child.

**Machine scored**—Most standardized achievement tests are scored by machine, because of the numbers of tests to be scored quickly and economically. This need restricts the format for student answers. Most machine-scored tests are made up of items on which students recognize or select a correct response (e.g., multiple choice or true-false) rather than create an answer of their own.

**Broad, general content**—**The** content of tests designed to be administered to all students will be broad and general when testing time is limited. The requirement that an achievement test can be taken by students of all skill levels in a given age group means that for every content area covered by the test, many items must be administered, ranging from low to high levels of difficulty. Most students will spend time answering extra items—some too difficult, some too easy—in order to accommodate all test takers.

#### Constraints

**The** design of standardized achievement tests for use with all students in a school system is therefore constrained by three factors: 1) the amount of testing time available which constrains test length, 2) the costs of test administration and scoring, and 3) the logistical constraints imposed by the large numbers of tests that must be administered and scored quickly. However, the tension between the economy and efficiency needs, and the desire for rich, individualized information, underlies much of the current testing debate.

Three major areas of technological development offer promise for expanding the range of possibilities for large-scale standardized achievement tests.

**Machine scoring**—As the technology advances, machines and computers may be able to score more complex and sophisticated responses by students (see ch. 7).

**Individual administration via computer**—**The** computer has considerable potential as a method for harnessing many of the important advantages of individualized test administration. These include the capability to adapt test items to match the proficiency of the student (allowing more detailed assessments in short time periods), and to record steps taken by the test taker. In essence, the computer may be able to replicate some of the important but expensive functions previously served by a trained testing examiner (see ch. 8).

**Sampling designs**—**The** technology of sampling, by which generalizable conclusions can be made based on testing of far fewer numbers of students, is an important development as well. The effectiveness of testing subgroups of children, or testing all children on a portion of the test, has been well demonstrated. This sampling methodology offers a practical avenue for trying some more expensive and logistically complex testing procedures, as every student in a system does not need to take the whole test.

SOURCE: Office of Technology Assessment, 1992.

ing, administering, and scoring the test. The goal of these procedures is to ensure that an individual receives a score that reflects his or her level of understanding and not the particular views or attitudes of persons administering or scoring the test. Thus, in theory an objective test is one on which the test taker will receive the same score regardless of who is involved in administering that test.<sup>8</sup>

### Reliability of Test Scores<sup>9</sup>

As used with respect to testing, reliability refers to the **consistency** of scores. If the goal is to estimate a child's level of mathematics achievement then the test should produce consistent results, no matter who gives the test or when it is given. If, at the end of 3rd grade, a student scores at the 90th percentile on Monday in mathematics achievement, but the 40th percentile when retested on Friday, neither score would instill much confidence. Scores can be inconsistent for a number of reasons: behavior varies from moment to moment, the content of a test varies, or the persons or procedures involved in scoring are variable.

The theoretical ideal for score reliability is 100 percent. In practice, though, it is impossible for an instrument that is calibrating human behavior to achieve this level of consistency. Any data from tests of human behavior contain some "noise" or error component that is irrelevant to the purpose of the test. The control of testing conditions through specification of procedures can reduce the variance in scores due to these irrelevant factors, and make the test a more reliable indicator. However, because no test is perfectly accurate and consistent, it should be accompanied by evidence of reliability. (When public opinion polls are reported, for example, they are usually accompanied by statements that indicate how much the given figures might be expected to vary, e.g., "this number might be expected to vary 4 points up or down." This statement provides information about the reliability of the poll estimate s.)

As tests are currently designed, there are three principal ways to conceptualize the reliability of test scores. Estimates of reliability can be obtained by examining the consistency of a test administered across different occasions. To what extent do scores obtained on one day agree with those obtained on a different day? This form of reliability is called stability. Secondly, consistency across content, either of different groups of items or forms of a test, can be examined. To what extent does performance on one group of subtraction items agree with performance on a second group of subtraction items intended to assess the same set of skills? This form of reliability can be assessed by alternate test forms or by indices of internal consistency. Finally, the extent to which consistent test scores will be produced by different raters can be assessed. To what extent do the scores assigned by one judge reading an essay test and using a set of designated rating criteria agree with those given by another judge using the same criteria? Indices of inter-rater reliability are used to assess such agreement.

Reliability is partly a function of test length. As a rule, the more items a test contains, the more reliable that test will be. As the number of items, or samples, incorporated in a score increases, the stability of that score will also increase. The effect of chance differences among items, as well as the impact of a single item on the total score, is reduced as a test gets longer. This is one of the reasons that multiple-choice and other short answer tests tend to be very reliable and consistent—many items can be answered in a short amount of testing time. As will be discussed in chapter 7, reliability of scores based on fewer and longer tasks is one of the important challenges faced by the developers of new performance assessments.

Reliability is particularly important when test scores are used to make significant decisions about individual students. Recall that any one test score is considered to be only an estimate of the person's "true" proficiency; this score is expected to vary somewhat from day to day. Reliability coefficients,

<sup>8</sup>While *Scoring* of certain tests can be made almost perfectly objective by use of machine-scoring technologies (see ch. 8), the *writing* of test questions, as well as the specification of what will be on the test and which is the right answer, remains a fundamentally subjective activity requiring a great deal of human judgment.

<sup>9</sup>The discussion of reliability and validity draw on Anastasi, op. cit., footnote 4; Anthony J. Nitko, *Educational Tests and Measurement: An Introduction* (New York, NY: Harcourt Brace Jovanovich, 1983); William A. Mehrens and Irvin J. Lehmann, *Measurement and Evaluation in Education and Psychology*, 3rd ed. (New York, NY: CBS College Publishing, 1984); and American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing* (Washington, DC: American Psychological Association, Inc., 1985).

**Box 6-E—Test Score Reliability: How Accurate is the Estimate?**

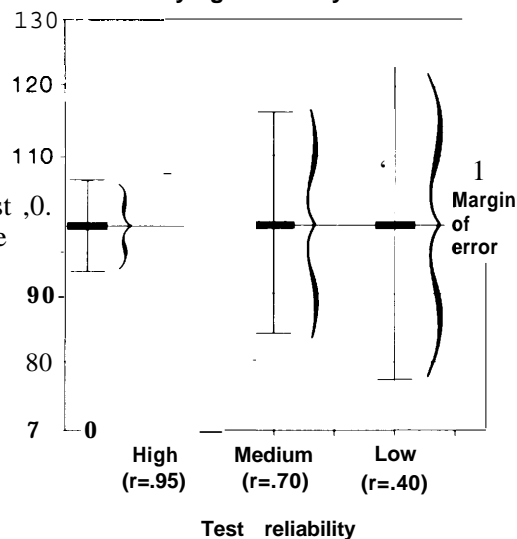
All test scores are estimates of proficiency. “Reliability” is a statistical indicator of the accuracy of those estimates: tests with higher reliability are, by definition, more accurate instruments. For example, if a test has a reliability coefficient of 0.85, this means that 85 percent of the variance in scores depends on true differences and 15 percent is attributable to other factors.

Scores therefore need to be accompanied with information about the test’s reliability. Suppose, for example, Test students took a test of arithmetic proficiency with high score reliability, e.g., 0.95. As shown in figure 6-EI, the range of error around scores on this test is relatively narrow: a score of 100 reflects a proficiency level of somewhere between 93 and 107. On a test with very low reliability, e.g., 0.40, the proficiency of a student who scores 100 may be anywhere from 77 to 123.

This information is particularly important when test scores are the basis of decisions about students. The likelihood of incorrect decisions increases when a test’s reliability is low: e.g., students could be denied remedial services based on an erroneously high score or retained in a special program because of erroneously low scores.

SOURCE: Office of Technology Assessment, 1992.

**Figure 6-EI—Error Ranges on Tests of Varying Reliability**



NOTE: Error ranges in this figure are based on the following statistical parameters: mean=100, standard deviation .15,  $p \leq 0.05$  for all tests.

SOURCE: Office of Technology Assessment, 1992.

which estimate error, allow one to set a range of likely variation or “uncertainty” around that estimated score. Box 6-E illustrates how great the variation around a score can get as the reliability of a test decreases.<sup>10</sup> Interpretation of individual scores should always take into account this variability. Small differences between the test scores of individual students are often meaningless, once error estimates are considered. When test scores are used for classification of people errors will be greatest for those whose scores are at or near the cutoff point.<sup>11</sup>

This suggests two important implications for the interpretation of individual scores in educational settings: 1) if a test score is used to make decisions about individual students, a very high standard of reliability is necessary,<sup>12</sup> and 2) using test scores alone to make decisions about individuals is likely to result in higher rates of misclassification or

incorrect decisions. With respect to educational decisions about individuals, test scores should always be used in combination with other sources of information about the child’s behavior, progress, and achievement levels.

### Validity Evidence for Tests

“It is a useful oversimplification to think of validity as truthfulness: Does the test measure what it purports to measure? . . . Validity can best be defined as the extent to which certain inferences can be made from test scores.”<sup>13</sup> Validity is judged on a wide array of evidence and is directly related to the purposes of the test.

Every test needs a clear specification of what it is supposed to be assessing. So, for example, for a test of reading proficiency, test designers first need to

<sup>10</sup>Reliability coefficients are based on the de-of relationship between two sets of scores. Correlation coefficients, generally signified with an “r,” range from 0.00 indicating a complete absence of relation to +1.00 and -1.00 indicating a perfect positive or negative relationship. The closer a reliability coefficient is to +1.00, the better.

<sup>11</sup>Nitko, op. Cit., footnote 9, p. 405.

<sup>12</sup>John Salvia and James E. Ysseldyke, *Assessment in Special and Remedial Education* (Boston, MA: Houghton Mifflin Co., 1988), p. 127.

<sup>13</sup>Mehrens and Lehmann, op. cit., footnote 9, p. 288.

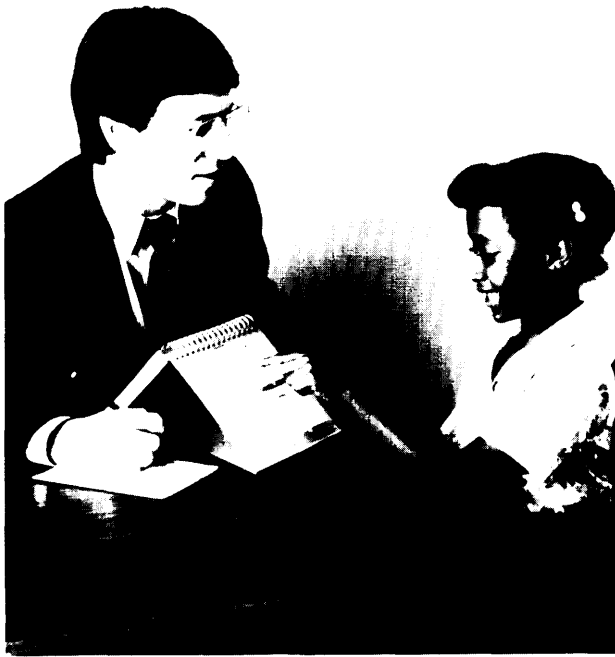


Photo credit: American Guidance Services

Some standardized tests, such as those used in special education evaluations, are individually administered by a trained examiner.

specify clearly what is meant by reading proficiency. Similarly, a test of diving skill needs to make clear what proficient dives look like. Before any testing can be done, a clear definition of the skills and competencies covered by the test must be made. There must be a definition of what the skill of interest looks like before anyone can decide how to test it. Once a method and a metric for assessing the skill has been chosen, validity evidence is gathered to support or refute the definition and the method chosen.

**EXAMPLE:** A geometry teacher, who knows nothing about diving, is drafted to take over as coach of a high school diving team when the regular coach is taken ill. While watching the varsity and the junior varsity (JV) teams practice, he tries to develop his own definition of a skilled dive; noticing that highly ranked divers enter the pool with only a slight splash while JV team members tend to make lots of waves, he designs a 1-10 rating scale to measure diving

proficiency by judging the height of the splash as the diver enters the pool. While his criterion for measuring skill may be *related to* “true diving skill,” it is not valid as the primary indicator of diving skill (as will be proven when he attempts to send his divers into statewide competition). In this case he has failed to define the trait of interest (diving skill) but rather jumped ahead to find an easy-to-measure indicator/correlate of diving skill. To carry this example farther, as the practice dives are rated on this scale, his divers begin to modify their dives in the attempt to increase their scores so that they might go to the State competition. They develop inventive ways to enter the water so that splashing is minimized. Slowly, their relative ranks (to each other) change and some JV members move up onto the varsity team. Finally, the best eight divers (judged on the 1-10 splash scale) are sent to statewide competition. Their scores are the lowest of any team and their awkward, gyrating dives send the spectators into an uproar. The most “truly” skilled divers from the team, who stayed home, never had a chance to compete.<sup>14</sup>

This example illustrates what can happen when an invalid measure is used. Often it is hard to define excellence or competence, and far easier to choose an easy-to-measure and readily available indicator of it. While many of these easy-to-measure characteristics may be correlated with excellence, *they do not represent the universe of characteristics that define competence in the skill of interest.* What can happen (as in this case) is that students practice to gain more skill in the measurable characteristic, often to the exclusion of other equally valid-but less readily measured-aspects of the skills. In this example, the coach should have first developed a definition of a skilled dive. Since statewide competition is a goal, he would do well to adopt the consensus definition and rating scale that is used by judges in the competition. This scale has developed validity over many years of use through a process of diving experts defining and describing: first, what skill in diving is and second, what level of skill one needs to get each score on a scale of 1 to 10.

The most often cited form of validity needed for achievement tests is called content validity. Establishing content validity is necessary in order to generalize from a sample to a whole domain—for example, a sample of science questions is used to

<sup>14</sup>Office of Technology Assessment, 1992.

generalize about overall science achievement. Does the content sampled by the test adequately represent the whole domain to which the test is intended to generalize? The tasks and knowledge included on a test of writing proficiency, for example, should represent the whole domain of skills and knowledge that educators believe to be important in defining writing proficiency. Since the whole domain can never be described definitively, the assessment of content validity rests largely on the judgment of experts. First the domain must be defined, then the test constructed to provide a representative sample across the domain.

There is no commonly used statistic or numerical value to express content validity. The traditional process for providing content-related validity evidence is a multifaceted one that includes review of textbooks and instructional materials, judgments of curriculum experts, and analysis of vocabulary. In addition, professionals from varying cultural and ethnic backgrounds are asked to review test content for appropriateness and fairness. The selection of test items is also influenced by studies of student errors, item characteristics, and evidence of differential performance by gender and racial-ethnic groups.

The content validity of an achievement test finally rests, however, on the match between the test content and the local curriculum.<sup>15</sup> Thus a school system selecting a test must pay careful attention to the extent to which test learning outcomes match the desired learning outcomes of the school system. "A published test may provide more valid results for one school program than for another. It all depends on how closely the set of test tasks matches the achievement to be measured."<sup>16</sup>

Another kind of validity evidence, called criterion-related, concerns the extent to which information from a test score generalizes to how well a person

will do on a different task. In this case, validity is established by examining the test's relation with another criterion of importance. For example, the Scholastic Aptitude Test (SAT), which is used to help make decisions about college admissions, is designed to predict a specific criterion, i.e., freshman grade point average (GPA). One kind of validity evidence required for any selection test is a demonstrated relation to the outcomes being predicted.<sup>17</sup>

A third kind of validity evidence, construct-related, has to do with providing evidence that the test actually measures the trait or skill it attempts to measure. Is a test of science achievement actually measuring knowledge of science and not some other skill such as reading achievement? Do scores on a mathematics achievement test really reflect the amount of mathematics a child has learned in school and not some other characteristic such as ability to work quickly under time pressure? Evidence for construct validity is gathered in multiple ways.

One common form of construct validity for achievement tests relates to whether or not performance on the test is affected by instruction. Since an achievement test is, by definition, intended to gauge the effects of a specific form of instruction, then scores should increase as a result of instruction. As the kinds of tests and tasks required of children on tests change, it will be important to conduct validity studies to make sure tests are sensitive to instruction. Care needs to be taken to assure that new tests designed to assess thinking skills or complex reasoning actually do assess the skills that can be taught in classrooms and learned by students.

Evidence that tests of specific skills such as reading comprehension, spelling, and vocabulary<sup>18</sup> are actually assessing the skills they are designed to measure is particularly important if those scores are going to be used to diagnose a child's strengths and

<sup>15</sup>Ibid.

<sup>16</sup>Norman E. Gronlund and Robert L. Linn, *Measurement and Evaluation in Teaching*, 6th ed. (New York, NY: MacMillan Publishing Co., 1990), p. 55.

<sup>17</sup>The Scholastic Aptitude Test (SAT) is not considered an achievement test, but rather a test of "developed abilities" which consist of "... broadly applicable intellectual skills and knowledge that develop over time through the individual's experiences both in and out of school." (Anastasi, op. cit., footnote 4, p. 330.) The SAT is not intended to serve as a substitute for high school grades in the prediction of college achievement; in fact, high school grades predict college grades as well, or slightly better than does the SAT. However, when test scores are combined with high school grades, prediction of college grades is enhanced slightly. This "third view" of college-bound candidates (supplementing grades and personal information from applications, interviews, and reference letters) was seen originally as a way to offset potential inequities of the traditional system; see also James Crouse and Dale Trusheim, "The Case Against the SAT," *Ability Testing: Uses, Consequences, and Controversies*, part I, Alexandra K. Wigdor and Wendell R. Garner (eds.) (Washington, DC: National Academy Press, 1982).

<sup>18</sup>The subtests that typically appear on survey achievement batteries include vocabulary, word recognition skills, reading comprehension, language mechanics (e.g., capitalization and punctuation), language usage, mathematics problem solving, mathematics computation, mathematics concepts, spelling, language, science, social studies, research skills, and reference materials.



weaknesses. Similarly, scores designed to assess “higher order thinking” need validity evidence to support the assumption that they are capturing something distinctly different from other scores assumed to include only “basic skills.” These other forms of construct validity have often been neglected by developers of standardized achievement tests.<sup>19</sup> Results of a recent survey of the technical characteristics of 37 published educational achievement tests indicate that while 73 percent of the tests presented information about content validity, only 14 percent presented criterion-related validity, and 11 percent construct validity evidence.<sup>20</sup>

Sometimes the argument is made that if a test **resembles the** construct or skill of interest, then it is valid. This is commonly referred to as face validity because the test looks like the construct it is supposed to be assessing. Because, for example, a test item seems to require complex reasoning, it is assumed to be an indicator of such reasoning. However, face validity is very impressionistic and is not considered sufficient kind of evidence for serious assessment purposes.<sup>21</sup>

The kinds of evidence discussed above constitute empirical or evidential bases for evaluating the validity of a test. Recently, however, some investigators have drawn attention to the importance of considering the consequential basis for evaluating the validity of test use. The questions posed by this form of validity are ethical and relate to the justification of the proposed use in terms of social values: “. . . should the test be used for the proposed purpose in the proposed way?”<sup>22</sup>

For example:

. . . tests used in the schools ought to encourage sound distribution of instructional and study time. . . . The worth of an instructional test lies in its contribution to the learning of students working up to the test or to next year’s quality of instruction. . . . The bottom line is that validators have an obligation to review whether a practice has appropriate consequences for individuals and institutions, and especially to guard against adverse consequences.<sup>23</sup>

## How are Achievement Tests Used?<sup>24</sup>

A precise description about how schools actually use achievement tests is difficult to obtain. Although there are many testing requirements imposed on children on their journey through elementary and secondary schools, it is difficult to say with any certainty how results are actually used, or by whom. Once a test is needed for a specific purpose such as determining eligibility for a compensatory education program, cost and time constraints often dictate that the test information is used for other purposes as well. In addition, the results of a test administration, once received by a school, are available to many people unfamiliar with the specific test administered. Test scores often remain part of a child’s permanent record and it is unclear how they might be used, and by whom, at some future point. It is difficult to prevent use of the test information for other purposes once it has been collected.

The multiple uses of achievement tests in school systems can be broadly grouped into three major categories.<sup>25</sup> (See table 6-1 for a summary of these functions.)

<sup>19</sup>James L. Waldrop, “Review of the California Achievement Tests, Forms E and F,” *The Tenth Mental Measurements Yearbook*, Jane Close Conoley and Jack J. Kramer (eds.) (Lincoln, NE: The University of Nebraska Press, 1989), p. 131.

<sup>20</sup>Bruce Hall, “Survey of the Technical Characteristics of Published Educational Achievement Tests,” *Educational Measurement: Issues and Practice*, spring 1985, pp. 6-14.

<sup>21</sup>Mehrens and Lehmann, op. cit., footnote 9; Roger Farr and Beverly Farr, *Integrated Assessment System: Language Arts performance Assessment, Reading/Writing*, technical report (San Antonio, TX: The Psychological Corp., 1991); Anastasi, op. cit., footnote 4.

<sup>22</sup>Samuel Messick, “Test Validity and the Ethics of Assessment” *American Psychologist*, vol. 35, No. 11, 1980, pp. 1012-1027. See also Samuel Messick, “Validity,” *Educational Measurement*, 3rd ed., Robert Linn (ed.) (New York, NY: MacMillan Publishing Co., 1989).

<sup>23</sup>Lee J. Cronbach, “Five Perspectives on the Validity Argument,” *Test Validity*, Howard Wainer and Henry I. Braun (eds.) (Hillsdale, NJ: Lawrence Erlbaum, 1988), pp. 5-6.

<sup>24</sup>This discussion of purposes draws on Jason Millman and Jennifer Greene, “The Specification and Development of Tests of Achievement and Ability” in Linn (ed.), op. cit., footnote 22, pp. 335-367; C.V. Bunderson, J.B. Olsen, and A. Greenberg, “Computers in Educational Assessment,” OTA contractor report, Dec. 21, 1990; J.A. Frechtling, “Administrative Uses of School Testing Programs,” in Linn (ed.), op. cit., footnote 22, pp. 475-485; and R. Darrell Bock and Robert J. Mislevy, “Comprehensive Educational Assessment for the States: The Duplex Design,” *CRESST Evaluation Comment*, November 1987.

<sup>25</sup>Although many authors have discussed these three major categories, these distinctions are drawn most directly from Lauren B. Resnick and Daniel P. Resnick, “Assessing the Thinking Curriculum: New Tools for Educational Reform,” *Future Assessments: Changing Views of Aptitude, Achievement, and Instruction*, B.R. Gifford and M.C. O’Connor (eds.) (Boston, MA: Kluwer Academic Publishers, 1989).

Table 6-I—Three Major Functions of Educational Tests

Functions	Examples
<p>1. Classroom instructional guidance</p> <p>Used to monitor and provide feedback about the progress of each student and to inform teaching decisions about <i>individuals</i> on a day-to-day basis</p>	<ul style="list-style-type: none"> <li>• Diagnose each student's strengths and weaknesses</li> <li>• Monitor the effects of a lesson or unit of study</li> <li>• Monitor mastery and understanding of new material</li> <li>• Motivate and organize students' study time</li> <li>• Adapt curriculum to progress as indicated by tests</li> <li>• Monitor progress toward curricular goals</li> <li>• Plan lessons that build on students' level of current understanding</li> <li>• Assign students to learning groups (e.g., reading group)</li> </ul>
<p>2. System monitoring</p> <p>Used for monitoring and making administrative decisions about aggregated <i>groups</i> of students (e.g., a school, instructional programs, curricula, district)</p>	<ul style="list-style-type: none"> <li>• Report to parents and school board about a school or district's performance</li> <li>• Make decisions about instructional programs and curriculum changes</li> <li>• Evaluate Chapter 1 programs</li> <li>• Evaluate experimental or innovative programs</li> <li>• Allocate funds</li> <li>• Evaluate teacher performance/school effectiveness</li> <li>• Provide general information about performance of the overall educational system</li> </ul>
<p>Selection, placement, and certification of students ("gatekeeping")</p> <p>Used to allocate educational resources and opportunities among individuals</p>	<p>Selection:</p> <ul style="list-style-type: none"> <li>• Admission to college or private schools</li> </ul> <p>Placement:</p> <ul style="list-style-type: none"> <li>• Place students in remedial programs (e.g., Chapter 1)</li> <li>• Place students in gifted and talented programs</li> </ul> <p>Certification:</p> <ul style="list-style-type: none"> <li>• Certify minimum competency for receipt of high school diploma</li> <li>• Certify mastery of a course of study (e.g., Advanced Placement examinations)</li> <li>• Make decisions about grade promotion</li> </ul>

SOURCE: Office of Technology Assessment, 1992.

The first broad category encompasses the kind of tests that can support and guide the learning process of each individual student in the classroom. These tests can be used to monitor and provide feedback about the educational progress of each student in the classroom, to diagnose areas of strength and weakness, and to inform teacher decisions about how and what to teach based on how well students are learning the material.

The second major function—system monitoring—encompasses the many managerial uses of tests to monitor the educational system and report to the public. In these uses, what is needed is aggregated information about the achievement of groups of students—from classrooms to schools, from districts

to States. School administrators use this data to make decisions among competing curricula or instructional programs and to report to the public about student achievement. In addition, test scores are increasingly being used as accountability tools to judge the quality of the educational system and those who work for it. Tests used as accountability tools are often intended to allow a public evaluation of whether or not standards are being met.<sup>26</sup>

The third broad category of uses is also managerial, called here selection, placement, and certification. Included in this broad category are tests used to make institutional decisions affecting the progress of individual students through the educational system. Comparable information is needed for each

<sup>26</sup>Frechtling, *op. cit.*, footnote 24.

Table 6-2-Consumers and Uses of Standardized Test Information

Consumer	Unit of analysis
<b>National level</b>	
Allocation of resources to programs and priorities .....	Nation, State
Federal program evaluation (e.g., Chapter 1) .....	State, program
<b>State legislature/State department of education</b>	
Evaluate State's status and progress relevant to standards .....	State
State program evaluation .....	State, program
Allocation of resources .....	District, school
<b>Public (lay persons, press, school board members, parents)</b>	
Evaluate State's status and progress relevant to standards .....	District
Diagnose achievement deficits .....	Individual school
Develop expectations for future success in school .....	Individual
<b>School districts--central administrators</b>	
Evaluate districts .....	District
Evaluate schools .....	Schools
Evaluate teachers .....	Classroom
Evaluate curriculum .....	District
Evaluate instructional programs .....	Program
Determine areas for revision of curriculum and instruction .....	District
<b>School districts--building administrators</b>	
Evaluate school.. .....	School
Evaluate teacher. ....	Classroom
Group students for instruction .....	Individual
Place students into special programs .....	Individual
<b>School districts-teachers</b>	
Group students for instruction .....	Individual
Evaluate and plan curriculum .....	Classroom
Evaluate and plan instruction .....	Classroom
Evaluate teaching .....	Classroom
Diagnose achievement deficits .....	Classroom, individual
Promotion and graduation .....	Individual
Place into special programs (e.g., gifted, handicapped) .....	Individual
<b>Educational laboratories, centers, universities</b>	
Policy analysis. ....	All units
Evaluation studies. ....	All units
Other applied research. ....	All units
Basic research .....	All units

SOURCE: Thomas M. Haladyna, Susan Bobbit Nolen, and Naney S. Haas, "Railings Standardized Achievement Test Scores and the Origins of Test Score Pollution," *Educational Researcher*, vol. 20, No. 5, June-July 1991, p.3.

individual student so that managerial decisions can be made about the allocation of additional resources, placement in instructional programs, and certification of mastery. Increasingly test scores have been used to make such decisions because they are perceived to provide clear, objective criteria. Thus, eligibility for a compensatory education program (e.g., Chapter 1) might be determined by a district policy that states a cutoff score below which children must score to qualify. Qualifying for an enrichment program might be contingent on scoring above some designated level on a standardized test.

The results of these tests clearly have significant implications for a student's progress through the school system.<sup>27</sup>

### Consumers of Achievement Tests

In addition to the many *uses* for achievement test-based information, there are many different consumers or users who need that information. The kind of information needed is often very different depending on who wants it. Table 6-2 summarizes the major consumers of test-based information as

<sup>27</sup>Ironically, while most of the supplementary resources allocated by schools are likely to be targeted to children scoring either quite low or quite high on these tests, the norm-referenced achievement tests routinely used by most school districts are designed to measure most accurately in the middle of the achievement distribution rather than at either the highest or the lowest ends.

well as the most common uses of each consumer.<sup>28</sup> Within the educational system there are multiple levels of need for test-based information including Federal, State, district, school, and classroom information. Policy makers and legislators need the information, as well as education departments. Teachers, parents, students, and the public also require test-based information about achievement.

Mandatory schoolwide testing programs, in which each child in a given grade takes the same test, have become routine. Some tests are required at the Federal level, e.g., for Chapter 1 accountability,<sup>29</sup> some mandated by the States, and others implemented by local school districts. Because most school districts want to keep testing requirements to a minimum, a test is often chosen that can serve as many uses and consumers as possible.

Figure 6-3 illustrates the mandated schoolwide tests given in grades 1 through 12 for three large school districts. State-mandated testing requirements, which have increased in overall numbers in recent years, account for only a fraction of the total testing burden. Additional tests (not listed in the table) are also administered to some subgroups of children who need to be screened for special services. For example, although some districts may use schoolwide tests to satisfy Federal-level Chapter 1 accountability requirements (Philadelphia uses the City Wide Test for this purpose), many children who receive Chapter 1 services will take tests in addition to those listed in the table.

Although the specifics of who actually uses test results and for what purposes remain difficult to document, evidence suggests that requirements regarding standardized achievement tests are imposed largely to serve the two broad managerial purposes—system monitoring; and selection, placement, and certification. There are few standardized tests designed explicitly to help teachers assess ongoing classroom learning and inform classroom practice. Furthermore, evidence also suggests that teachers find the results of existing standardized achievement tests only generally useful for classroom practice. In



Photo credit: Arana Sonnier

Teachers need tests that are closely matched to instruction and that provide detailed information about student progress on a frequent basis. This kind of information, which can help teachers influence learning and guide instruction, is very different from the kind of information school administrators need to monitor school systems.

one study that interviewed teachers, 61 percent reported that standardized tests have little effect on their instructional decisionmaking.<sup>30</sup>

Current achievement tests do a good job of assessing a student general level of knowledge in a particular content domain. . . . A low score relative to a student grade placement on, say, a reading comprehension test is apt to be a valid indicator that a student will have difficulty reading and understanding assignments in the typical textbooks used at the grade level. Such global information, however, is more likely to confirm what the teachers already know about the student than to provide them with new insights or clear indications of how best to help the student. The global score simply does not reveal anything about the causes of the problem or provide any direct indications of what instructional strategies would be most effective.<sup>31</sup>

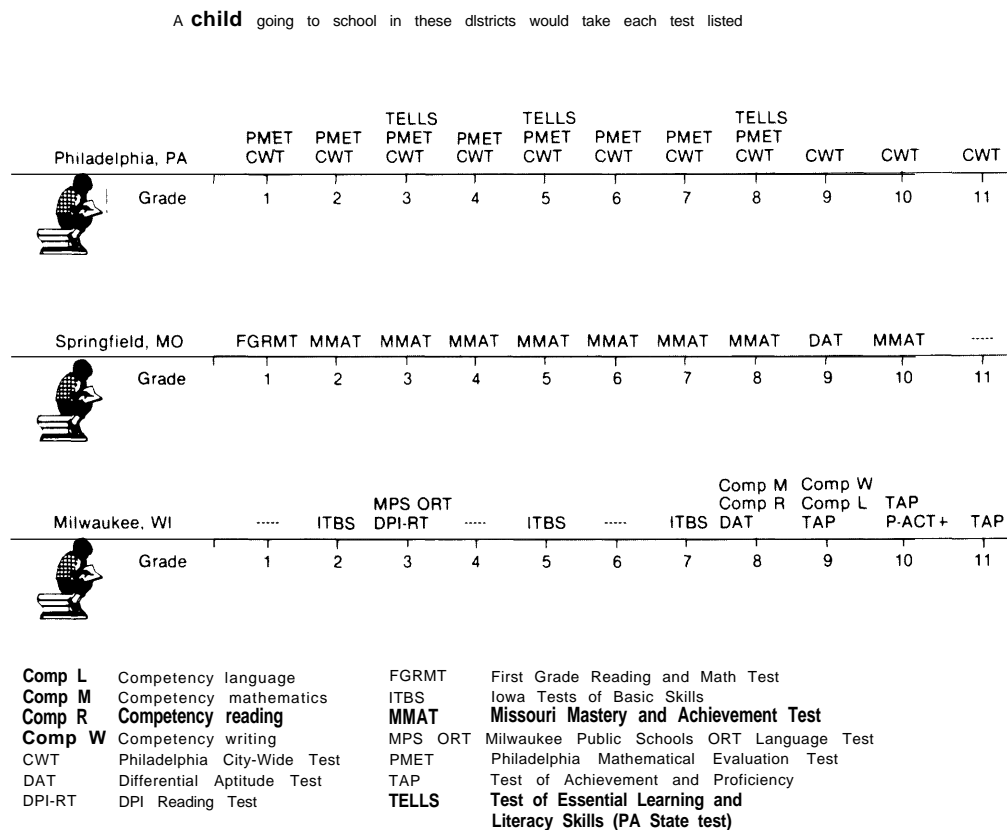
<sup>28</sup>See also Bock and Mislevy, *op. cit.*, footnote 24, for a similar list and analysis of test consumers.

<sup>29</sup>Chapter 1 is a Federal compensatory education program serving low-achieving students from low-income schools. See ch. 3 for a fuller discussion of the testing and evaluation requirements under Chapter 1.

<sup>30</sup>Robert B. Ruddell, "Knowledge and Attitudes Toward Testing: Field Educators and Legislators," *The Reading Teacher*, vol. 389, 1985, pp. 538-543.

<sup>31</sup>Robert L. Linn, "Barriers to New Test Designs," *The Redesign of Testing for the 21st Century* (Princeton, NJ: Educational Testing Service, Oct. 26, 1985), p. 72.

Figure 6-3-Testing Requirements: Three District Examples



NOTE: If students have special needs or are in supplementary programs (e.g., Chapter 1 or gifted programs) they will usually take additional tests.

SOURCES: Milwaukee Public Schools, "Summary Report and Recommendations of the Assessment Task Force," unpublished report, June 2, 1989; Springfield Public Schools, 1990; Nancy Kober, "The Federal Framework for Evaluation and Assessment in Chapter 1, ESEA," OTA contractor report, May 1991.

Teachers desire diagnostic tests that are precise, closely matched to curricula and instruction and timely. Achievement tests of the kind now widely used do not match these criteria.<sup>32</sup>

Part of the reason that few existing standardized tests are applicable for classroom use, however, has to do with local control of curriculum. Achievement tests are designed to match the goals and objectives of the content being taught; the validity of an achievement test rests largely on the degree to which it mirrors the content being taught in the classroom. A test that contains a great deal of content not covered by the curriculum in a particular school is said to be “content invalid” for that school. Teachers, because they know what they are teaching, can design tests that are well aligned with the curriculum. If an examination is designed at a great distance from the local classroom (as commercially produced and published tests are bound to be) it is less likely to reflect the specific curricular content of the classroom; these tests will largely reflect only those broad content areas and skills that are common across school settings and on which there is implicit consensus.<sup>33</sup> Thus, tests that are precise and closely matched to curricula, and therefore useful to teachers, will need to be designed at the local level, close to where specific curricular goals and objectives are set. “Generic” *standardized achievement tests as currently designed cannot be both specific enough to assist teachers on an ongoing basis and generic enough to be useful to large numbers of school systems.*

Most mandated, standardized testing is put in place for managerial purposes and not for purposes related to shaping directly day-to-day learning processes in classrooms. Since such tests are generally given once a year, they can offer teachers a “snapshot” of a child’s achievement at one particu-

lar point in time, but offer little information about the ongoing, ever-changing *process* of a child’s learning and development.<sup>34</sup>

The social success of testing in many ways is a product of the bureaucratization of education. Testing seems not so important in the stuff of teaching and learning, where surely there must be much personal contact, but rather in the interstices of our educational institutions—entry into elementary school, placement in special classes, the transition from elementary to secondary school, high school leaving and college going.<sup>35</sup>

### Test Misuse

It is difficult to make general statements about the misuses of tests, because each test has to be evaluated with respect to its own specifications and technical evidence regarding the validity of its use for specific purposes.<sup>36</sup> Many different tests are used by school systems, some commercially designed, some designed by districts or States. However, results of one survey of mathematics teachers shed some light on the uses of well-known commercial achievement tests. In this survey, three commercial tests were found to account for 44 percent of district testing requirements. In districts where these three tests were used about two-thirds of the teachers reported their use by the district to group students by ability and to assign students to special programs. However, technical reviews of these three tests have suggested that evidence is lacking regarding inferences about student diagnosis and placement for these tests.<sup>37</sup> One reviewer cautioned about one of these tests that: “. . . although useful as an indicator of general performance, the usefulness of the test for diagnosis, placement, remediation or instructional planning has not been validated.”<sup>38</sup>

<sup>32</sup>Leslie Salmon-Cox, “Teachers and Standardized Achievement Tests: What’s Really Happening?” *Phi Delta Kappan*, vol. 62, No. 9, 1981, p. 634.

<sup>33</sup>See, e.g., Roger Farr and Robert F. Carey, *Reading: What Can be Measured?* 2nd ed. (Newark DE: International Reading Association, Inc., 1986), p. 149.

<sup>34</sup>The majority of districts test at the end of the school year and the results are often received too late to be of help to that year’s classroom teacher. Some districts test more than once a year.

<sup>35</sup>Walter Haney, “Testing Reasoning and Reasoning About Testing,” *Review of Educational Research*, vol. 54, No. 4, 1984, p. 641.

<sup>36</sup>See also Robert L. Linn, Center for Research on Evaluation, Standards and Student Testing, University of Colorado at Boulder, “Test Misuse: Why Is It So Prevalent?” OTA contractor report, September 1991; Larry Cuban, Stanford University, “The Misuse of Tests in Education,” OTA contractor report, Sept. 9, 1991; and Nelson Noggle, “The Misuse of Educational Achievement Tests for Grades K-12: A Perspective,” OTA contractor report, October 1991.

<sup>37</sup>T. Romberg, E.A. Zarinnia and S.R. Williams, *The Influence of Mandated Testing on Mathematics Instruction: Grade 8 Teachers’ Perception* (Madison WI: National Center for Research in Mathematical Sciences Education, March 1989).

<sup>38</sup>Peter W. Airasian, “Review of the California Achievement Tests, Forms E and F,” Jane Close Conoley and Jack J. Kramer (eds.), *The Tenth Mental Measurements Yearbook* (Lincoln, NE: The University of Nebraska Press, 1989), pp. 719-720.

Although most standardized achievement tests are not designed to be used as selection or placement instruments on which to base judgments about future proficiency or capability, there are few mechanisms to prevent such uses. Tests that are going to be used for selection should be designed and validated for that purpose. Tests designed to be used as feedback mechanisms to inform the learning process should not be used to make significant decisions about an individual's educational career unless additional evidence can be provided substantiating this use. However, there are few safeguards available to make sure this does not happen.

One of the most consistent recommendations of testing experts is that a test score should never be used as the single criterion on which to make decisions about individuals. Significant legal challenges to the over-reliance on IQ test scores in special education placements led to an exemplary federally mandated policy on test use in special education decisions. In Public Law 94-142, Congress included several provisions designed to protect students and ensure fair, equitable, and non-discriminatory assessment procedures. Among these were:

- decisions about students are to be based on more than performance on a single test,
- tests must be validated for the purpose for which they are used,
- children must be assessed in all areas related to a specific or suspected disability, and
- evaluations should be made by a multidisciplinary team.<sup>39</sup>

This legislation provides, then, a number of significant safeguards against the simplistic or capricious use of test scores in making educational decisions. Similar safeguards are needed to prevent over-reliance on single test scores to make educational decisions about all students, not just those in special education programs.<sup>40</sup>

Other examples of test misuse arise when results of available tests are used in the aggregate to make unsupportable inferences about educational effectiveness. The use of college admissions tests (SAT and the American College Testing program-ACT)

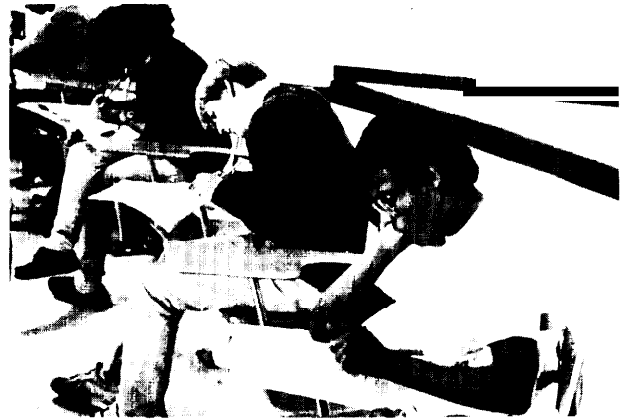


Photo credit: Educational Testing Service

**Some standardized tests are used to make significant decisions about the progress of individual students through the educational system. These tests must meet very high technical standards and are most subject to scrutiny and legal challenge.**

to compare the quality of education in various States, as in the "Wall Charts" produced by the U.S. Department of Education, is one prominent example. The SAT is taken by different numbers and samples of students (none of them randomly selected) in each State. Further, inferences about the achievement levels of high school seniors should be made only from a test designed to sample what high school seniors have been taught. The SAT is not designed for this purpose—it is designed to predict success (grade point average) in the freshman year of college. College admissions tests are designed for a distinctly different purpose than informing policymakers interested in educational quality.<sup>41</sup> In some respects it is similar to using a test of reading achievement to draw conclusions about mathematics achievement; although the two are likely to show some relation to one another, it would be erroneous to draw conclusions and make decisions about mathematics based on test scores in reading.

## Changing Needs and Uses for Standardized Tests

Current disaffection with the widely used existing standardized tests rests largely on three features of those tests: 1) most are norm-referenced and thus

<sup>39</sup>Salvia and Ysseldyke, op. cit., footnote 12.

<sup>40</sup>See ch. 2 for further discussion of test misuse and mechanisms for enforcing appropriate testing practices.

<sup>41</sup>See Robert L. Linn, "Accountability: The Comparison of Educational Systems and the Quality of Test Results," *Educational Policy*, vol. 1, No. 2, June 1987, pp. 181-198, for further discussion of the problems involved in using test scores to compare educational quality across States.

scores are based on comparing students to one another; 2) most are exclusively made up of multiple-choice items; and 3) their content does not adequately represent local curricula, especially those parts associated with thinking and reasoning skills. Most of the new developments in test design and alternative forms of assessment reflect a move away from this one dominant testing technology. What features do innovators seek in other designs?

### What Should the Yardstick Be?

Traditional test theory and techniques of test construction have been developed on the assumption that the purpose of a test is to discriminate among individuals. If the purpose of a test is to compare each individual to a standard, then it is irrelevant whether or not the individuals differ from each other.<sup>42</sup>

Recent attempts to develop alternative tests represent a move away from the traditional testing model built on comparing individuals to one another. Instead, new testing developments represent attempts to extend the criterion-referenced model of testing and design ways to assess students against criteria and goals for achievement.

There are two main reasons that existing norm-referenced tests tend to provide broad coverage of a limited number of content areas. First, these tests are designed to be taken by students of all skill levels in a given grade; this means that for every content area covered by the test, many items must be administered, ranging from low to high levels of difficulty. Most students will spend scarce testing time answering extra items—some too difficult, some too easy—included in order to accommodate all test takers. This means that fewer content areas can be covered in a limited amount of testing time. Second, NRTs must concentrate on those content areas that are common to most schools throughout the country. In essence, the content areas represented on NRTs represent broad and generally implicit national consensus about the core skills that children should know at each grade level. If these tests are primarily tests of basic skills, as many have argued, it maybe because it is these skills that are common to the

majority of curriculum frameworks throughout the country. Because of the way NRTs are developed, the content areas included can only represent a subset of the content areas covered in any particular school. Arizona, for example, found that only 26 percent of their curriculum goals were covered in the NRT they had been using. Thus, existing NRTs will only assess a limited set of content areas and only in a very general way. However, they can provide a basis for comparing children across the Nation on that common general content.

Comparing children across the Nation on what they have been taught, without setting any standards or goals as to what they should have been taught, entails testing only those skills for which there is an implicit national consensus—which is also likely to be the “least common denominator” of academic content. Local control over curricula means that each district can decide what skills and knowledge fourth graders should have, for example. To compare them fairly, one can only use a test that represents content all children have been taught. However, if one is willing to arrive at some kind of consensus about what children should know at various age levels, then tests can be designed to represent those areas.<sup>43</sup>

Criterion-referenced tests (CRTs) can provide specific information that is directly tied to the curricula being delivered in the classroom. Most tests need to be developed locally to achieve this level of specificity. Many States have, in recent years, implemented a CRT statewide program in order to assess progress on State-mandated goals and skills. However, many people, from policymakers to parents, also want a method for referencing how students are doing with respect to the education of the whole Nation. Parents and policymakers want assurance that children are not just getting the set of skills and knowledge that would make them successful in Wyoming, for example, but rather that the received education is preparing children for the national workplace and postsecondary educational institutions. Because States and districts continually need to evaluate their own goals and curriculum, data comparing their students to students across the

---

<sup>42</sup>Mehrens and Lehmann, *op. cit.*, footnote 9, p. 210

<sup>43</sup>Another important aspect of the design of norm-referenced tests has to do with the way items are finally selected to appear on the test. ‘One of the most important criteria for deciding whether to retain a test item is how well that item contributes to the variability of test scores.’ Rudner et al. (eds.), *op. cit.*, footnote 2, p. 12. In this model, items that are too easy or too difficult maybe eliminated from the test even if those items are related to important learning goals. For example, information that has been mastered by all children of a given age may not appear on the test because this information does not describe the differences in what they know.



Nation can provide an important perspective on the relative success of their educational efforts. At the present time, nationally norm-referenced standardized achievement tests are the only mechanism available for achieving this type of “national calibration.”<sup>44</sup> Thus many States and districts will adopt an overall testing program that uses both an NRT and a CRT. One testing program (CRT) can describe how the State is doing with respect to its own curricular goals, the other (NRT) program can describe how children in the State are achieving relative to all children in the country.<sup>45</sup>

### How Much is Enough? Setting Standards

It can be difficult to evaluate what either a CRT or NRT score *means* without reference to some standard or decision about how much is enough. If a child has mastered 70 percent of a given skill, how is she doing? This score means something different to her teacher if most other children in her class know 100 percent than if most know 50 percent. Or if the school district expects 100 percent mastery of this skill in first grade or fifth grade. Often, therefore, cutoff scores are set to establish mastery levels.

In discussions of testing, this represents the more technical meaning of the word “standard.”<sup>46</sup> In this case:

... a standard is an answer to the question “How much is enough?” There are standards for many kinds of things, including the purity of food products, the effectiveness of fire extinguishers and the cleanliness of auto exhaust fumes. When you choose a passing score, you are setting a standard for performance on a test.<sup>47</sup>

The most familiar testing example comes from minimum competency testing; a passing score is set, based on some criteria for competency, above which students are certified and below which they are not.

The answer to “how much is enough?” is almost always “it depends.” How safe is safe enough and how clean is clean enough are issues that have occupied consumer safety and environmental protection advocates and policymakers for years. Choosing a passing score on a test is rarely clear-cut. Any standard is based on some type of judgment. In testing, the choice of a passing score or scores indicating levels of proficiency will be largely reliant on judgments. In testing, “. . . it is important that these judgments be:

1. made by persons who are qualified to make them;
2. meaningful to the persons who are making them; and
3. made in a way that takes into account the purpose of the test.”<sup>48</sup>

Because of the error inherent in any individual test score, however, it is virtually impossible to choose a passing score that will eliminate mistakes or wrong decisions. Some test takers will pass when they should have failed and some will fail when they should have passed. When setting passing scores or standards it is important to consider the relative likelihood, importance, and social value of making both of these kinds of wrong decisions.<sup>49</sup>

A second, more general use of the term standard is also being employed in many of the current discussions about testing.

As the history of the word reminds us, a “standard” is a set of values around which we rally; we “defend” standards. (The “standard” was the flag held aloft in battle, used to identify and orient the troops of a particular king.) . . . Standards represent . . . desirable behaviors, not the best typical behavior.<sup>50</sup>

This meaning of standard draws more from the dictionary definition of a standard as “. . . something established by authority, custom, or general

<sup>44</sup>See Linn, *op. cit.*, footnote 41, pp. 181.19s, for further discussion of various options by which State and national comparisons might be made.

<sup>45</sup>See also the profiles of Arizona and Kentucky State testing programs in ch. 7.

<sup>46</sup>Webster’s defines this meaning as “. . . something set up and established by authority as a rule for the measure of quantity, weight, extent, value or quality.” *Webster’s Ninth New Collegiate Dictionary* (Springfield, MA: Merriam Webster, 1988), p. 1148.

<sup>47</sup>Samuel A. Livingston and Michael J. Zieky, *Passing Standards: A Manual for Setting Standards of Performance on Educational and Occupational Tests* (Princeton, NJ: Educational Testing Service, 1982), p. 10.

@Ibid., p. 12.

<sup>49</sup>For analysis and discussion of technical problems in the setting of cutoff scores see, e.g., Robert Guion, “Personnel Assessment, Selection, and Placement,” *Handbook of Industrial and Organizational Psychology*, vol. 2, M. Dunnette and L. Hough (eds.) (Palo Alto, CA: Consulting Psychologists Press, 1991), pp. 327-397.

<sup>50</sup>Grant Wiggins, “‘Standards’ Should Mean ‘Qualities,’ Not Quantities,” *Education Week*, vol. 9, No. 18, Jan. 24, 1990, p. 36.

consent as a model or example.”<sup>51</sup> A standard, in this sense, is an exemplar—. . . whether few, many, or all students can meet or choose to meet it is an independent issue. . . .<sup>52</sup>

An example of this kind of standard that is now widely cited is the *Curriculum and Evaluation Standards for School Mathematics* prepared by the National Council of Teachers of Mathematics (NCTM). This document contains a series of standards intended to be criteria against which schools can judge their own curricular and evaluation efforts. For example, the first standard reads as follows:

**Standard 1: Mathematics as Problem Solving**

In grades K-4, the study of mathematics should emphasize problem solving so that students can—

- \* use problem-solving approaches to investigate and understand mathematical content;
- \* formulate problems from everyday and mathematical situations;
- \* develop and apply strategies to solve a wide variety of problems;
- \* verify and interpret results with respect to the original problem;
- \* acquire confidence in using mathematics meaningfully.<sup>53</sup>

The specifics about how to test or assess this standard or about “how much is enough?” are not specified in the NCTM document. Instead it provides a common framework and a set of exemplars toward which educators and students can work—such standards describe what optimal performance looks like and what is desirable for students to know. Without clear standards for performance, many students are left struggling to understand the criteria on which they are being evaluated. Box 6-F, excerpted from a contemporary play, highlights one aspiring athlete’s struggle to ascertain the criteria or standards by which his performance as an athlete is being judged. Box 6-G describes some of the issues involved in setting and maintaining standards.

### What Should the Tests Look Like?

Currently almost all group-administered standardized achievement tests are made up of multiple-choice items; increasing dissatisfaction with multiple-choice technology as the single method for assessing

#### Box 6-F—Helping the Student Understand Expectations: The Need for Clear Criteria

The need for explicit standards and criteria in learning is aptly described in this letter excerpted from the play *Love Letters*. The letter is written by a teen-age boy about his performance in crew.

I’m stroking the 4th crew now. Yesterday, I rowed number 2 on the 3rd. Tomorrow I may row number 6 on the 2nd or number 4 on the 4th. Who knows? You get out there and work your butt off, and the launch comes alongside and looks you over, and the next day they post a list on the bulletin board saying who will row what. They never tell you what you did right or wrong, whether you’re shooting your slide or bending your back or what. They just post the latest results for all to see. Some days I think I’m doing really well, and I get sent down two crews. One day I was obviously hacking around, and they moved me UP. There’s no rhyme or reason. I went to Mr. Clark who is the head of rowing and I said, “Look, Mr. Clark. There’s something wrong about this system. People are constantly moving up and down and no one knows why. It doesn’t seem to have anything to do with whether you’re good or bad, strong or weak, coordinated or uncoordinated. It all seems random.” And Mr. Clark said “That’s life, Andy.” And walked away. Well maybe that’s life, but it doesn’t have to be life. You could easily make rules which made sense, so the good ones moved up and the bad ones moved down, and people knew what was going on. I’m serious.<sup>1</sup>

<sup>1</sup>From *Love Letters*, a play by A.R. Gurney.

achievement has led to considerable current experimentation with other item types and testing formats. Although the pros and cons of multiple-choice items are being widely and hotly debated, this testing format has many valuable characteristics.

The multiple-choice item has achieved virtual dominance of the large-scale testing market primarily because of its psychometric and administrative properties. Although expensive and difficult to develop, multiple-choice items are efficient to administer and score, particularly when items and answers are kept secure. Large numbers of students can be tested simultaneously and their tests scored and returned within a relatively short period of

<sup>51</sup>Webster’s Ninth New Collegiate Dictionary, Op. Cit., footnote 46.

<sup>52</sup>Wiggins, op. cit., footnote 50, p. 25.

<sup>53</sup>National Council of Teachers of Mathematics, *Curriculum and Evaluation Standards for School Mathematics* (Reston, VA: 1989), p. 23.

<sup>54</sup>A typical standardized achievement test battery can be scored and reported back to schools in about 6 weeks.

time.<sup>54</sup> These tests can also be administered without any special training or equipment. The answers can be objectively scored—thus seeming to avoid any judgment or subjectivity in scoring and potential controversy that might result.

The measurement properties of multiple-choice items also make them very efficient. Many items can be administered in a relatively short amount of testing time, providing much information and making composite scores highly stable and reliable. The large number of items also allows each content domain assessed to be represented by multiple questions, which increases both the reliability and validity of the test. Because large numbers of items can be pretested efficiently, a large pool of good items with empirical description of their difficulty levels (and other item parameters of concern in the design of tests) can be developed. Items in this pool can also be tested for statistical evidence of bias. Finally, multiple-choice items have been found to perform as well as other, less efficient kinds of items (e.g., essays) for specific functions such as predicting freshman college grades.<sup>55</sup> The dominant testing technology of the present—multiple-choice items in a norm-referenced test—has been shown to be a very efficient technology for some specific purposes, in particular those purposes that require ranking individuals along a continuum. However, this is only one of many educational uses for achievement tests.

The educational advantages of multiple-choice items, the ways in which they enrich or enhance learning, are harder to articulate. Historically, educational examinations consisted of oral or written questions used to demonstrate mastery of content taught. Most other industrialized countries do not use multiple-choice examinations in education.<sup>56</sup> Multiple-choice items were pressed into service in this country when more efficient methods of testing large numbers of students were needed (see ch. 4). Each step in the historical process of examining—from oral to written examinations, then from written to multiple-choice—has taken us farther away from the actual skills, such as oral and written expression, that we want children to develop. Critics of multiple-choice items argue that we spend considerable time



Photo credit: Bob Daemmrich

These elementary school students are taking a multiple-choice achievement test that requires filling in the correct “bubble” on a separate answer sheet. Although such tests have certain advantages, many educators believe that negative effects on classroom practice indicate a need for new testing approaches.

training students in a skill not required in life, namely answering multiple-choice questions. As one analyst has observed: “. . . most of the important problems one faces in real life are ill-structured, as are all the really important social, political, and scientific problems in the world today. But ill-structured problems are not found in standardized achievement tests.”<sup>57</sup> Many educators are now arguing that achievement tests need to consist of items and tasks that are more “authentic”—i.e., are made up of skills that we actually want children to practice and master, such as producing and explaining how they reached the answer, writing a logical argument, drawing a graph, or designing a scientific experiment. These efforts are described at length in the next chapter.

One of the consistent themes of the debate throughout the last 70 years has been to ask whether more open-ended items (e.g., essays) really measure

<sup>55</sup>See, e.g., Brent Bridgeman and Charles Lewis, “Predictive Validity of Advanced Placement Essay and Multiple-Choice Examinations,” paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL, April 1991.

<sup>56</sup>A major exception is Japan, which does as much (if not more) multiple-choice testing than does the United States. See ch. 5 for discussion.

<sup>57</sup>Norman Fredericksen, “The Real Test Bias: Influences of Testing on Teaching and Learning,” *American Psychologist*, vol. 39, No. 3, March 1984, p. 199.

### Box 6-G—Setting and Maintaining Standards

Few tests in this country have attempted to provide interpretations of scores with respect to broad standards of performance. Most judgments about how well a child or school is doing have been made through the use of norms—essentially a standard based on average performance. The current effort by the National Assessment of Educational Progress to establish national proficiency levels of performance-basic, proficient and advanced-in mathematics is one such attempt.<sup>1</sup>

Consider two different methods that could be used by a teacher to grade the tests of his students. He could decide to grade them all relative to one another; in this method he looks overall the answers that have been provided and assigns the highest grade to those students with the highest scores and the lowest grade to the lowest scores. This is a norm-referenced scoring system. Several problems arise with this system. First, there is no objective referent—all of his students' answers may still be better than the best answer given in the class next door. Second, all of his students may have mastered the material of interest; if all have mastered it the actual differences that underlie a high and a low score mean very little, and will reflect very fine-grained and perhaps unimportant distinctions in their understanding. Thus, the drawback of this procedure is that a student's performance is evaluated solely with respect to the performance of others.

The second method would be to judge the work against some standard reflecting what his students should be able to do. The teacher determines what an excellent, an average, and a poor answer would look like. AU students are then judged relative to that standard. This is how many teachers assign letter grades. The most widely cited problem with a standard-based scoring system is that it is hard to equate standards across teachers. Different teachers hold different expectations for what their students should be able to do and what excellence looks like. However, reference to some absolute standard of proficiency is in many ways the most meaningful kind of score, particularly if one wants to compare progress across time or to raise the absolute level of achievement among students.

Some educational examinations, particularly in European countries, have attempted to set central standards and have used various mechanisms to maintain the consistency of the standards. In Great Britain, for example, the new national assessment involves a system of moderation of teacher judgments; initially, teachers are trained to make judgments about student performance on a number of standardized tasks. During the administration of these tasks at the end of the year, a moderator is sent to the schools to observe teachers, rate a subsample of students with the teacher, discuss discrepancies in judgments, and in various other ways maintain the consistency with which the standards are being applied by teachers in the school.<sup>2</sup>

<sup>1</sup>See ch. 3 for a further discussion of standard setting by the National Assessment of Educational Progress (NAEP).

@are **Burstall**, National Foundation for Educational Research, London, personal communication February 1991. See also Department of Education and Science and the Welsh Office, *Curriculum* (London, England: 1987).

different traits, skills, or abilities than multiple-choice items. As one reviewer states:

The enduring question for the [multiple] choice type items is whether or not these seemingly artificial contrivances measure the same thing as the more "natural and direct" free-response types of item. Popular opinion on this question is rather well formulated and almost universally negative, i.e., the two types of items do not measure the same thing. One can hear multiple-choice and true-false questions castigated in nearly any teachers' lounge in the country on a daily basis, and they are lampooned with regular frequency in cartoon strips. . . . But at

the root of the question of whether free-response and choice-type tests are measuring the same thing (trait, ability, level of knowledge) is an empirical one, not a philosophical or polemical one.<sup>58</sup>

Few data are available comparing the extent to which tests in different formats provide the same information or different information. Results of a few studies that shed light on this topic are somewhat mixed. In some areas, the research evidence suggests that multiple-choice and open-ended items measure essentially the same skills.<sup>59</sup> However, other research suggests that the extent to which open-ended or multiple-choice tests get at different

<sup>58</sup>Thomas P. Hogan, University of Wisconsin, Green Bay, "Relationship Between Free-Response and Choice-Type Tests of Achievement: A Review of the Literature," paper prepared for the National Assessment of Educational Progress, 1981.

<sup>59</sup>Ibid.; and Millman and Greene, op. cit., footnote 24.

Similarly, the International Baccalaureate Program has been developed to confer a degree on high schools students worldwide. This program can be adopted by all high schools and is used at a number of schools in the United States. In order to maintain the comparability of the credential across schools, teachers, and countries, the program has a very detailed set of curricular requirements for various courses. Teachers are carefully trained in the criteria for grading and judging performance of students in each discipline. Teachers must have their examinations approved by the central administrative program. After an examination has been given and graded, the teacher sends several student examinations—one receiving the highest score, one the middle score, and one the lowest score—to the central administrative program where standards for grading are carefully matched. Feedback is provided to the teacher if his grading standards are not in line with the central program standard.<sup>3</sup>

Recent developments in psychometric theory and its application to large-scale achievement testing also provide some encouraging evidence of the possibility of calibrating test items designed at the State or local level to a common scale. Group-level item-response theory may provide the technical model by which a shared pool of items could be created for different States or districts. A State or district would not be limited to those items but would include a sufficient number of these items so that the rest of their test could be calibrated to national norms or standards.<sup>4</sup> Such a model still requires, however, some degree of consensus about the content and curricular areas to be tested.

“Trustworthy comparative data, . . . demands a degree of agreement about the curriculum that many may consider to be a threat to local control. It is one thing to agree that arithmetic should be assessed, or even that the assessment should include applications of concepts such as ratios and percents. It may be something else to agree on the grade in which the assessment of specific skills such as these should take place or on the appropriate items.<sup>5</sup>

For subjects such as literature—what books should students read and at what age?—or social studies, these issues become even more thorny.

<sup>3</sup>Carol M. Dahlberg, coordinator, International Baccalaureate Program, Montgomery High School, Rockville, MD, remarks at OTA Workshop on Examination Systems in Other Countries and lessons for the U. S., Mar. 27-28, 1991.

<sup>4</sup>Robert L. Linn, “Accountability: The Comparison of Educational Systems and the Quality of Test Results,” *Educational Policy*, vol. 1, No. 2, June 1987, pp. 181-198; and R. Darrell Bock and Robert J. Mislevy, “Comprehensive Educational Assessment for the States: The Duplex Design,” *CRESST Evaluation* November 1987.

<sup>5</sup>Linn, *op. cit.*, footnote 4, p. 196.

skills will depend on the subject matter being tested. Evidence is strong, for example, that essay tests of writing provide different information than do multiple-choice tests of writing.<sup>60</sup> In part, the potential usefulness of open-ended items will depend on the purpose of the particular test and the kind of information needed.

### Multiple Choice: A Renewable Technology?

Because of concerns related to efficiency, reliability, and economy, many researchers and test developers think that the multiple-choice test will probably always have some role to play in the assessment of achievement. Therefore, educators and psychometricians have become interested in exploring ways

to improve the multiple-choice items that currently dominate standardized achievement tests. A number of State assessment programs have put efforts into developing multiple-choice items that seem to require more complex thinking skills and are more consistent with their changing educational goals.

For example, Michigan recently decided to move away from an exclusively skill-based approach to reading. New statewide reading objectives were developed consonant with a redefinition of reading as a process that involves constructing meaning through a dynamic interaction between the reader, the text, and the context of the reading situation. A new approach to assessing these goals *was also* needed, so the State embarked on developing new

<sup>60</sup>R.E. Traub, “On the Equivalence of the Traits Assessed by Multiple-Choice and Constructed-Response Tests,” *Construction Versus Choice in Cognitive Measurement*, R.E. Bennett and W.C. Ward (eds.) (Hillsdale, NJ: L. Erlbaum Associates, in press); and Edys S. Quellmalz, “Designing Writing Assessments: Balancing Fairness, Utility and Cost,” *Educational Evaluation and Policy Analysis*, vol. 6, No. 1, spring 1984, pp. 63-72. It should also be noted that much of the research that does exist about item differences has been based on college or college-bound students and 4’ . . . hence those of (a) above average ability, (b) beyond the years of rapid cognitive development, and (c) from predominantly middle-class, White, Western cultural background.” Hogan, *op. cit.*, footnote 58, p. 46. Some of the field studies conducted as part of the National Assessment of Educational Progress can and will provide much needed data about the performance of a diverse population of elementary and secondary students.

tests to be used with grades 4, 7, and 10. Michigan's innovative reading assessment program involves many changes in the tests—including the use of stories drawn from children's literature and other primary sources instead of short excerpted passages or ones written for the test—while still employing a multiple-choice format for answering the questions. Such questions are designed to assess “constructing meaning” and “knowledge about reading” as well as factors typically not tested such as a child's familiarity with the topic of the story and his or her effort and interest in the testing questions.<sup>61</sup>

A point that is consistently made by those who design educational tests is that multiple-choice items are not restricted to assessing only basic skills or the memorization of facts.<sup>62</sup> Multiple-choice items, if carefully crafted, can be used to assess very high levels of expertise—for example in admissions tests for graduate education (Law School Admission Test, Graduate Record Exam) and board certification examinations for physicians. The ACT Science Reasoning Test, which is part of the ACT used for college admissions, uses multiple-choice items to assess interpretation, analysis, evaluation, reasoning, and problem-solving skills required in the natural sciences. Each unit on the test presents scientific information—in the form of graphs, results of experiments, or descriptions of conflicting scientific theories—that the student must interpret. According to the test designers, advanced knowledge in the subjects covered by the test (biology, chemistry, physics, and the physical sciences) is not required; instead the test emphasizes scientific reasoning skills.<sup>63</sup> The National Assessment of Educational Progress (NAEP) has also put considerable effort into developing multiple-choice items to measure thinking skills such as solving problems and conducting inquiries in science, conceptual understanding and problem-solving in mathematics,

and evaluating information and constructing meaning in reading. See figure 6-4 for examples of items drawn from these and other multiple-choice tests designed to assess more complex thinking skills.

Recent research and development efforts have suggested additional ways that multiple-choice tests might be designed to reflect complex processes of learning and development:

- One effort to assess science understanding has focused on trying to describe the various “mental models” that children hold before they master the correct understanding of basic scientific principles. Multiple-choice items, such as the one in figure 6-5, are then designed to represent these various mental models; each distracter (or incorrect choice) represents a commonly held misconception about a scientific principle. Wrong answers can be examined by the teacher to discern what misconceptions each child may hold and better focus instruction.<sup>64</sup>
- Similarly, if free-response answers given by children to all kinds of open-ended tasks can be analyzed, then the kinds of misunderstandings and errors commonly made by children can be described. This information can be used to write distracters that reflect these errors (not just to “trick” students) and may then be useful in diagnosing mistakes and error patterns.
- Researchers for some time have explored ways of giving partial credit for partial understanding on multiple-choice questions. One method of doing this involves giving different weights or points to different answers that are written to reflect incorrect, partial, and complete understanding of the solution. Partial credit scoring procedures are particularly relevant for diag-

<sup>61</sup>For more information on the new Michigan reading tests see Edward Roeber and Peggy Dutcher, “Michigan's Innovative Assessment of Reading,” *Educational Leadership*, vol. 46, No. 7, April 1989, pp. 64-69; and Edward D. Roeber, Caroline S. Kirby, Geraldine J. Coleman, Peggy A. Dutcher, and Robert L.C. Smith, *Essential Skills Reading Test Blueprint*, 5th ed. (Lansing, MI: Michigan Department of Education, Michigan Educational Assessment Program, July 1989).

<sup>62</sup>William A. Mehrens, “Using Performance Assessment for Accountability Purposes: Some Problems,” paper presented at the annual meeting of the American Educational Research Association @ Chicago, IL, 1991; Anastasi, op. cit., footnote 4; Thomas M. Haladyna, “Context-Dependent Item Sets,” *Educational Measurement: Issues and Practice*, in press; Millman and Greene, op. cit., footnote 24; and Lewis R. Aiken, “Writing Multiple-Choice Items to Measure Higher Order Educational Objectives,” *Educational and Psychological Measurement*, vol. 42, No. 3, autumn 1982, pp. 803-806.

<sup>63</sup>The American College Testing program, *The ACT Assessment Test Preparation Reference Manual for Teachers and Counselors* (Iowa City, IA: December 1990).

<sup>64</sup>Richard J. Shavelson, Neil B. Cary, and Noreen M. Webb, “Indicators of Science Achievement: Options for a Powerful Policy Instrument,” *Phi Delta Kappan*, vol. 71, No. 9, May 1990, pp. 692-697.

Figure 6-4--Sample Multiple-Choice Items Designed To Measure Complex Thinking Skills

<p style="text-align: center;"><b>Thinking Skill:<sup>a</sup></b> <b>Knowing Science</b></p> <p style="text-align: center;"><i>Grade Levels: 4, 8, 12</i></p> <table style="margin-left: auto; margin-right: auto;"> <tr> <td style="text-align: center;">Always True</td> <td style="text-align: center;">Sometimes True</td> <td style="text-align: center;">Never True</td> </tr> </table> <p>Scientists should report exactly what they observe . . . . *</p> <p>Belief is the main basis for scientific knowledge . . . . . *</p> <p>Knowledge is the goal of scientific work . . . . . *</p> <p>Scientific knowledge can be questioned and changed . . . . .</p> <p>Knowledge discovered in the past is used in current scientific work . . . . . *</p> <p>Scientists who do experiments find answers to their questions . . . .</p> <p style="text-align: center;"><i>Grade Level: 4</i></p> <p>The methods of science can be used to answer all of the following questions EXCEPT:</p> <p>*(A) Are puppies more beautiful than spiders? (B) How many oak trees grow in Pennsylvania? (C) Which laundry detergent cleans best? (D) What are the effects of lead pollution on trout?</p>	Always True	Sometimes True	Never True	<p style="text-align: center;"><b>Thinking Skill:<sup>c</sup></b> <b>Summarizing Ideas</b></p> <p><b>Read</b> the sentence. Then choose the essential phrase that should be included in research notes for a paper on the subject.</p> <p>Despite the fact that Puritan forces in England objected to plays and tried to interfere with performances, theatrical entertainment enjoyed great popularity in Shakespeare's time, both with the public and with the members of the royal court.</p> <p style="margin-left: 40px;">A royal court enjoyed plays during Shakespeare's time</p> <p>* B plays popular despite objection and interference by Puritans</p> <p style="margin-left: 40px;">C theatrical entertainment very popular with the public</p> <p style="margin-left: 40px;">D Puritans object to public performances</p>
Always True	Sometimes True	Never True		
<p style="text-align: center;"><b>Thinking Skill:<sup>b</sup></b> <b>Applying Principles</b></p> <p style="text-align: center;"><i>Grade 8</i></p> <p>If the law of supply and demand works, the farmer will obtain the highest price for crops when</p> <p style="margin-left: 40px;">A. both supply and demand are great . B. both supply and demand are low . C. supply is great and demand is low . *D. supply is low and demand is great .</p>	<p style="text-align: center;"><b>Thinking Skill:<sup>c</sup></b> <b>Comprehension</b></p> <p><b>Read the question and</b> then choose the best answer.</p> <p>Which of these is most like an excerpt from a myth?</p> <p>* A And so the turbulent sea suddenly grew calm as Father Neptune urged his steeds forward and flew off toward the setting sun.</p> <p>B Gold coins were reported to have come from an ancient Phoenician ship that sank off the island during Homeric times.</p> <p>C We lowered the sails but the <i>Moon Goddess</i> still lurched violently on the crashing waves as we prepared to ride out the storm.</p> <p>D Retrace the voyage of Ulysses in a 2 1-day adventure that takes you from Asia Minor to the islands and mainland of Greece.</p>			

\* Correct answers for multiple-choice items are indicated by an asterisk (\*).

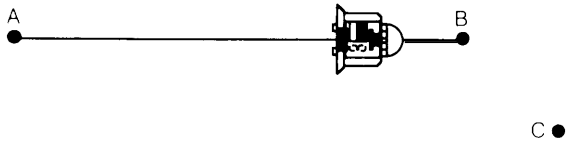
<sup>a</sup>SOURCE: National Assessment of Educational progress, *Science Objectives: 1990 Assessment*, booklet No. 21-S-1() (Princeton, NJ: 1989), pp. 45-46.

<sup>b</sup>SOURCE: Connecticut State Department of Education, *Connecticut Assessment of Educational progress 1982-83: Social Studies summary and Interpretations Report* (Hartford, CT: 1984).


<sup>c</sup>SOURCE: CTB/McGraw-Hill, *Comprehensive Test of Basic Skills (CTBS) Class Management Guide: Using Test Results* (Monterey, CA: 1990), pp. 68, 70. These are sample items that do not appear on an actual test.


**Figure 6-5-Sample Multiple-Choice Item With Alternative Answers Representing Common Student Misconceptions**


A spaceship is drifting sideways in space from point A to point B. It is not affected by outside forces. At point B, its engine fires to produce a constant thrust at a right angle to AB. At point C, the engine is shut off again.





Which of the following (1, 2, 3, 4, or 5) best represents the path of the spaceship?

1. 

2. 

3. 

4. 

5. 

The correct answer is 5.

NOTE: The alternatives presented represent both the correct mental model of the effect of forces on a spaceship and a variety of possible answers based on known, erroneous mental models that children hold.

SOURCE: R.J. Shavelson, N.B. Carey, and N.M. Webb, "Indicators of Science Achievement: Options for a Powerful Policy Instruments," *Phi Delta Kappan*, vol. 71, No. 9, May 1990, p. 697.

nostic tests designed to describe a student's strengths and weaknesses.<sup>65</sup>

- The complex multiple-choice item is a widely used format in medical and health professions' testing programs where many questions have more than one right answer. In this item type, four or five answers are presented and the student can select any number of correct responses from none to all.<sup>66</sup>
- Another way that multiple-choice items can be used to measure more complex understandings is to group a series of them together based on a

common set of data. The data may be in the form of charts, graphs, results of experiments, maps, or written materials. Students can be asked "... to identify relationships in data, to recognize valid conclusions, to appraise assumptions and inferences, to detect proper applications of data, and the like."<sup>67</sup>

### Redesigning Tests: Function Before Form

Test use in schools has been increasing. Much of the increase in the volume of school-based testing in the last decade has come from its rising popularity as

<sup>65</sup>Millman and & op. cit., footnote 24; Thomas M. Haladyna, "The Effectiveness of Several Multiple-Choice Formats," *Applied Measurement in Education*, in press. For a discussion of ways in which test theory will have to develop and change in order to accommodate the measurement of problem-solving strategies and misconceptions see Robert J. Mislevy, *Foundations of a New Test Theory*, ETS Research Report RR 89-52-ONR (Princeton, NJ: Educational Testing Service, October 1989).

<sup>66</sup>Haladyna, op. cit., footnote 65. This item type has been found to have a number of technical problems. Haladyna recommends the related five-option "multiple true-false" item.

<sup>67</sup>Gronlund and Linn, op. cit., footnote 16, p. 193.



Table 6-3-Functions of Tests: What Designs Are Needed?

	Classroom instructional guidance	System monitoring	Selection, placement, and certification
<i>Who needs to be describe d....., . . . . .</i>	Individuals	Groups of students	Individuals
<i>"Stakes" or consequences attached . . . . .</i>	Low	High or low	High
<i>Characteristics of the test needed</i>			
Comparability of information. . . . .	Low	High	High
Impartial scoring (not teachers) . . . . .	No	Yes	Yes
Standardized administration . . . . .	No	Yes	Yes
<i>Type of information needed</i>			
Detailed v. general . . . . .	Detailed	General	General
Frequency . . . . .	Frequently during a single school year	Once a year or less	Once a year or less
Results needed quickly . . . . .	Yes	No	No
<i>Technical requirements</i>			
Need for high test reliability (internal insistency and stability) . . . . .	Can vary	Depends on size of group if low stakes: content	Very high Content
Type of validity evidence . . . . .	Content	If high stakes: content and construct	Additional validity evidence must be demonstrated for the specific purpose (e.g., certification = criterion validity, selection = predictive validity)

SOURCE: Office of Technology Assessment, 1992; adapted from Lauren B. Resnick and Daniel P. Resnick, "Assessing the Thinking Curriculum: New Tools for Educational Reform," paper prepared for the National Commission on Testing and Public Policy, August 1989 (To appear in B.R. Gifford and M.C. Connor (eds.), *Future Assessments: Changing Views of Aptitude, Achievement, and Instruction* (Boston, MA: Kluwer Academic Publishers, in press).)

an accountability tool for policymakers interested in a measure of system effectiveness (see ch. 2). The available testing technology—norm-referenced multiple-choice tests—has been pressed into service even when the properties of this technology were not well matched to the needs of the users. Similarly, there has been increasing interest in the role that tests can play in fostering learning and knowledge acquisition in the classroom. For tests to have educational value to the student in the classroom, educators argue, the tests must be frequent, provide feedback in a timely fashion, and make clear the expectations and standards for learning. A single testing technology no longer seems enough for the needs of multiple users. How, then, should we redesign achievement tests to better serve multiple testing needs?

Table 6-3 summarizes the characteristics of tests required for each of the three main functions of testing. Consider first the system monitoring function of tests. In this case only groups of students need to be described, that is classrooms, schools, districts, or States. Individual scores are not needed. This means that sampling methodologies can be used—a representative subset of students can be tested and accurate information obtained. One of the advan-

tages of a sampling methodology is that no individual scores are available, thus preventing their use for unintended purposes such as selecting students for special programs or grouping students according to ability. One of the drawbacks sometimes cited for sampling, however, is that students may not be particularly motivated to give their best performance when they are not going to receive personal scores (see ch. 3).

In system monitoring, managerial uses can include information that has both high and low stakes. Purely informational uses (without consequences) may include program evaluation and curricular evaluation. Similarly, some administrators may want information about how their system is doing but may not attach any particular rewards, sanctions, or expectations to the test scores; test results would have a "temperature taking" function. NAEP is an example of a test designed to provide nationally representative information of this type. However, increasingly tests are being used for accountability purposes—rewards and consequences are attached to the results of those tests and they are being used as a lever to motivate improvement. When this happens, the informational value of the test can be compromised. Attention is readily focused on test

performance as a goal of instruction; in this case improvement in test scores may or may not signal growth in real achievement.<sup>68</sup>

Many of the characteristics of tests designed for monitoring systems are those expected from standardized achievement tests. It is very important that the results obtained from these tests be comparable across students and that they can be aggregated in a meaningful way. This means that the tests must be standardized in administration and scoring. Impartial scoring is very important. The monitoring of systems requires general information at occasional intervals (usually once a year or less). The results are not needed immediately.

Tests used for selection, placement, or certification differ from tests used for system monitoring in several major ways. First, each student must receive a score. Second, the kinds of decisions these tests are used to make are almost always high stakes—they can have significant consequences for an individual's educational career. Tests used for selection, placement, and certification must meet exceptionally high standards of comparability, reliability, and validity. As with tests used for monitoring systems, impartial scoring and standardized administration are required; similarly the information required is general, needed infrequently (once a year or less) and not required quickly.

The third major difference is in the kind of validity evidence required. Tests for selection, placement, or certification must be validated for each of those specific uses. Thus certification tests need criterion-related validity evidence particularly related to the "cutoff scores" that are established to certify mastery. Selection tests need predictive validity evidence demonstrating that test results relate to future performance or ability to benefit from a particular resource or intervention. In the current debate about redesigning tests, there is little discussion by educators or measurement specialists about needing or using various new test designs for selection. In part, this may be due to a fairly widespread and entrenched belief that selection tests are not appropriate for elementary school and, for the most part, not within secondary school either.<sup>69</sup>

Tests designed for classroom use are the most divergent in their design requirements (see table 6-3), differing significantly both from existing and new tests designed to serve managerial functions. Tests used by teachers to monitor learning and provide feedback need to provide detailed information on a frequent basis, as quickly as possible. Because classroom tests are very closely related to the goals of instruction, time spent on testing need not be considered "wasted time." As testing at the classroom level becomes more integrated with instruction, the time constraints so often in-posed on tests can be relaxed considerably because time spent on tests is also time spent learning. Because these tests do not carry high stakes and because they are not going to be used to make comparisons among students or schools, they are free of many of the stringent requirements of standardization, impartial scoring, and need for comparability. However, the more that teachers or school systems want these classroom level tests to be useful for other purposes, i.e., to make high-stakes decisions about individuals or to aggregate the information across classrooms or schools, the more that these classroom tests will need to incorporate features that provide comparability and standardization. **It is difficult to prevent** the misuse of information once that information has been collected. One of the dangers, therefore, in relaxing technical standards for classroom tests is that the use of the scores cannot be restricted or monitored appropriately once they are obtained.

How can the various functions of testing and design requirements be coordinated with one another? Most investigators working in test design today believe that one test cannot successfully serve all testing functions.

Many of the features of tests that can effectively influence classroom learning are very different from the requirements of large-scale managerial testing. Many testing experts believe that we need two distinct types of tests to serve these two functions

---

<sup>68</sup>For a discussion of the "Lake Wobegon Effect" and other evidence about how gains in test scores can be attained without affecting "real achievement," see ch. 2.

<sup>69</sup>Haney, *op. cit.*, footnote 35.

because the requirements are so divergent.<sup>70</sup> The Pittsburgh school district, for example, has developed a diagnostic testing system, called Monitoring Achievement in Pittsburgh (MAP), which is characterized by tests closely aligned with curricula, brief and frequent administration of those tests, and rapid turnaround of results. These test results are then used to inform instruction, as teachers can see whether an objective that has been covered has, in fact, been learned by the class and tailor instruction accordingly. Pittsburgh uses a different test for system monitoring; analyses have suggested that recent gains on this traditional norm-referenced test are largely due to the effects of MAP.<sup>71</sup>

## Conclusions

No testing program operates in a void. The effects of any testing program on the school system as a whole, or of different tests on one another, need to be continually monitored. The effect of other testing requirements, imposed by the State or a special program such as Chapter 1, may also affect the impact of a new test or new reform program. The

consequences of a given test—to the individual student, the teacher, the school—will heavily influence the effects of that test on learning and instruction. A beautifully designed and educationally relevant test may have no impact if no one looks at its scores; the poorest quality test available could conceivably influence much of a school's educational climate if the stakes attached to it are high.

What a test looks like—the kinds of tasks and questions it includes—should depend on the intended purpose of the test. As the next chapter will illustrate, test formats can vary widely from multiple-choice to essays to portfolios. Different types of testing tasks will be more or less useful depending on the purpose of the test and the type of information needed. The purpose of a test and a definition of what it is intended to assess need to be carefully determined *before* test formats are chosen. Moreover, critical issues such as bias, reliability, and validity will not be resolved by changing the format of the test.

<sup>70</sup>Paul G. LeMahieu and Richard C. Wallace, Jr., "Up Against the Wall: Psychometrics Meets Praxis," *Educational Measurement: Issues and Practice*, vol. 5, No. 1, spring 1986, pp. 12-16; and Educational Testing Service, "Instructional and Accountability Testing in American Education: Different Purposes, Different Needs," brochure, 1990.

<sup>71</sup>LeMahieu and Wallace, op. cit., footnote 70; and Paul G. LeMahieu, "The Effects on Achievement and Instructional Content of a Program of Student Monitoring Through Frequent Testing," *Educational Evaluation and Policy Analysis*, vol. 6, No. 2, summer 1984, pp. 175-187.