These days, standardized tests (especially standardized *achievement* tests) are often misused. In Chapter 15 you'll learn why it is unwise to try to evaluate educators' effectiveness based on students' scores on many such tests. Nonetheless, standardized tests *do* have an important, educationally useful role to play. This chapter focuses on the appropriate uses of standardized tests.

The chapter will conclude with a consideration of two aptitude tests that have a whopping impact on students' lives. You'll be looking at the two examinations that often function as determiners of a student's continued education. That's right, we'll be dealing with the two most widely used college entrance exams: the SAT and the ACT.

If you're a current or would-be secondary school teacher, you'll immediately recognize the need for you to know about the SAT and the ACT. After all, many of your students will soon be taking one or both of these tests. But if you're an elementary school teacher, or are preparing to be one, you might be thinking, "What do college entrance exams have to do with me?" Well, all teachers—elementary and secondary—ought to understand at least the most significant facts about these two examinations. Indeed, even teachers of primary-grade kids will find parents asking questions such as, "How can we get our child really ready for those important college entrance tests when the time comes?" Every teacher should be familiar with any truly significant information associated with the teaching profession. Thus, because most students and their parents will, at some point, want to know more about college entrance exams, it seems that every single teacher—and even married ones!—should know the basic SAT and ACT facts you'll learn about in this chapter.

Standardized Tests

A ***standardized test*** is a test that is designed to yield either norm-referenced or criterion-referenced inferences and that is administered, scored, and interpreted in a standard, predetermined manner. Almost all *nationally* standardized tests are distributed by commercial testing firms. Most such firms are for-profit corporations, although there are a few not-for-profit measurement organizations, such as the Educational Testing Service (ETS), that distribute nationally standardized tests. Almost all nationally standardized tests, whether they're focused on the measurement of students' aptitude or of their achievement, are chiefly intended to provide norm-referenced interpretations.

Standardized achievement tests have also been developed in a number of states under the auspices of state departments of education. In some instances, these state-chosen standardized tests may have been acquired from a confederation of states, for instance, the states affiliated with the Smarter Balanced Assessment Consortium (SBAC)—or even from smaller consortia of only a few collaborating states. These statewide tests (clearly intended to be administered, scored, and interpreted in a standard, predetermined fashion) have usually been

installed to satisfy a legislative mandate that establishes an educational accountability program of some sort. In certain instances, important decisions about individual students are made on the basis of a student's test performance. In many states, for example, if a student does not pass a prescribed statewide basic-skills examination by the end of high school, the student is not awarded a state-sanctioned diploma—even though the student has satisfied all other curricular requirements. In other instances, although no contingencies for individual students depend on how a student performed on a test, the results of student tests are publicized by the media on a district-by-district or school-by-school basis. The test results thus serve as an indicator of local educators' effectiveness, at least in the perception of many citizens. These state-sired standardized *achievement* tests are generally intended to yield criterion-referenced interpretations. Educational *aptitude* tests are rarely, if ever, developed by state departments of education.

Although standardized tests have traditionally consisted almost exclusively of selected-response items, in recent years the developers of standardized tests have attempted to incorporate a certain number of constructed-response items in their tests. Standardized tests, because they are intended for widespread use, are developed with far more care (and cost) than is possible in an individual teacher's classroom. Even so, the fundamentals of test development that you've learned about in earlier chapters are routinely employed when standardized tests are developed. In other words, the people who create the items for such tests attempt to adhere to the same kinds of item-writing and item-improvement precepts that you've learned about. The writers of multiple-choice items for standardized tests worry, just as you should, about inadvertently supplying students with clues that

Decision Time

Which Test to Believe

Each spring in the Big Valley Unified School District, students in grades 5, 8, 10, and 12 complete nationally standardized achievement tests in reading and mathematics, as well as a nationally standardized test described by its publishers as "a test of the student's cognitive aptitude." Because William White teaches eighth-grade students in his English classes, he is given the task of answering any questions his eighth-graders' parents about the

...ent question

the achievement test. For examp
Mrs. Wilkins (Wanda's parents)
this: "If Wanda scored at the 90
aptitude test and only at the 65
achievement test, does that m
enough study time? Putting it
continued, "should we really b
test's results or the achievem
Which is Wanda's 'true' test

If you were William an
to answer the questic
parents, what would

give away the correct answer. The writers of short-answer items for standardized tests try to avoid, just as you should, the inclusion of ambiguous language anywhere in their items.

There are, of course, substantial differences in the level of effort associated with the construction of standardized tests and the construction of classroom tests. A commercial testing agency may assign a flotilla of item writers and a fleet of item editors to a new test-development project, whereas teachers are fortunate if they have a part-time aide or, possibly, a malleable spouse to proofread teacher-made tests to detect typographical errors.

Group-Focused Test Interpretation

Although the bulk of this chapter will be devoted to a consideration of score-reporting mechanisms used to describe an individual student's performance, you'll sometimes find you need to describe the performance of your students as a group. To do so, you'll typically compute some index of the group of scores' *central tendency*, such as when you determine the group's ***mean*** or ***median*** performance. For example, you might calculate the *mean raw score* or the *median raw score* for your students. A ***raw score*** is simply the number of items that a student has answered correctly. The *mean*, as you probably know, is the arithmetic average of a set of scores. For example, the mean of the scores 10, 10, 9, 8, 7, 6, 3, 3, and 2 is 6.4 (found by summing the nine scores and then dividing that sum by 9). The *median* is the midpoint of a set of scores. For the nine scores in the previous example, the median is 7 because this score divides the group into two equal parts. Means and medians are useful ways to describe the point at which the scores in a set of scores are centered.

In addition to describing the central tendency of a set of scores (via the mean and/or median), it is also helpful to describe the *variability* of the scores—that is, how spread out the scores are. One simple measure of the variability of a set of students' scores is the ***range***. The range is calculated by simply subtracting the lowest student's score from the highest student's score. To illustrate, suppose the highest test score by students in your class was 49 correct out of 50, earned by Hortense (she always tops your tests; it is surprising she missed one item). Suppose further that the lowest score of 14 correct was, as usual, earned by Ed. The range of this set of scores would be 35—that is, Hortense's 49 minus Ed's 14.

Because only two scores influence the range, it is less frequently used as an index of test-score variability than is the standard deviation. A ***standard deviation*** is a kind of average. More accurately, it's the average difference between the individual scores in a group of scores and the mean of that set of scores. The larger the size of the standard deviation, the more spread out are the scores in the ***distribution***. (That's a posh term to describe a set of scores.) Here is the formula for computing a standard deviation for a set of scores:

$$\text{Standard Deviation } (\{S.D.\}) = \sqrt{\frac{\Sigma(X - M)^2}{N}}$$

where $\Sigma(X - M)^2$ = the sum of the squared raw scores (X) − the mean (M)
N = the number of scores in the distribution

Here's a step-by-step description of how you compute a standard deviation using this formula. First, compute the mean of the set of scores. Second, subtract the mean from each score in the distribution. (Roughly half of the resulting values, called *deviation score*, will have positive values, and roughly half will have negative values.) Third, square each of these deviation scores. This will make them all positive. Fourth, add the squared deviation scores together. Fifth, divide the resulting sum by the number of scores in the distribution. Sixth, and last, take the square root of the results of the division you did in the fifth step. The square root that you get is the standard deviation. Not too tough, right?

To illustrate the point that larger standard deviations represent more spread in a distribution of scores than smaller standard deviations, compare the two fictitious sets of scores on a 10-item short-answer test presented in Figure 13.1. Both sets of scores have a mean of 5.0. The distribution of scores at the left is much more homogeneous (less spread out) than the distribution of scores at the right. Note that the standard deviation for the more homogeneous scores is only 1.1, whereas the standard deviation for the more heterogeneous scores is 3.2. The larger the standard deviation, therefore, the more distant, on average, will be the distribution's scores from the distribution's mean.

You may have occasions to describe the scores of an entire group of the students you teach. Those descriptions might portray your students' performances on standardized tests or on teacher-made tests. If you get at all comfortable with means and standard deviations, those two indices usually provide a better picture of a score distribution than do the median and range. But if you think means and standard deviations are statistical gibberish, then go for the median (midpoints are easy to identify) and range (ranges require only competence in subtraction). With group-based interpretations out of the way, let's turn now to interpreting individual students' scores from the kinds of standardized tests commonly used in education.

Figure 13.1 Two Fictitious Sets of Tests Scored with Equal Means but Different Standard Deviations

More Homogeneous Scores:	*More Heterogeneous Scores:*
3, 4, 4, 5, 5, 5, 5, 6, 6, 7	0, 1, 2, 4, 5, 5, 6, 8, 9, 10
Mean = 5.0	Mean = 5.0

Individual Student Test Interpretation

Two overriding frameworks are generally used to interpret students' test scores. Test scores are interpreted in *absolute* or *relative* terms. When we interpret a student's test score *absolutely*, we infer from the score what it is that the student can or cannot do. For example, based on a student's performance on test items dealing with mathematics computation skills, we make an inference about the degree to which the student has mastered the array of computation skills represented by the test items. The teacher may even boil the interpretation down to a dichotomy—namely, whether the student should be classified as having mastered or as not having mastered the skill or knowledge being assessed. A mastery-versus-nonmastery interpretation represents an *absolute* interpretation of a student's test score. Classroom teachers often use this absolute interpretive approach when creating tests to assess a student's knowledge or skills based on a particular unit of study.

When we interpret a student's test score *relatively*, we infer from the score how the student stacks up against other students who are currently taking the test or who have already taken the test. For example, when we say that Johnny's test score is "above average" or "below average," we are making a relative test interpretation because we use the average performance of other students to make sense of Johnny's test score.

As pointed out earlier, this chapter focuses on how teachers and parents can interpret scores on standardized tests. Because almost all standardized test scores require relative interpretations, the three interpretive schemes to be considered in the chapter are all relative score-interpretation schemes. The vast majority of standardized tests, whether achievement tests or aptitude tests, provide relative interpretations. Accordingly, teachers need to be especially knowledgeable about relative score-interpretation schemes.

Percentiles

The first interpretive scheme we'll consider, and by all odds the most commonly used one, is based on ***percentiles***, or as they are sometimes called, *percentile ranks*. Percentiles are used most frequently in describing standardized test scores, because percentiles are readily understandable to most people.

A percentile compares a student's score with those of other students in a *norm group*. A student's percentile indicates the percent of students in the norm group that the student outperformed. A percentile of 60, for example, indicates that the student performed better than 60 percent of the students in the norm group.

Let's spend a moment describing what a norm group is. As indicated, a percentile compares a student's score with scores earned by those in a norm group. This comparison with the norm group is based on the performances of a group of individuals who have already been administered a given examination. For instance, before developers of a new standardized test publish their test, they

usually administer the test to a large number of students who then become the norm group for the test. Typically, different norm groups of students are assembled for all the grade levels for which percentile interpretations are made.

Figure 13.2 shows a graphical depiction of a set of 3000 students' scores such as might have been gathered during the norming of a nationally standardized achievement test. Remember, we refer to such students as the norm group. The area under the curved line represents the number of students who earned scores at that point on the baseline. You'll notice that for a typical norm group's performance, most students score in the middle, and only a few students earn very high or very low scores.

In fact, *if* the distribution of test scores in the norm group is *perfectly* normal, then, as you see in Figure 13.3, over two-thirds of the scores (represented by the

Figure 13.2 A Typical Norm Group

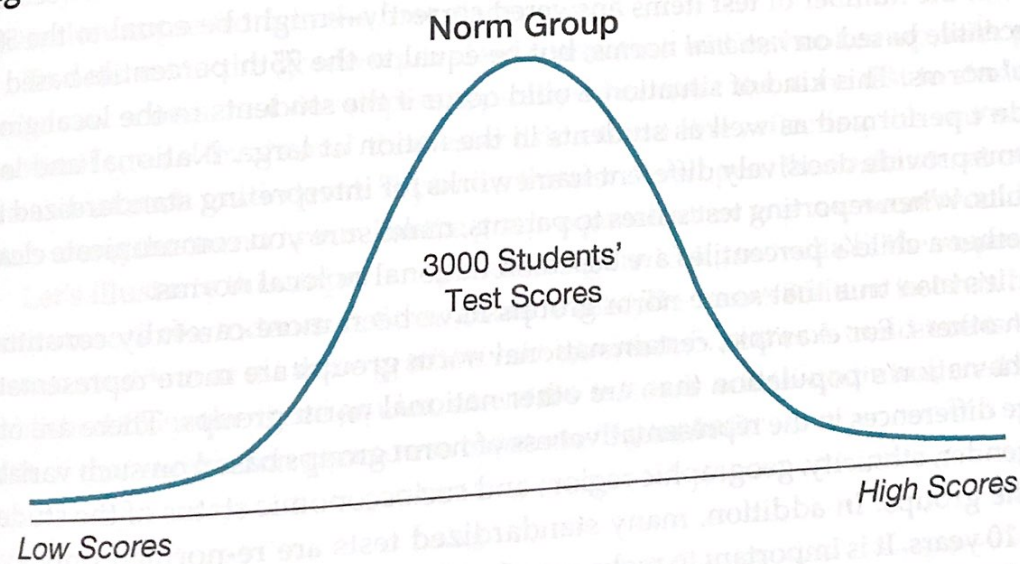

Figure 13.3 A Normally Distributed Set of Norm-Group Scores