7

Interpreting and Using National and State Large-Scale and Standards-Based Tests

This chapter is about the use of what I refer to as "large-scale" assessments (tests), primarily to distinguish them from what is developed and used by teachers in their classrooms. For many decades educators have used standardized achievement and ability tests, which are characterized as being "large-scale" because of their wide usage. More recently, states have developed their own large-scale standards-based tests. Both of these kinds of assessments are used for many purposes:

- To identify students who may be eligible to receive special services
- To monitor student performance from year to year
- To identify students' academic strengths and weaknesses
- To identify and monitor achievement gaps
- To determine readiness for new academic work
- To select students for special programs
- To improve teaching
- To evaluate programs
- To give feedback to students and parents

- To compare schools and students
- To evaluate curriculum
- To determine if performance standards for grade promotion and graduation have been demonstrated
- To evaluate teachers
- To evaluate principals
- To evaluate and accredit schools

Some of these purposes, such as monitoring student performance through time and predicting performance, have been used to justify the administration and reporting of standardized ability (aptitude) and achievement tests for decades. In recent years, use of large-scale assessment for the last four purposes listed above has accelerated and intensified. When large-scale assessments, whether at the state, local, or national level, are used to deny students promotion in grade or graduation, to evaluate school personnel, or to accredit schools, the results of the assessments have serious implications. As previously noted, these are called "high-stakes" assessments (of course, it is not the assessments that are high stakes, but the way the results are used). Because high-stakes assessments are commonly employed at all levels, it is imperative to understand appropriate uses and limitations when interpreting the scores.

This chapter will first examine different types of large-scale tests and how they are administered and prepared for. I will also review the nature of the scores that are reported and how to interpret them. Finally, we will look at how large-scale test scores can be used to improve instruction.

TYPES OF LARGE-SCALE AND STANDARDIZED TESTS

Table 7.1 summarizes five ways of classifying large-scale tests. Each of these categories describes characteristics that influence the makeup of the assessment and the way results are reported and interpreted. With some exceptions, the categories are independent of one another. For example, a test that is national in scope could serve an achievement or aptitude function, use norm- or criterion-referenced interpretations, and be administered individually or to groups of students. The most important difference in large-scale assessments is in function.

Table 7.1 Ways of Classifying Characteristics of Large-Scale and Standardized Tests

Function	*Scope*	*Interpretation*	*Level*	*Publisher*
Achievement Ability	National State District	Norm-referenced Criterion/Standards-referenced	Individual Group	Federal government Commercial State District

Let's consider in further detail the nature of assessments that are typical in each of these two major types.

Standardized Achievement Tests: What Do Students Know?

Standardized achievement tests (standardized tests) have been a mainstay of American education for decades. The purpose of these tests is to measure how much students have learned in specific, well-defined content areas such as reading, mathematics, science, social science, and English. The tests are typically published by companies for profit, for use throughout the country. This has some important implications. Because the tests need to appeal to a broad spectrum of potential users, the material that is covered is common to most school districts. The advantage of broad coverage is that these tests measure outcomes and content that are shared by most schools in the nation. This allows significant comparisons with the achievement of other students throughout the country. A disadvantage of such broad coverage of content, however, is that there may not be a good match between the test and the local curriculum. Close inspection of the test objectives and types of items is needed to determine the extent of the match.

The most common type of national norm-referenced standardized test is the *survey battery*, which consists of a comprehensive set of subject matter tests, all normed on the same group. This type of norming allows comparisons between the subject matter areas tested to help determine students' relative strengths and weaknesses. That is, it can be determined that a student is strong in computational skills and weak in reading comprehension. It is not possible to make such determinations from different tests. Common norm-referenced standardized achievement survey batteries include the following:

- *TerraNova* California Achievement Tests
- Iowa Test of Basic Skills
- Metropolitan Achievement Test
- Stanford Achievement Test Series

In addition to providing survey batteries, publishing companies also develop tests in specific subjects, such as mathematics and reading. These tests cover a content area with greater depth and breadth. Although a subject-specific test uses the same types of items as a survey battery, there are many more questions on each subject than are found on a survey battery, which allows greater specification of strengths and weaknesses within a subject. Often, these tests are used as a follow-up to a survey battery that may have identified possible difficulties. Reading tests, for example, are used extensively to focus on a number of specific reading skills. Thus, although a survey battery may provide scores in language, reading comprehension, and vocabulary, a subject-specific reading test may examine students' decoding skills, their ability to identify the main idea in a passage, and their ability to use context to understand the meaning of words.

The primary purpose of achievement test batteries is to *survey* student learning and obtain an overall performance score for each content area as well as major categories within each content area. The results are relatively broad or general. As such, they do not typically provide teachers with much information that is helpful for instruction. To address this limitation, test publishers have developed *diagnostic batteries* for the major content areas. Diagnostic batteries, typically given in mathematics, reading, and language, provide criterion-referenced interpretations to identify a student's specific strengths and weaknesses in a particular subject. These results help teachers make instructional decisions such as whether students need remediation. For example, the Metropolitan Achievement Test series includes a norm-referenced reading survey test and a criterion-referenced reading diagnostic test. The diagnostic test provides scores in the following areas:

- Visual discrimination
- Letter recognition
- Auditory discrimination
- Sight vocabulary
- Phoneme/grapheme: consonants
- Phoneme/grapheme: vowels
- Vocabulary in context
- Word part clues
- Rate of comprehension
- Skimming and scanning
- Reading comprehension

As the nature of the list indicates, the diagnostic scores focus on areas that parallel instruction. Many testing companies offer tailor-made criterion-referenced diagnostic tests for states or districts. The state or district selects the objectives it wants measured from a large bank of objectives provided by the publisher. Items measuring the objectives are then pulled from a large bank of items provided by the publisher. For example, the Multiscore System has more than 1,500 objectives and 5,500 test items.

Some achievement batteries attempt to serve both survey and diagnostic purposes. This is accomplished by giving a score in major categories and then breaking down student performance on specific items corresponding to the diagnostic information that is useful for instruction. Because there are only a few items in each area, however, be wary of claims that the test can be both norm-referenced and criterion-referenced. In general, a minimum of 5 items is needed to make a reliable interpretation of the student's skill level or knowledge. Yet some test reports will contain 3 or 4 items for a specific area or skill, indicate how many were answered correctly, and then provide a judgment about competence. These summaries can provide an initial indication of achievement, but there needs to be further assessment to be sure about the student's strengths and weaknesses. Remember, test publishers want to sell as many tests as possible, so they do whatever they can to appeal to potential buyers

who want a test that serves both norm-referenced and criterion-referenced purposes. Survey battery norm-referenced interpretations are better than survey criterion-referenced interpretations.

Standards-Based Large-Scale Tests

The most recent and very influential trend in testing is for both large and small test publishers to provide assessments that are aligned with state standards. Such tests are customized to states and school districts so that they measure state standards and, at least according to the publishers, help design more effective instruction. This has opened a very large and profitable market for test publishers who heretofore were concerned with national standardized achievement and ability tests.

In contrast to national achievement tests, however, which tend to be norm referenced, state standardized tests are usually criterion-referenced or standards-referenced, hence the name *standards-based*. The purpose of these tests is typically related to specific learning outcomes, rather than to the broader skills and knowledge tapped by national norm-referenced tests. The results of these state tests often have direct consequences for students and schools. For students, specified levels of performance may be required to graduate from high school or even to be promoted to the next grade. School accreditation may depend on students obtaining certain scores.

Most major test publishers also claim that their tests can be used for formative assessment, designed to provide information to improve achievement and evaluate instruction as well as summative achievement. Consider what Pearson, a large test publisher, says its formative assessments can do:

> Formative assessment is used to guide student instruction and learning, diagnose skill or knowledge gaps, measure progress, evaluate instruction, and report NCLB-related data . . . to determine what concepts require more teaching and what teaching techniques require modification . . . to use results to evaluate instruction strategies, curriculum and teachers, and make adjustments for better student performance. (Pearson Education, 2007)

The company goes on to show how their myriad services can meet these goals, such as alignments to state and national standards, test banks, tools to create classroom assessments, school and class comparisons, student response pads (clickers), test generators, and special reports to indicate progress over time. Teachers are able to build tests from items written by "experts." Such claims are, by and large, just that—claims. The marketing is directly addressing the need for school and student accountability. But the buyer must beware! The profit motive may be corrupting the essence of formative assessment by suggesting that the test results are applicable to individual teachers. Most importantly, these systems rarely provide instructional correctives, and often teachers don't need a constant stream of tests to know which students are struggling.

Also, such systems provide so much data that teachers and administrators are swamped with scores, buried with reports that contain student and item performance information.

National Assessment of Educational Progress (NAEP)

A fairly longstanding federal testing program, the National Assessment of Educational Progress (NAEP), has become a very important indicator of national, state, and district achievement. Since 1969, NAEP has been known as "the nation's report card" because representative sampling from the nation has been used to track student achievement in reading, writing, mathematics, science, citizenship, history, art, social studies, and additional subjects. Performance data are reported for the nation and for various subgroups categorized by region, gender, race/ethnicity, parental education, type of school, and type and size of community. Instructional practices are also surveyed and related to the achievement scores. Samples are selected carefully for each administration to allow longitudinal analyses of the results. This provides one of the few national student achievement indicators that can pinpoint progress toward increasing performance through many years. Current NAEP data are reported as scale scores (0 to 500) and according to the percentage of students placed in one of three reporting categories: basic, proficient, and advanced. These categories represent levels of achievement, and different scores are needed to be classified in one of these three.

NAEP scores are used for criterion-referenced interpretation. While NAEP is large-scale and standardized, norm-referenced data are not provided. Rather, results are presented according to the number and percentage of students achieving different levels of performance. It is critical to understand the meaning of each of these levels when using NAEP scores. There has been controversy about the labels and what they mean, especially in light of the fact that state findings may differ significantly. Recently, the discrepancy between NAEP and state results for No Child Left Behind (NCLB) has been so great that one is left wondering what is going on. Some actual comparisons of percentages of students labeled "proficient" in four states in 1997 and 2005 are presented in Table 7.2. While the "gap" has narrowed, we don't know by these numbers alone which is most accurate. In fact, many have used the NAEP scores as evidence that schools are measuring up.

The greatest value of NAEP is giving the country a standardized measure of student proficiency across years. Trends reported by NAEP are more important than the actual percentages of students reported to be in each level.

Standardized Ability Tests

Standardized ability tests measure a student's cognitive ability, potential, intelligence, reasoning, or capacity to learn. The purpose of these tests is to predict future performance or behavior. The ability that is measured is determined by both in-school and out-of-school experiences. This is how these tests differ from achievement tests. By including more out-of-school experiences, a broader

Table 7.2 1997 NAEP and State Percentages of Fourth Graders at the "Proficient" Level

State	*NAEP*	*State Test*
Wisconsin	35	88
North Carolina	30	65
Georgia	26	67
South Carolina	20	82

set of skills is measured. But the difference between standardized achievement and ability tests is one of degree. There is considerable overlap in what is covered, and often the same or similar items are used in both types.

Ability tests are developed to enable prediction of future performance by assessing current general ability (not innate capacity that cannot change). An understanding of the general ability level of students is helpful in designing appropriate instruction and grouping of students. Suppose one class, overall, has a low ability level, and another class has a high ability level. Would it make sense to use the same instructional approach and materials for both classes? Would it be reasonable to give the same homework assignments to each class?

An important aspect of standardized ability tests is the nature of the theory that is used as a basis for defining *ability*. Many theories can be used, each of which results in a unique conceptualization and interpretation. For many years, ability tests were designed to assess general cognitive ability, or intelligence. More recently, two trends have emerged. First, the language associated with these tests has changed. It is now common to call these assessments *ability* tests rather than aptitude tests (e.g., school ability, cognitive ability, or learning ability), despite little change in the nature of the initial aptitude tests. This change in language implies that *ability* communicates both innate and experiential influences, whereas *intelligence* tends to put the focus on innate, inherited characteristics.

Second, theories of abilities have changed considerably in the last three decades, stressing new capabilities and conceptualizations. Although early aptitude tests, such as the Stanford-Binet, were based on Binet's theory of intelligence, Charles Spearman's development of *g* (general factor) and specific factors, and Thurstone's primary mental abilities theory, later developments by Robert Sternberg (1985) and Howard Gardner (1993) have offered new insights into the nature of abilities. Gardner, for example, postulates that there are *multiple intelligences*, including musical, interpersonal, bodily/kinesthetic, linguistic, and intrapersonal. Sternberg's triarchic theory of intelligence includes the internal world of the individual (further divided into metacomponents, performance components, and knowledge-acquisition components), experiences of the individual, and external contextual abilities as three aspects of intellectual functioning. These new conceptualizations imply that although there is undoubtedly a general ability for abstract thinking, evidence for validity needs to be grounded in appropriate theory that may measure more specific capabilities that are relevant for predicting performance.

Group Ability Tests

Most ability assessments administered in schools are group tests, in which all students respond to written questions at one time. The items are usually multiple-choice to allow for efficient machine scoring. Three widely used group ability assessments are the Test of Cognitive Skills, the Otis-Lennon School Ability Test, and the Cognitive Abilities Test. The Cognitive Abilities Test for Grades K through 12 (Grades 3–12 for the Multilevel Edition) is a good example of the types of abilities that are measured. There are nine subtests in the Multi-Level Edition, grouped into three categories (verbal, nonverbal, and quantitative) that are used for reporting the results. There is also an overall, or *composite*, score, but no scores for the nine subtests.

Both the verbal and quantitative areas stress skills that are not directly taught in school, but the items require the use of skills that are learned in and outside school. The nonverbal section focuses on reasoning skills and is a good measure of reasoning abilities for language deficient students or poor readers.

Group ability tests are used primarily as screening devices. They are designed to identify students whose abilities deviate substantially from the norm. Individual assessment is typically carried out for students with suspected deficiencies.

Individual Ability Assessments

An individual ability assessment is conducted by a trained test examiner with one examinee. It is an oral test administered face-to-face. With a one-on-one administration, the results are usually more dependable and informative than are group tests. The examiner can take into consideration such variables as examinee motivation, handicapping conditions, and persistence. These tests are used routinely in the identification of educational disabilities that qualify a student to receive special education services and placement in special programs.

Four commonly used individual ability assessments are summarized in Table 7.3. The Stanford-Binet Scale and Wechsler Scales have a long tradition of use in schools. Each of these scales is focused solely on abilities. Together, the scales provide global measures of intelligence in a few major areas. The Kaufman Assessment Battery and Woodcock-Johnson III Tests of Achievement are more recently developed tests that include both ability and achievement scales. The use of results from these two assessments is similar to that of the Stanford-Binet and Wechsler scales, but because there is a measure of achievement, more direct comparisons can be made between achievement and ability. This provides a more complete picture for diagnostic purposes. As with all standardized tests, each of the individual ability tests measures different skills. This makes it difficult to compare the scores from two or more of these tests. It also means that appropriate interpretations of the results occur only if it is clear what certain scores mean. This is best understood when those making the interpretations have knowledge about the nature of the specific items used in the test.

Table 7.3 Summary of Common Individual Aptitude Assessments

Stanford-Binet Scale, Fourth Edition	*Wechsler Intelligence Scale for Children—Revised (IV)*	*Kaufman Assessment Battery for Children, Second Edition*	*Woodcock-Johnson III Tests of Achievement*
Published in 1985, the fourth edition updates the content but retains the basic structure of previous editions. Given one-on-one to individuals age 2 to 23, the test has 15 subtests grouped into four areas: quantitative reasoning, verbal reasoning, abstract/visual reasoning, and short-term memory.	Revised and restandardized in 2003, the WISC-IV is designed for use with individuals age 6 to 16 years of age. Administered one-on-one, the WISC-IV contains 10 subtests, 5 verbal and 5 performance.	Revised in 1983, the KABC provides a comprehensive assessment of both intelligence and achievement for children age 3 to 12. Sixteen subtests are combined into three regularly administered scales. Intelligence is assessed with three scales. The Kaufman is marketed as "culturally fair" ability test.	Individually administered battery of tests to assess intelligence and academic achievement of individuals age 4 through adulthood. Revised in 1989, the WJ-R contains a cognitive battery of 21 subtests (7 in the standard battery, 14 in the supplementary battery) to give a measure of intelligence, and 14 achievement subtests (9 in the standard battery, 5 in the supplementary battery).

ADMINISTERING LARGE-SCALE AND STANDARDIZED TESTS

Because most large-scale and standardized testing in schools takes place in classrooms, teachers will be responsible for administering the tests to students. It is important for the teachers to follow the directions carefully and explicitly. Instructions for the test administrators are provided in writing by the test publisher. These must be strictly followed, including any time limitations. Even if the teachers have administered similar tests in the past, the written directions should be read. The directions will indicate what to say to students, how to respond to student questions, and what to do while students are working on the test. Some portion of the instructions will probably direct the teachers to read directly from the instructions to the students, word for word as specified.

During the test, teachers may answer student questions about the directions or procedures for answering questions but should not help students in any way with an answer or what is meant by a question. Students should not be told to "hurry up" or "slow down." Teachers need to suspend their role as instructors and take on the role of test administrators. This isn't easy because most teachers want to help students do their best, especially with high-stakes tests that may reflect on the teacher. Whether the standardized assessment is national, state, or district in scope, the importance of strictly following directions cannot be overemphasized.

While observing students as they take the test, teachers may see some unusual behavior or events that could affect the students' performance, such as interruptions or students acting out. These behaviors and events should be recorded for use in any subsequent interpretation of the results.

PREPARING STUDENTS TO TAKE LARGE-SCALE AND STANDARDIZED TESTS

For students to do their best on large-scale tests, they need to have good test-taking skills for the types of items and format that are used. These skills help familiarize students with the format of the questions and give them strategies for answering the questions. As pointed out in Chapter 4, these "test-taking" skills are important because they help to ensure validity of the inferences that are drawn from the results. That is, you don't want a situation in which students have done poorly, in part, because of a lack of these skills. Here are some important test-taking skills for taking standardized and standard-based tests:

- Read or listen to directions carefully.
- Read or listen to test items carefully.
- Set a pace that will allow adequate time to complete the test.
- Bypass difficult items and come back to them later (do easy items first).
- Make informed guesses, rather than omitting items.

- Eliminate as many options as possible before guessing.
- Follow directions for marking answers carefully.
- Check to be sure that the item number in the booklet matches the item number on the answer sheet.
- Check answers if time permits.
- Review item formats and strategies to get the answer.
- Look for grammatical clues to the right answer.
- Read all answers before selecting one.

It is also best to create an appropriate climate or classroom environment for taking the test. This begins with a teacher's attitude toward the test. If the teachers convey to students that the test is a burden, unnecessary, or even an unfair imposition, students may adopt a similar attitude and may not try as hard as they can to do well. Teachers should impart an attitude of challenge and opportunity. Comments that add pressure or result in pretest jitters, such as saying what will happen if the scores are low, should be avoided. Teachers should emphasize to students that they should try to do their best and that this effort is more important than receiving a high score. Telling the students that the results are important and will be combined with other information will reduce anxiety, which could severely impair performance. If students appear overly anxious, their behavior should be noted to be included when interpreting the results. Some students may need counseling or other special services if test anxiety is serious.

Because most standardized tests use items that are fairly difficult, prepare students for this level of difficulty so that they are not easily discouraged. Give them practice items and short practice tests that simulate the difficulty of the items. Motivate students by explaining how the results will help them by improving teaching, learning, knowledge of themselves, and planning for the future.

A proper physical environment will support students' best efforts. Students need adequate work space, lighting, and ventilation. The room should be quiet, without distractions, and the test should be scheduled to avoid interruptions, such as school announcements. A sign such as "Testing—Do Not Disturb" should be placed on the door. The seating arrangement in the classroom should minimize distractions and cheating. Visual aids in the room that could help students should be removed. If possible, tests should be scheduled in the morning because students can usually focus better then than in the afternoon. Table 7.4 lists some "do's and don'ts" regarding test preparation.

INTERPRETING LARGE-SCALE AND STANDARDS-BASED TEST SCORES

On a test, students answer questions and get a certain percentage of the items correct. As pointed out in Chapter 6, this is referred to as the student's *raw score.* Although many large-scale tests report raw scores for subscales, the vast majority

Table 7.4 Teacher Do's and Don'ts When Preparing Students for Large-Scale Tests

Do	*Don't*
Teach to the test	Teach the test
Improve students' test-taking skills	Use the standardized test format for classroom tests
Establish a suitable environment	Describe tests as a burden
Motivate students to do their best	Tell students that important decisions will be made solely on the results of a single test
Explain why tests are given and how the results will be used	Use previous forms of the same test to prepare students
Give practice items and tests	Convey a negative attitude about the test
Tell students they probably won't know all the answers	
Tell students not to give up	
Tell students to skip and come back to hard items	
Allay student anxiety	
Have a positive attitude about the test	

Source: Adapted from McMillan, J. H. *Classroom Assessment: Principles and Practice for Effective Instruction* (2nd ed.). Boston: Allyn & Bacon. Copyright © Allyn & Bacon, Incorporated. Used with permission.

of scores that are used are modifications or transformations of the raw scores into *derived scores.* There are two types of derived scores, those that refer to the percentage correct, or *absolute derived score,* and those that are reported as a comparison with how others did on the same assessment, or *relative derived scores.* These two types of scores relate closely to criterion/standards-referenced and norm-referenced interpretations, respectively. We will begin with a discussion of norm-referenced scores—those that indicate relative position.

Norm-Referenced Scores

Percentile Scores

One type of norm-referenced derived score, percentile, was introduced in Chapter 6. As a reminder, percentile scores indicate the percentage of students in the reference group (norm group) who were outperformed. For example, a percentile score of 75 means that the student scored as well as or better than 75 percent of the students in the reference group. This gives meaning to the relative position of the score, but it does not tell much about the degree of difference between the scores. That is, the difference between scores of 80 and 85, in

raw score units, is not the same as the difference between percentile scores of 50 and 55. This property makes it difficult to do much with percentiles beyond a simple description of results. For instance, because of unequal intervals, percentile scores should not be averaged.

Standard Scores

Standard scores are derived scores, transformed from raw scores, that are expressed as units of standard deviation. Each set of scores has a fixed or permanent mean and standard deviation with roughly equivalent units between the same number of scale points (e.g., a 5-point difference in performance means the same thing regardless of the location of the difference on the entire scale). This allows appropriate statistical analyses, such as averaging and comparing groups.

The simplest and most easily calculated standard score is the z score. A z score is expressed as units of standard deviation above or below the mean. The mean of the distribution is 0, and the standard deviation is 1. A z score can be calculated from any raw score as long as the mean and standard deviation of the raw score distribution is known:

$$z \text{ score} = (X - \bar{X})/SD$$

Where
X = any raw score
$\bar{X}$ = raw score distribution mean
SD = raw score distribution standard deviation

For example, the z score for a raw score of 70 in a distribution that has a mean of 80 and a standard deviation of 10 is –1: (70–80)/10. Because the z score distribution has a standard deviation of 1, these scores can easily be transformed to other standard scores that will have only positive values. These are the types of standard scores reported with standardized tests. The new mean and standard deviation unit for the converted scores are usually determined arbitrarily, which can be confusing to the public. Different companies and states will adopt unique scales. Some of these are summarized in Table 7.5.

Grade Equivalent Scores

A grade equivalent (GE) or grade norm score indicates student performance in relation to grade level and months of the school year, assuming a 10-month year. Thus, a GE of 4.6 refers to fourth grade, in the sixth month. As with other norm-referenced tests, GEs indicate a student's standing in relation to the reference group. The number is based on the performance of reference group

Table 7.5 Types of Standard Scores

Standard Score	*Description*
Normal Curve Equivalent (NCE)	Scores range from 1 to 99, with a mean of 50 and a standard deviation of 21.06. NCEs are similar to percentile rank at the mean (50) and ends of the distribution (1 and 99). In between, however, NCEs are not equivalent to percentile.
T-score	Distribution with a mean of 50 and a standard deviation of 10.
Stanine	Scores that correspond to nine areas of the normal curve. The mean of the distribution is 5, with a standard deviation of about 2. There is a different percentage of scores in the range identified by each stanine.
Standard Age Score (also IQ scores)	Widely used with ability tests. The distribution has a mean of 100 and a standard deviation of 15 or 16.
Developmental Standard Score (growth score)	Allows year-to-year comparison of progress with a scale that is continuous across many grade levels (e.g., mean score at second grade is 175, mean score at fourth grade is 200, and mean score at sixth grade is 225).
SAT	Distribution with a mean of 500 and a standard deviation of 100 for quantitative and verbal tests.

students on the test at different grade levels. If the reference group of students in grade 4 achieved a mean score of 32 items correct, then any student who also got 32 items correct would receive a GE of 4.6. For a GE of 4.0, students would need to correctly answer less items, say 28.

GE scores appear to be easily interpreted because of the common sense referent to grade level (year and month in school). There are important limitations to these types of scores, however. Consider Tom, a third grader who obtained a GE score in mathematics of 4.7 at the beginning of the school year. Does this mean that Tom should be promoted to fourth grade, or that he could do as well as fourth graders, or that he is performing "above" grade level? The answer is no to each question. It can be said with confidence, however, that Tom has performed about the same on the test as students in the norming group who are in the seventh month of the fourth grade. Compared with other third graders in the reference group, Tom is above average, but his score does not imply that he would be successful with fourth-grade material or should be promoted to fourth grade.

GE scores are helpful in explaining student strengths and weaknesses. For example, consider Jennifer's scores in three areas during a two-year period that includes fourth and fifth grades:

	4th Grade	5th Grade
Reading comprehension	5.6	5.8
Language	5.2	6.7
Mathematics	7.7	8.6

It is clear that, relatively speaking, Jennifer is much stronger in mathematics than in either reading comprehension or language, that there is little difference between reading comprehension and language, and that her overall performance is above average when compared with the reference group. It is also clear that she is making good progress in language and mathematics but not in reading comprehension.

Although GE scores can be helpful in identifying strengths and weaknesses, and in examining growth, a number of important cautions limit what can be concluded from the scores.

1. It is incorrect to interpret the GE as the grade level at which the student is performing. That is, a GE of 7.2 for a fifth grader does not mean that the student is performing at the seventh-grade level; it means that the student is performing the same as a typical seventh grader taking the same test.

2. Most GE scores are extrapolated beyond and interpolated between the actual data provided by students in the reference group. For example, a sixth-grade test may be given to a sample of sixth graders only during the 2nd and 10th month of the year, yet scores are calculated to correspond to every month of the school year between these times (interpolated) and also calculated to correspond to GE scores prior to and after sixth grade (extrapolated). This means that many GE scores are only estimates of student performance.

3. A unit of 1 GE should not be the standard by which progress is evaluated. This assumes uniform growth throughout a year, when in reality, students learn at different rates.

4. GE scores do not indicate at what grade level students should be placed. Grade placement depends on local objectives and the performance of all students in the school. A third grader who scores a GE of 5.0 on a test shows strong mastery of the material, but this does not mean that skipping a grade would be appropriate.

5. Because GE scores are based on the normal distribution, half the students in the reference group are expected to be above the score and half below it. Expecting all students to be above grade level may not be consistent with the achievement levels of the students or local conditions. Above-average students (in comparison with the reference group) would be

expected to achieve a GE score that places them "above" grade level, whereas a below-average class might be expected to simply reach the GE score that is consistent with their grade level.

6. Extremely high or low GE scores are problematic because of a lack of reliability for such scores and their heavy dependence on extrapolation.
7. The methods used to establish GE scores tend to exaggerate the importance of small differences in the number of items answered correctly.
8. GE scores from different tests cannot be compared because different tests of the same or similar content do not measure the same thing and because different reference groups will influence relative standing. Significant comparisons can be made only within the same test battery.

Standards-Based Test Scores

The second way scores are reported on large-scale assessments is in relation to some standard, criterion, or level of performance. With NCLB, all states have developed learning standards or objectives and assess the extent to which students demonstrate competency that meets the standards and objectives; standards-based scores and criterion-referenced interpretations are common.

The basis for interpreting standards-based scores is the number of items answered correctly or the judgment of an expert who reviews a sample of student work, such as a writing sample. The raw score or expert judgment is used to determine placement into two or more categories, such as the following:

- Pass/fail
- Meets/fails to meet
- Advanced, proficient, basic, novice
- Not proficient, proficient, advanced
- Minimal, partial, satisfactory, extended
- No attempt, inadequate, satisfactory, competent, exemplary

The score that is reported corresponds to such categories, so meaning is directly dependent on what is meant by terms such as *pass* and *proficient* and *advanced*. Accurately interpreting the scores, then, requires understanding how the standards (or benchmarks) were set and what is meant by each level. Let's consider the most basic type of standard-based score—pass or fail. Suppose a school district has identified a set of fifth-grade mathematics competencies that must be demonstrated for students to obtain a passing score on the end-of-year fifth-grade mathematics test. Once the competencies have been identified, test items would need to be generated to measure the competencies. Suppose 20 questions are developed to measure each competency. How many of the 20 items would a student need to answer correctly to be judged competent? Seventy-five percent? Fifteen correct of 20? Half the items? Twelve items? Seventeen items? All 20 items?

The determination of the standard involves making a judgment about the number of items that need to be answered correctly to classify the score as "pass." Who makes the judgment? Typically, such judgments are made by experts in the content area, though final determination of "cut scores" for state tests is made by state boards of education.

Experts and policy makers review the items and make decisions about whether students with competence in the area assessed would be able to answer the items. Of course, there is some variation in the judgments of different individuals, so usually there is some type of averaging of the judgments of many individuals. Table 7.6 illustrates how many items must be answered correctly to reach proficient and advanced levels in the Virginia assessment program for several subjects in Grades 3 and 8.

A common approach to standards-based scores is to use a scale to report different levels of competency. These scales are arbitrary and often are unique to a given test or state assessment program. In Virginia, for instance, scores on statewide competency tests, which are standards based, are reported using a scale of 0 to 600, with a score of 400 indicating proficient. The score indicating proficient stays the same for all content areas, although the number of items answered correctly is different.

Standards-based scores, then, depend not only on how well students do but also on the nature of the judgments made by those setting the standards. Accurate interpretations can be made only after inspecting the items and descriptions of what words such as *proficient* and *advanced* mean and by knowing about the individuals who set the standards.

To promote accurate interpretations, large-scale test developers release sample items and examples of student work that have been judged. These items can be reviewed to give some idea of the level of performance needed.

Table 7.6 Cut Scores Established in 2006 for the Virginia Assessment Program

Test	*Pass/Proficient*	*Pass/Advanced*
Grade 3		
English: Reading	23 out of 35 items	31 out of 35 items
Mathematics	35 out of 50 items	45 out of 50 items
Grade 4		
English: Reading	23 out of 35 items	31 out of 35 items
Mathematics	35 out of 50 items	43 out of 50 items
Grade 8		
English: Reading	28 out of 45 items	40 out of 45 items
Mathematics	32 out of 50 items	42 out of 50 items

Source: Virginia Department of Education (2007).

The more removed the test is from a local setting, the more likely it is that those setting the standards will bring perspectives and values to that process that are inconsistent with local perspectives and values. National-level tests, such as the NAEP, are further from the classroom than a state test, and a state test is further from the classroom than a district test. Thus, those setting the standards on a national or state test are much less informed about local curriculum and values than those in the district who set standards only for schools and students in that district.

Difficulty of items is an important factor in standard setting, which is why sound inferences depend on knowledge of the items. There can be great variability in the difficulty of items that measure the same competency, objective, or standard. Given this variability, just knowing that a student answered 70 percent of the items correct is insufficient. You also need to know if these were hard or easy items. Obviously, getting 70 percent correct with easy items means something different from 70 percent with difficult items. One approach to judging the nature of the standard is to compare scores from the assessment with other performances of the students. This is essentially a check on the validity of the inferences drawn from the scores. It can provide a type of anchor for interpretation. For instance, suppose your brightest and highest-performing students don't "pass" the test. These are students who have demonstrated strong achievement in similar areas in class, yet they fail to show adequate performance. Because there is good evidence that they know the skills, it may be best to explore student motivation to do well on the test and to examine the test specifications and items in greater detail to determine a reasonable explanation for the discrepancy. It may be that what you thought was adequate knowledge and skill was not, or that the test is assessing areas you did not emphasize with your students.

As previously mentioned, many norm-referenced standardized tests purport to provide criterion-referenced information. Be wary of using norm-referenced tests in this way. It is best to use norm-referenced tests for what they are designed to do—show comparisons with other students, identify strengths and weaknesses, and show growth. To make sound decisions about whether students have obtained specific knowledge and skills, criterion-referenced assessments, which are designed for that purpose, are best.

One additional and important limitation to some standardized assessments, both norm- and criterion/standards-referenced, is the common use of a multiple-choice item format that allows machine scoring of student responses. This format, along with the need to be broad in coverage, results in the measurement of mostly low-level skills and knowledge (Marzano & Kendall, 1996). Such tests tend to assess isolated facts and only rudimentary understanding. Students select, rather than produce, a response on these types of tests. The multiple-choice format also gives the false impression that there is a right and wrong answer for all questions.

One of the dilemmas of standards-based large-scale assessment is that if items are constructed to require application, analysis, synthesis, and other reasoning skills, the tests measure general ability in addition to knowledge and

understanding of the content area. In Virginia, for example, the scores on the Grade 8 English tests are strongly correlated to scores on both science and mathematics. Although incorporating reasoning skills with content may be desirable from one perspective, it complicates interpretation of the results from these assessments. If a student obtains a low score on a mathematics test that correlates strongly with performance in English, is the correct conclusion that the student is weak in mathematics or weak in applying math skills to these types of tests that require competence in reading comprehension to understand the question? Do low scores mean more work is needed in mathematics, in doing the types of items that are on the test, or in reading comprehension? As I have already stressed, the scores from these tests make most sense when interpreted in light of other indicators of student performance.

INTERPRETING LARGE-SCALE AND STANDARDIZED TEST REPORTS

Standardized Achievement Tests

Many types of reports can be produced from standardized achievement tests, including reports for parents, individual students, classes, schools, and school districts. Because the reports are designed to provide as much information as possible on a single page, they may appear complicated and difficult to understand. There is typically a large number of different scores, and often graphs are provided. For a comprehensive battery of tests, scores are usually reported for each skill as well as each subskill. A good approach to understanding the reports is to first consult the test manual and/or interpretation guide to find examples of explanations of actual scores. The manual or interpretation guide is also important for understanding the meanings of the labels used for the skills (e.g., math computation, measurement, and language). Most publishers of large-scale standardized assessments do a good job of explaining what each part of the report means.

Each test publisher has a unique format for reporting results and usually has unique types of scores. Different formats are used to summarize the scores. A single report may include a listing of all students in a class, the class as a whole, a skills analysis for the class or individual students, individual profiles, growth charts, and other formats. Some reports will include scores for only major content areas, whereas others will include subscale scores or even results for individual items. Also, different types of norms may be used. All this means that each report contains different information, organized and presented in unique formats. So the first step in understanding a report is to identify the nature of the information presented, then find an explanation for it in an interpretive guide.

Two examples of state-level standards-based reports are shown in Figures 7.1 and 7.2. The Virginia Student Performance Report shows scaled scores that are unique to these tests and a proficiency level summary for each of five major

content areas. The reporting category scaled scores are also reported, ranging for Elizabeth Tomlinson (fictitious) from 31 to 44. The reporting category scores are not tied to proficiency levels and provide only a general idea of student strengths and weaknesses and a general indication of performance compared with other students statewide (35 is the state average). The reporting category scores do not add up to the total test score.

The Virginia School Summary Report (Figure 7.2) summarizes the number and percentage of students obtaining each of the four ratings in three writing domains. The mean scale scores for the test and reporting categories are indicated, as well as the number and percentage of students in the school who are categorized as fail/does not meet, pass/proficient, or pass/advanced. For this testing period, 38 percent of the students obtained a passing score. The written expression and usage mechanics scores show that these areas contribute in approximate equal amounts to the total score (62 percent show consistent or reasonable control for written expression; 65 percent show consistent or reasonable control in usage mechanics).

USING LARGE-SCALE AND STANDARDIZED TEST RESULTS TO IMPROVE INSTRUCTION

The results of large-scale testing can be used for planning prior to instruction and as a way to evaluate the effectiveness of instruction after content and skills have been taught. Any use of standardized scores should be done with the understanding that the results provide only one of many sources of information. Large-scale test results should always be interpreted in the context of other evidence provided by classroom assessments and teacher observation. It is also important to understand the specific nature of the content or skills that are assessed.

Prior to instruction, results from standardized tests may provide a good indication of the general ability level of the students in the class. This information can be used to help establish reasonable, realistic expectations for students and to influence the nature of instructional materials. Expectations should not be fatalistically low or unreasonably high. If the results from a reading readiness test indicate that the class is lacking in ability to read, then they, should not be the sole or even major determinant of instructional practices.

The scores in various subtests can be compared to identify strengths and weaknesses, which can help determine the amount of instruction to give in different areas. Students whose achievement is much lower than what might be expected on the basis of ability testing may need further testing, special attention, or counseling. What constitutes "much lower"? Generally, a discrepancy of 10 percentile points may be sufficient. If the percentile "bands" for achievement, which show the probable range of actual student knowledge or skill, do not overlap with the ability bands, then a significant discrepancy is identified.

Norm-referenced standardized tests are useful for selection and placement into special programs. This occurs at both ends of the achievement/aptitude distribution. Students are selected to receive special services in part on the basis

Student ID#s are optional.

Virginia

Standards of Learning Assessments

STUDENT PERFORMANCE REPORT
GRADE 5 TESTS

STUDENT NAME: ELIZABETH TOMLINSON
DOB: 10/24/85
GENDER: FEMALE
ID# 043156327690
ETHNICITY: WHITE
CLASS: M. SMITH
SCHOOL: LAKESIDE ELEMENTARY - 5678
DIVISION: NEWTOWN - 123

GRADE: 05
TEST DATE: SPRING 1998

TEST REPORTING CATEGORIES	# of ITEMS BLANK& MULTIPLE MARKED	SCALED SCORE	PROFICIENCY LEVEL SUMMARY
English: Reading/Literature and Research	0	433	PASS/PROFICIENT
Use word analysis strategies.	0	31	
Understand a variety of printed materials/resource materials.	0	35	
Understand elements of literature.	0	39	
Mathematics	0	406	PASS/PROFICIENT
Number and Number Sense	0	37	
Computation and Estimation	0	37	
Measurement and Geometry	0	39	
Probability and Statistics	0	33	
Patterns, Functions, and Algebra	0	35	
History and Social Science	0	361	
History	0	33	
Geography	0	33	
Economics	0	33	
Civics	0	33	
Science	0	400	
Scientific Investigation	0	32	
Force, Motion, Energy, and Matter	0	43	
Life Processes and Living Systems	0	34	
Earth/Space Systems and Cycles	0	30	
Computer/ Technology	0	484	
Basic Understanding of Computer Technology	0	43	
Basic Operational Skills	0	39	
Using Technology to Solve Problems	0	44	

Notes:

This *Student Performance Report* (SPR) displays scores for a student on all tests and their reporting categories except for *English: Writing*, which is reported on a separate SPR. The SPR shows, for each SOL test, the

- number of items to which the student did not respond (BLANK) and those where the student marked more than one answer (MULTIPLE-MARKED),
- scaled scores earned by the student on each test as a whole and its reporting categories, and
- proficiency level attained by the student (Pass/Advanced, Pass/Proficient, Fail/Does Not Meet).

Figure 7.1 Example of State Standards-Based Test Individual Report

Source: Virginia Department of Education (2007).

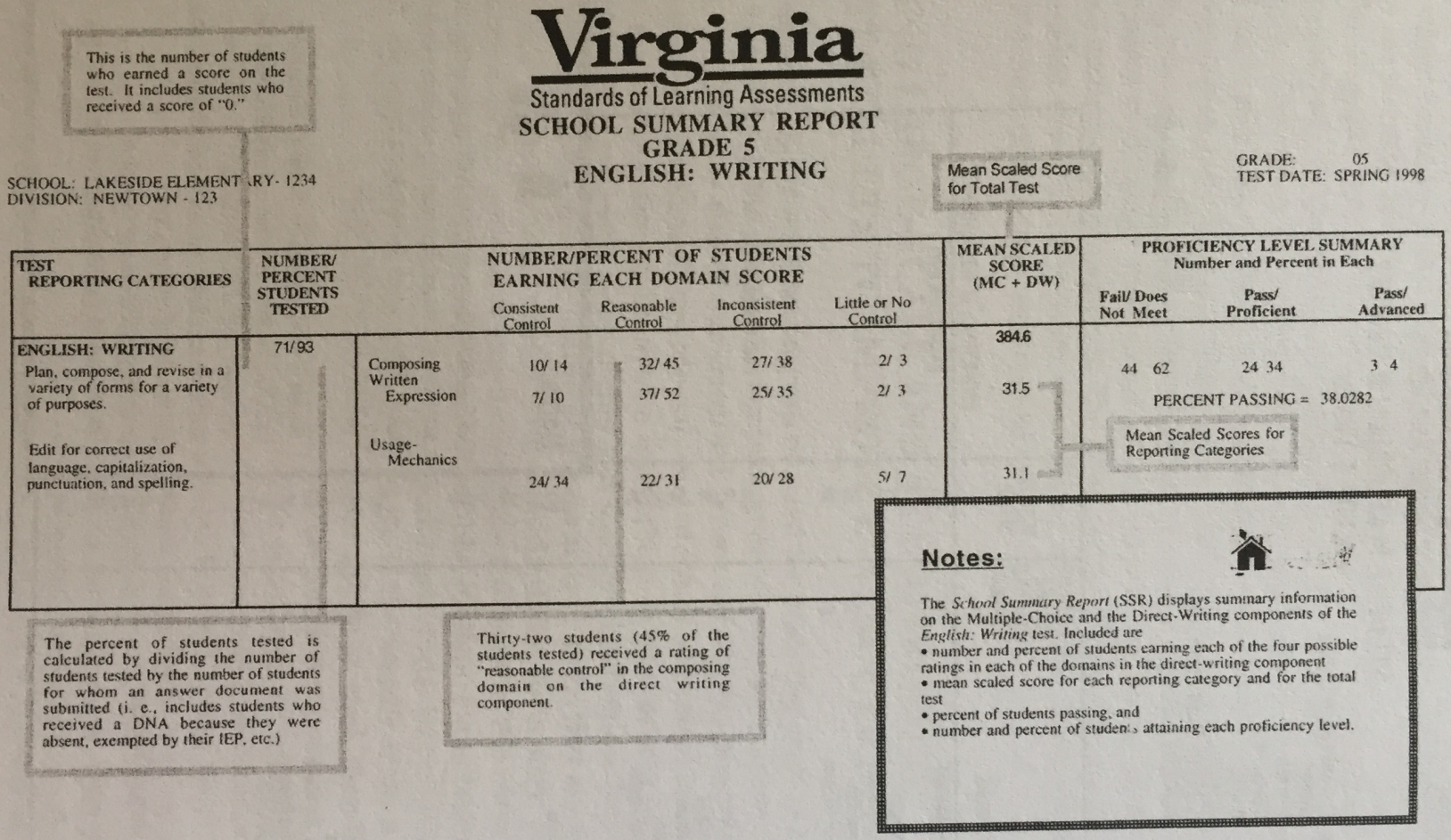

Figure 7.2 Example of State Standards-Based Test School Report

Source: Virginia Department of Education (2007).

of standardized test results; other students are placed in "gifted" programs or other advanced programs because they are identified as the brightest or most knowledgeable.

Large-scale tests given at the end of instruction, like external accountability tests for NCLB, can be used to evaluate the effectiveness of instruction and curriculum. It is expected that students should score well in areas that have been stressed in the instructional program. If not, the scores act like a temperature gauge, indicating that further information needs to be gathered, such as other test performance data and a review of the match between specific test items and subscales and what was taught. When large-scale test data can be gathered through several years, it is possible to evaluate programs by examining trends in areas that have been emphasized and in areas that have not been the focus of instruction.

When students are tested each year, the scores can be used to indicate areas within the curriculum that need further attention. If an area shows a consistent pattern of low scores despite the area being stressed in the classroom, the specific methods of teaching may need to be examined.

Perhaps the most serious misuse of large-scale tests is to evaluate teachers. This is a misuse of the results for several reasons: (a) Standardized tests are not designed to evaluate teaching or teachers; (b) the content and skills tested will not have a perfect match with local curriculum or what individual teachers stress in the classroom; (c) each year brings a unique group of students to a teacher, with knowledge, skills, motivation, and group chemistry that may be different from other years; (d) it is difficult, if not impossible, to isolate the influence of differences between teachers (most of what students experience is common); and (e) a large-scale test provides only one indication of student performance.

In summary, most teachers will find large-scale test results helpful *as long as the scores are used with a full understanding of their limitations and as a supplement to data gathered directly from students day to day.* Often, these tests simply corroborate what teachers already know, but sometimes new information is provided that can have a positive influence on teaching practices. On a larger scale, to comprehensively evaluate programs and schools, standardized tests have value because of their technical soundness and their ability to identify strengths and weaknesses, show gains from year to year, and compare programs to establish effectiveness.