CHAPTER 1

AN INTRODUCTION TO CONTEMPORARY EDUCATIONAL TESTING AND MEASUREMENT

LEARNING OUTCOMES

After completing this chapter, the student will be able to:

1. Explain why tests are only tools that can be used appropriately or inappropriately.
2. Explain why test scores are fallible.
3. Distinguish between formative, interim or benchmark, and summative assessment.
4. Describe how technical adequacy, test user competency, and the extent to which a test aligns with its intended purpose can limit or enhance a test's usefulness.
5. Describe how educational test usefulness is affected by diversity considerations.
6. Explain why a single test score alone should not be used for important educational decision making.
7. Compare and contrast testing, assessment, and the assessment process.
8. Compare and contrast the various types of educational tests and assessments.
9. Explain why testing and assessment skills are vital to today’s classroom teacher.

CHANCES ARE that some of your strongest childhood and adolescent memories include taking tests in school. More recently, you probably remember taking a great number of tests in college. If your experiences are like those of most of the students who come through our educational system, you probably have very strong or mixed feelings about tests and testing. Indeed, some of you may swear that you will never test your students when you become teachers, unless of course you are required by law to do so! If so, you may think that test results add little to the educational process and fail to reflect learning, that testing may turn off students, or that tests do not measure what they are supposed to measure. On the other hand, others may believe that tests are necessary and vital to the educational process, and are critical for accountability purposes. For you, tests may represent irrefutable evidence that learning has occurred and that goals and objectives are being met. Rather than viewing tests as deterrents that turn off students, you may see them as motivators that stimulate students to study and provide both students and decision makers with vital and objective feedback about student achievement.
Between those who feel positively about tests and those who feel negatively about them lies a third group. Within this group, which includes the authors of this textbook, are those who see tests as tools that can make important contributions to the process of evaluating pupils, curricula, and teaching methods, but who question the status and power too often given to individual tests and test scores. We are concerned that test users and consumers of test results (e.g., teachers, parents, the media, administrators, policy makers, and other decision makers) often uncritically accept test scores without careful consideration of how useful the test scores may actually be for whatever decision may be at hand.

Whatever your feelings may be about educational testing, it is critical that prospective teachers be aware of important recent developments that have altered significantly the testing landscape at national, state, district, and local levels. In Chapter 2, we will review these developments to give you an historical context that we hope will enable you to understand how today’s testing picture has developed over the last several decades. Before we review those developments, we will clarify what our perspective is about the utility of educational tests, assessment, and the broader assessment process in this chapter. We then discuss briefly the specific and general purposes of educational testing, and the various types of educational tests that exist. We conclude this chapter with a discussion of the impact of educational testing on the general education classroom teacher.

TESTS ARE ONLY TOOLS; THEIR USEFULNESS CAN VARY

One of the driving reasons we originally decided to write this textbook was the concern we had at that time about some common misconceptions. These included the ideas that many teachers, administrators, parents, and other decision makers had about the utility of tests and test scores, and especially the uncritical acceptance of individual test scores for important educational decision making. Uncritical acceptance of test scores by decision makers concerned us for five reasons. First, tests are only tools, and tools can be appropriately used, unintentionally misused, and intentionally abused. Second, tests, like other tools, can be well designed or poorly designed. Third, both poorly designed and well-designed tools in the hands of ill-trained or inexperienced users can be dangerous. Fourth, the usefulness of a well-designed tool, even in the hands of a competent user, can be limited if the tool, or test, is used for an unintended purpose or population. In other words, just as there is no “one-size-fits-all” tool (not even the venerable Swiss army knife!), no single test is appropriate for all purposes and all persons. Fifth, even when a test is well designed and is appropriately used by a competent examiner (i.e., for the purpose and populations it was designed for), the test can only provide us with some of the information we may want or need to make the best possible educational decision about a student.

“Wait a minute!” you may say. “All this makes it sound like you are saying that tests are not useful for educational decision making, even if they are well constructed and properly used.” Not so! We are not saying test results are useless, unimportant, or unhelpful. We are saying that it is important to recognize that the usefulness of tests, like the usefulness of all tools, depends on a variety of factors. Let us explore some of these factors next.
WHY WE DEVELOPED THIS TEXT: ENHANCING TEST USEFULNESS

By helping you learn to design and use tests and test results appropriately, we hope you will be less likely to misuse tests and their results and be better able to recognize and avoid using poorly designed tests with questionable technical accuracy. We also hope that you will become mindful of how the purpose of testing and the population to be tested can affect a test’s usefulness. Finally, we hope that you will grasp the importance of considering multiple sources of information obtained from multiple informants along with test results to make important educational decisions about students. Let us turn to a more detailed explanation of how each of these points can affect the usefulness of a test for educational decision making.

Technical Adequacy

A critically important factor that affects a test’s usefulness is its technical adequacy. Today’s teachers must not only administer tests developed or required by their state or federal requirements (e.g., the No Child Left Behind Act, or NCLB), but also develop their own tests to inform day-to-day instructional decision making. Much of this text is devoted to helping you develop teacher-constructed (or teacher-made) tests with good technical adequacy and to help you understand how to use the results of teacher-made tests to evaluate instructional effectiveness and to provide data to help teachers complement and adjust instruction. We also intend to help you learn to evaluate the technical adequacy of commercial or state-mandated tests (i.e., developed by test publishers) required for accountability purposes so that you know how to use the results from those tests, when appropriate, to evaluate and inform instruction and educational decision making.

The technical adequacy of a test includes evidence of its validity (see Chapter 15) and its score reliability (see Chapter 16). Validity evidence helps us determine whether the test is measuring what it is intended to measure, and score reliability indicates the extent to which test scores are consistent and stable. In general, we strive to use tests with the strongest validity and score reliability evidence. However, contrary to common wisdom, these factors are not fixed characteristics of a test, even if the test is well established, widely used, and respected. This is because a test’s validity and score reliability can be affected by many factors, including the competency of the test user, whether the test is being used for the purpose it was developed, the person or population it is used with, and even the testing conditions (e.g., noisy rooms, poor lighting, and timing errors) (see Chapters 15–19).

This is why we said before that no test, even one with well-established reliability and validity, is a “one-size-fits-all” test that is equally useful for all test users, purposes, and populations. Thus, it is inappropriate to speak of the “validity of a test” or the “reliability of a test,” as though validity and reliability are permanent, unchanging characteristics of the test. Nevertheless, this is exactly what many test users believe. Because test usefulness can vary, it is most appropriate to speak of the evidence of a test’s validity and score reliability for a particular use and with a particular population, when administered by a competent test user. The need to consider test user competency, a test’s intended use, and the intended population when discussing the test’s technical adequacy and usefulness emerged from deliberations among measurement experts from the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) over several years when they developed the third edition of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999). Often referred to as the Standards, the importance of those factors...
was reinforced when the fourth edition was released (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 2014). The Standards are widely regarded as one of the most authoritative and influential guidelines in the testing, measurement, and assessment arena. Next, let us consider several examples that should help you understand why it is so important to consider the competency of a test user and a test’s validity and reliability evidence within the context of a test’s intended use and the intended population.

**Test User Competency**

Evidence of a test’s usefulness can vary depending upon the competency of the people administering, scoring, and interpreting the test. An electric drill (the corded type, not the battery-powered ones that always seem to need recharging) can be very useful in the hands of a competent electrician who is skilled in carefully drilling holes in a wall while avoiding the electrical and water lines behind the wall. The same drill may be far less useful, and even dangerous, in the hands of a child or in the hands of an adult who acts like a child! Does this mean that a child or an incompetent adult could not drill a hole in the wall? Of course not; it simply means that the competent tool user will make better use of the tool, just as a competent test user will likely make better use of a test. Could a child use the drill to drill holes in the wall? Probably. Could the child avoid all the electrical and water lines behind the wall? And could the child avoid drilling a hole in its hand or electrocuting itself? We can only hope! In short, an electric drill’s, or a test’s usefulness, varies depending upon the competency of the person using it.

**Matching the Test’s Intended Purpose**

A screwdriver is intended to be used to drive screws. Nonetheless, who has not used a screwdriver as an ice pick, a lever or pry bar, a chisel, a paint mixing stick, a means to poke an older sibling in the eye, or for some other purpose? Did it work? Probably. Did it work as well as an ice pick, a lever or pry bar, a chisel, or a sharp stick would have? Probably not. In short, the screwdriver’s usefulness depends on whether you are using it for its intended purpose. It also is important to know whether the purpose of testing is specific or more general.

**Specific Purposes**  Like other tools, tests have been designed for many specific measurement purposes (achievement in various academic content areas, intellectual and personality functioning, vocational aptitudes, etc.). Like other tools, a test’s usefulness (i.e., the evidence of its validity and reliability) can vary, depending upon how well the current purpose of testing matches the specific purpose for which the test was developed. A test designed to identify individuals with above-average ability to quickly and accurately recognize typographical errors in a document may have excellent validity and score reliability when it is used to predict a potential employee’s ability to quickly and accurately recognize typographical errors in a book manuscript. On the other hand, the validity of the same test may be substantially lower if the test is used to predict a person’s ability to actually write a book (a very different skill, believe us!), or to predict a person’s ability to program a video game. In those cases, the test’s usefulness is more limited; it was not designed to assess writing or programming ability. This does not mean it is useless for this purpose, but there may be better tools ... oops, we mean tests, that would be more useful.

**General Purposes: Formative, Interim, and Summative Assessment**  In addition to being designed for a wide variety of specific content areas (e.g., assessing reading vocabulary, comprehension, spelling, mathematics, algebra, and general science), educational tests also can be designed to be administered at different times over the school year. The timing of test
WHY WE DEVELOPED THIS TEXT: ENHANCING TEST USEFULNESS

administration refers to the more general purposes of testing: formative (any time), interim (two to three times per year), and summative (end of year or semester) assessment. Historically, summative tests/assessments have been considered to be the most important tests in education, although their utility for day-to-day instructional decision making is limited. In recent years, formative and interim assessments have rapidly gained popularity in classroom testing.

Summative tests are administered after some period of instruction is completed (this can vary widely, e.g., a unit on vertebrates in biology, a semester of physics, or a year of algebra). Summative tests are intended to provide a measure or gauge of student learning following the completion of a unit of instruction. Summative tests are lengthy and are used to meet annual state and federal accountability requirements, to assign grades, to evaluate curriculum effectiveness, and to compare annual year-to-year performance among students, schools, and districts. Summative tests/assessments can be very useful if the purpose of testing is to inform us about broad achievement trends after instruction has been completed. However, summative tests/assessments may not be very useful if the purpose of testing is to evaluate acquisition of specific skills after a lesson, or the effectiveness of instruction on a day-to-day basis. Summative tests are simply not designed to be sensitive to small, day-to-day specific changes in achievement; rather, they are designed to measure larger and broader changes in achievement.

Formative tests/assessments are intended to assess the effectiveness of instruction on an ongoing basis (e.g., have specific skills been acquired?), and to inform day-to-day instructional decision making (e.g., move on to the next step in the curriculum, review or re-present the content using a different approach/medium, or provide instruction in a different setting). Formative assessments are intentionally brief (e.g., from one to several minutes) and are intended to be administered frequently, unlike summative tests that are typically administered annually or every few years.

In the past, most formative assessments were constructed by classroom teachers to support the learning process and inform instruction on a classroom by classroom basis. However, over the last decade, the adoption of the Response to Intervention (RTI) model, which we will describe in Chapter 3 and more recently, and the Common Core State Standards (CCSS), which will be discussed in detail in Chapter 2, have prompted a surge in formative assessment products developed by commercial test developers. However, questions have been raised about the validity, quality, and utility of recent, commercially developed formative assessments (Molnar, 2014).

Interim assessments reside between summative and formative assessments. They are longer than formative assessments but not as long as summative assessments. They are intended to be administered at more than one point during the academic year, at regular intervals (e.g., fall, winter, and spring). They are not intended to be as comprehensive as summative assessments, nor are they intended to be as narrowly focused as formative assessments. They are intended to measure student growth or progress over time and can be useful in differentiating instruction and evaluating instructional program effectiveness (i.e., both special education and general education programs). When they are used to identify students at risk for poor performance on an annual summative test, they are referred to as benchmark tests.

Matching Diverse Test-Takers to the Test

The U.S. population has become increasingly diverse in recent years, and there is no reason to expect that this trend will diminish any time soon. A wide range of cultural, linguistic, and academic backgrounds are common in today’s classrooms (Banks & Banks, 2013). Yet, the technical adequacy of many educational tests and assessments was established based on test development samples that included primarily, if not entirely, Caucasian, Hispanic American, and African-American students. Would we expect the technical adequacy of these tests to be the same when used with populations from different cultural, linguistic, and academic backgrounds (e.g., Middle Eastern and
Indonesian learners, limited English-speaking learners, and both higher and lower socioeconomic learners? Before you answer this question, let us return to the example of the electric drill.

Did you ever try to drill a hole in metal with a drill bit designed for drilling into wood? If you did, you will not make that mistake again! Specialized drill bits have been developed to enhance usefulness when drilling into diverse surfaces (e.g., wood, metal, concrete, and ceramic). Thus, a wood bit works best for drilling into wood, a metal bit for drilling into metal, and so on. Would we expect that one bit would work equally well for all diverse surfaces? Of course not. To be most useful, the drill bit must match the surface into which you are drilling the hole.

Things are no different with tests. For example, a test may be designed to measure certain characteristics for a particular group. We would expect the test to be most useful when used for similar groups, but like a single drill bit, we would not expect the test to be equally useful for all groups. Let us consider a multiple-choice test used to assess end-of-year achievement in a tenth-grade American History class (quick, is this a formative or a summative test?). The class is composed largely of two groups: native English-speaking U.S. children who have taken many multiple-choice tests and recent immigrants from Nicaragua who speak little English and have had little formal schooling. Excellent evidence may exist for the test’s technical adequacy (e.g., reliability and validity) when it is used to assess achievement for the English-speaking students. However, evidence for the same test’s reliability and validity may be less impressive (or even non-existent) when used to assess achievement for the limited English proficient (LEP), recently emigrated students from Nicaragua. In this case, linguistic (lack of familiarity with the English language) and cultural (lack of experience with multiple-choice tests in their native Nicaragua) factors may seriously limit the usefulness of the test. This consideration does not necessarily mean the test should not be used at all with this population. It does mean we should always be thoughtful and try to select the test that is most useful for the population we are testing—if there is one available—and to be very careful in interpreting results.

Although using a test that is not a great “fit” is not ideal, it is a matter of practicality. We simply lack tests that have strong evidence for their validity and usefulness with all populations and for all purposes for which tests are used. For example, school children in the Houston, Texas, public schools speak almost 200 different languages and have a similarly wide range of cultural backgrounds. Because of the diverse cultural, linguistic, and academic backgrounds of these students, it follows that the usefulness of the tests used to evaluate these diverse children will vary. This leads us to our next point.

Even though we should always strive to select the test that is most appropriate for the group(s) to be tested, we cannot always achieve this goal. When we cannot match the purpose and the group, we should try to be especially thoughtful and careful in interpreting test results. What else can we do?

Test Results and Diversity Considerations

What should you do when the group being tested does not match the characteristics of the sample used in its development? Depending on whom you ask, you will get a variety of suggestions. Here are ours. In such situations, the results of a single test administered at a single point in time should never be used alone to make important decisions—even when the technical adequacy, test user competency, and purpose criteria we have just described have been met. Instead, we recommend that testing should be part of a thoughtful, multifaceted approach to assessment, with input provided over time by multiple informants (i.e., teachers and other trained personnel). Simply put, in our diverse society there can be no “one-size-fits-all” test or assessment, and this applies to summative, interim, and formative tests.
That, as they say, is the theory (or perhaps wishful thinking on our part). Reality is very different. For more than two decades now, promotion, graduation, and other high-stakes educational decisions (e.g., ranking of schools as exemplary, acceptable, or in need of improvement) are commonly made based entirely, or primarily, on summative test scores obtained at a single point in time, in spite of the increasingly diverse nature of our society. This phenomenon is largely attributable to the rapid spread of the high-stakes testing (HST) movement since the mid-1990s and the passage of the No Child Left Behind Act (NCLB) (discussed in detail in Chapter 2).

That said, efforts have been undertaken to make accommodations for culturally, linguistically, and academically diverse test-takers (Flanagan, Ortiz & Alfonso, 2013). In some cases, tests have been translated or otherwise modified in an effort to better align them with diverse populations (Malda, van de Vijver & Temane, 2010). Nevertheless, the technical adequacy and appropriateness of these tests after such modifications, as well as the impact on results of the use of interpreters, and variations in examiner and situational variables have all proven difficult to determine (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999). In view of our nation’s increasing diversity, further study of the impact of such modifications is clearly needed. Fairbairn and Fox (2009) provide a summary of the relevant issues, and they also offer test development suggestions for English-language learners.

The importance of making decisions based on more than a single test result is not a concern limited to testing with diverse populations. Even when a test has technical adequacy, the test user is competent, and the purpose and population are appropriate, we still do not recommend making important educational decisions based on a single test administered at a single point in time. Instead of relying on such a limited “snapshot” or photograph (or JPEG) of student achievement for important decision making, we recommend that test results should be considered to be part of a broader “video” or process of measurement called assessment. We will describe the process of assessment in the next section and also distinguish between testing and assessment. See Box 1-1 about the Waco, Texas, public schools for an example of the controversial use of test results from a single test at a single point in time to make important educational decisions.

The situation described in Box 1-1 is not unusual. Well-intended educators and others continue to rely solely or primarily on test results from a single point in time to make important, high-stakes educational decisions. At times, they may have little choice because federal, state, or district requirements mandate “one-size-fits-all” policies that are tied to scores from a specific, “approved” test, without regard for the extent to which the validity and reliability of the scores from this test may vary for diverse populations of students, or for a different purpose than the one for which the test was developed.

To sum up, our position is that tests are only tools that can be appropriately used, abused, or misused. To minimize inappropriate test use, it is important to carefully consider the (1) evidence of a test’s technical adequacy, (2) competency of the test users, (3) extent to which the purpose of testing matches the purpose for which the test was developed, and (4) degree to which the test-takers match the group that was used to establish the technical adequacy of the test. Furthermore, we encourage you to consider additional background, historical, and observational data, especially when the test is administered to a group that differs from the test’s development sample, and when the test is used to make high-stakes educational decisions (Rhodes, Ochoa & Ortiz, 2005). In short, these situations call for a thoughtful and comprehensive assessment process rather than simply a testing/assessment snapshot.
BOX 1-1

WACO, TEXAS, SCHOOLS USE STANDARDIZED TEST SCORES ALONE TO MAKE PROMOTION DECISIONS

Concerned with possible negative effects of social promotion, the Waco, Texas, public schools decided to utilize standardized test scores as the basis for promotion decisions beginning with first graders in 1998. As a result, the number of students retained increased from 2 percent in 1997 to 20 percent in 1998 (The Waco Experiment, 1998). The Waco schools are not alone in curtailing social promotion. The Chicago public schools, in the midst of a wide-ranging series of educational reform initiatives, retained 22,000 students in 1994, with 175,000 retained in 1998 (Newsweek, June 22, 1998).

Social promotion is a practice that purports to protect student self-esteem by promoting students to the next grade so that they may stay with their classmates even when they are not academically ready for promotion. Educational, psychological, political, fiscal, cultural, and other controversies are all associated with social promotion. What has come to be known by some as the “Waco Experiment” also raised a number of measurement-related issues.

Even though the Waco schools’ decision was doubtless well intended, their policy may have overlooked the fact that the utility of test scores varies with age, with test results for young children being less stable and more prone to error than those for older children. A relatively poor score on a test may disappear in a few days, weeks, or months after additional development has occurred, regardless of achievement. In addition, older children are less susceptible to distractions and, with years of test-taking experience under their belts, are less likely to be confused by the tests or to have difficulty completing tests properly. All these factors can negatively affect a student’s score and result in a score that underrepresents the student’s true level of knowledge.

Furthermore, a single standardized test score provides only a portion of a child’s achievement over the school year, regardless of the grade level. As we will see when we consider the interpretation of standardized test results in Chapter 19, a number of student-related factors (e.g., illness and emotional upset) and administrative factors (e.g., allowing too little time and failing to read instructions verbatim) can negatively affect a student’s performance on the day the test was taken. Thus, making a decision that so substantially affects a child’s education based on a single measure obtained on a single day rather than relying on a compilation of measures (tests, ratings, observations, grades on assessments and portfolios, homework, etc.) obtained over the course of the school year seems ill-advised.

On the other hand, using data collected on a single day and from a single test to make what otherwise would be complex, time-consuming, and difficult decisions has obvious attraction. It appears to be expedient, accurate, and cost-effective and to be addressing concerns about the social promotion issue. However, it also may be simplistic and shortsighted if no plan exists to remediate those who are retained. As noted in a June 12, 1998, editorial in the Austin American-Statesman, “Failing students who don’t meet a minimum average score, without a good plan to help them improve, is the fast track to calamity.” Nevertheless, this trend has not diminished since we first reported on it in our sixth edition. Indeed, reliance on the use of test scores to make high-stakes promotion decisions has increased across the nation. Several states have now adopted versions of Florida’s retention policy, enacted by then Governor Jeb Bush in 2002–2003 to combat social promotion. In these states, students who do not pass the states’ high-stakes test must be retained, although there are often several “good cause” exemptions from this policy that soften this practice (Robelen, 2012).

THE DIFFERENCE BETWEEN TESTING/ASSESSMENT AND THE ASSESSMENT PROCESS

Today, the terms tests and assessments are commonly used interchangeably. Indeed, some seem to have eliminated the word “testing” from their vocabularies and replaced it with the word “assessment” because they believe that use of the word “assessment” is less evaluative, threatening, or negative than use of the word “testing.” In any case, we too will consider the terms
testing and assessment to be synonymous. However, we believe a clear distinction needs to be made between tests and assessments and the assessment process.

Tests and Assessments

The terms tests and assessments typically refer to single measures that yield results at a single point in time. There are exceptions, and some of these will be discussed in Chapters 9 and 10 (i.e., performance and portfolio assessments that combine multiple types of measures into a single aggregated rating). It is from the results of tests and assessments that we attempt to measure learning or to quantify some attribute or characteristic (e.g., intellectual ability and level of anxiety). Educational tests/assessments may be either formative, interim, or summative, depending upon whether they are used to measure day-to-day changes in learning (i.e., formative), over the course of months or a semester (i.e., interim), or learning over a more extended time frame (i.e., summative).

Assessment Process

The assessment process may span days, weeks, an entire semester, the entire school year, or longer. Formative, interim, and summative assessments are typically part of this broad assessment process, but are not end points. The assessment process is a means to an end, a comprehensive evaluation made up of many testing and assessment components and relevant background and contextual information. A comprehensive assessment process may include the following:

a. Traditional (i.e., summative and interim) test results from one or more multiple-choice, true-false, matching, or essay tests, or performance/portfolio assessments.

b. Progress monitoring (i.e., formative) results from less-traditional tests, such as curriculum-based measurement (CBM) or other brief, repeated measures.

c. A variety of other measurement procedures (e.g., observations, checklists, rating scales— included in the supplemental chapters on the textbook website at http://www.wiley.com/ college/kubiszyn).

d. The findings from all these assessments are integrated with relevant background and contextual information (e.g., language proficiency and cultural considerations—also covered later in the text, as well as the five factors that can affect a test’s usefulness that we described at the beginning of this chapter) to help ensure that educational decisions are as appropriate and as valid as possible.

Therefore, you can see that from our perspective, testing is only one part (i.e., like a snapshot or photograph) of the process of assessment that may include multiple photographs or segments (i.e., like a slide show, movie, or video) that reflect multiple types of information obtained from multiple informants at multiple points in time. Taken together, these components can provide us with a far richer and, we believe, more valid and accurate description of the individual than we can possibly obtain from any of the individual components alone. Figure 1.1 further clarifies the distinction between testing/assessment, and the assessment process. To avoid confusion later, be sure to understand the distinction we make between testing/assessment (as components of the assessment process) and the broader and richer assessment process itself, as depicted in Figure 1.1.

For an introductory chapter, we have exposed you to a lot of terms and concepts. If you are having any trouble absorbing all of this material, do not despair, but do take the time to review
CHAPTER 1 AN INTRODUCTION TO CONTEMPORARY EDUCATIONAL TESTING AND MEASUREMENT

Testing/Assessment

1. Tests (or assessments) are developed or selected, administered to the class, and scored.
2. Test results may then be used to make decisions about a pupil (assign a grade, recommend for an advanced program), instruction (repeat, review, move on), curriculum (replace, revise), or other educational activities.

An Assessment Process

1. Information is collected from tests and other measurement instruments (portfolios and performance assessments, rating scales, checklists, and observations).
2. This information is critically evaluated and integrated with relevant background and contextual information, including the five factors that can affect a test’s or assessment’s usefulness.
3. The integration of critically analyzed test results and other information results in a decision about a pupil (assign a grade, recommend for an advanced program), instruction (repeat, review, move on), curriculum (replace, revise), or other educational factors.

FIGURE 1.1 The distinction between testing/assessment and the assessment process.

and master what we have covered so far because these terms and concepts are important to your success as we go forward. Here are some of the main points we have made:

- Tests/assessments are only tools that can be well or poorly designed.
- Even well-designed tests can be misused or abused.
- Competent test users are less likely to misuse or abuse tests (we want you to become one of these!).
- The usefulness of tests/assessments can vary depending upon the purposes of testing.
- The usefulness of tests/assessments can vary depending upon the characteristics of the persons who are being tested, or assessed.
- Obtaining multiple forms of assessment information from multiple informants at multiple time points enriches our understanding of pupils.
- Important educational decisions should not be made based on a single test taken at a single point in time.
- A test/assessment is only one part of a much broader and more comprehensive process of assessment. Today, this process increasingly includes formative and interim data as well as summative assessment data.

Next, we will build on this foundation by introducing you to the various types of tests/assessments used in education and explaining how they differ from each other.

TYPES OF TESTS/ASSESSMENTS

So far, we have clarified the notion that tests are only tools, and we have described some of the factors that can affect the usefulness of these tools. Then we distinguished among formative, interim, and summative tests, and between tests/assessments and the much broader assessment process.
Next, we clarify some technical test-related terminology. The terms we introduce here will be referred to over and over again in the text. Even though it is important to understand as many of these terms as you can at this point, if you are like most students, you will need to return to this section repeatedly as you work your way through the text.

“You can’t tell the players without a scorecard.” Yes, this is an old baseball expression, so you are probably wondering what it is doing in a testing textbook. A baseball team has many players, and the scorecard, or roster, helps you to identify the players and the positions they play (or what purpose the players are best used for). Similarly, you will be introduced to many types of tests and assessments over the remainder of the text, and the next section of the chapter will serve as a “scorecard,” or reference, to help you identify and distinguish among the various types of tests and their best uses. In almost every case, the various types of tests and assessments will be discussed in much more detail later in the text. We should also note that these types of tests are used for formative, interim, summative, or in some cases for all three general assessment purposes. For now, our intention is only to highlight the major differences among objective, essay, and performance/portfolio assessments; standardized and teacher-made tests; norm-referenced and criterion-referenced tests; and the newest entry into classroom testing, curriculum-based measurements (CBMs).

### Objective, Essay, and Performance/Portfolio Tests/Assessments

The distinctions among these three types of tests have to do with the types of responses produced by test-takers, and how consistently, objectively, and reliably these responses can be scored. Tests that include item response formats that can be scored consistently and objectively are referred to as objective tests. Objective tests include multiple-choice, true-false, and matching formats, and student responses are typically limited to “bubbling in” electronically scanned answer sheets, or circling or underlining on test booklets. Even though objective test items can be used to measure higher-order learning and thinking (see Chapter 7), too often they are written to measure only factual knowledge.

Responses to essay test items vary in length from a single paragraph to several pages. For this reason, they are more difficult to score consistently and objectively. However, enabling students to respond in flexible and creative ways can make essay items well suited to measure analytic, composition, and higher-order thinking skills. A third item response format is the completion item. These items may require the test-taker to supply a word or phrase. Even though completion items are generally scored more consistently than essay items, completion items may not be able to be as objectively scored as objective test items. Completion items tend to measure factual knowledge.

Performance and portfolio examinations are intended to reduce pressures to test solely for factual knowledge and to provide a stimulus to introduce more extended thinking and reasoning activities into the curriculum (Borich, 2015; Borich & Tombari, 2004). Responses may be products, processes, collections, or other combinations of activities. They may include having students videotape portions of their own projects, conducting interviews with the student in order to probe for understanding and thinking abilities, or making a visual inspection of a product in order to determine whether it has the required characteristics for successful operation. The intent is usually for the learner to carry out the activities actually used in the real world (such as measuring the tensile strength of a building material, estimating the effects of pollutants on aquatic life, designing circuitry for a microprocessor, or assembling a folder or portfolio of “works in progress” to reflect growth in critical or integrative thinking or other skills over the course of a semester or school year). As with essay tests, these measures can be difficult to score consistently, objectively, and reliably. They will be considered in depth in Chapters 9 and 10.
Teacher-Made and Standardized Tests

Teacher-made tests are developed, administered, and scored by teachers, and often consist of completion, true-false, matching, multiple-choice, and essay items. Teacher-made tests are often flexible, or variable, in terms of their administration and scoring procedures, and in the amount of attention given to their construction. Different teachers may be more or less careful in constructing their tests, may allow more or less time for the test to be taken, and may be more or less stringent in grading the test. Even though teachers may use the same test repeatedly, they may also make modifications to the test each time they use it. In the latter case, the usefulness of performance comparisons over time would be reduced.

Standardized tests (described in more detail in Chapters 19 and 20) are typically developed by test construction professionals over lengthy intervals and at considerable expense to standardized test publishers, and state education agencies. The developmental process typically includes pilot testing, and then multiple drafts of the test are evaluated over several years to enhance item quality and the overall quality of the test. Standardized tests are called so because they must be administered and scored in a consistent manner, no matter who administers them, or where or when they may be administered. Standardized test users are specifically trained to follow exactly the same administration procedures, and scorers also are trained to always follow the exact same scoring procedures. Standardized achievement tests often include multiple-choice, matching, and true-false items. Essay items are also increasingly common in standardized tests.

Norm-Referenced Tests (NRTs) and Criterion-Referenced Tests (CRTs)

Norm-referenced tests (NRTs) are typically standardized tests developed by commercial test publishers (e.g., the SAT), sometimes in conjunction with state education agencies. They are designed to enable us to compare the performance of students who currently take the test with a sample of students who completed the test when it was being developed. The sample of students who completed the test in the past is called a norm group (or normative group or sample). NRTs tend to measure broad educational goals and are usually lengthy (hours long in duration). If the purpose of the NRT is to enable current student performance to be compared to national averages, or norms, the normative sample must be carefully selected to proportionately represent the key characteristics of the national population (e.g., age, gender, ethnicity, parental income, and geographic region). However, current student performance may also be compared to local or state norms. In these cases, the samples may or may not be similar to national norms, depending on how closely local or state samples compare to key characteristics of the national population. Where local or state norms differ significantly from national norms, it is possible that current student performance may be above average when compared to national norms, for example, but below average compared to local norms (e.g., if the local norms come from an unusually high-achieving school or district). In any case, for an NRT, the meaning given to a student’s performance is always relative to that of the norm group (e.g., “Marie is above average” or “Joel is well below average,” where average refers to the norm group’s performance). NRTs will be discussed more fully in Chapters 5, 19, and 20.

CRTs may be standardized or teacher made, and enable a different kind of comparison. Compared to NRTs, CRTs are typically shorter in length and narrower in focus. Instead of comparing current student performance to other students (i.e., the norm group), CRTs enable comparisons to an absolute standard or criterion. CRTs help us to determine what a student can or cannot do. Rather than stating that “Marie is above average,” CRTs enable us to make judgments about a student’s (or group of students’) level of proficiency or mastery over a skill or set of skills (e.g., “Marie is
able to spell the words in the third-grade spelling list with greater than 80% accuracy,” or “Joel is less than 10% accurate when subtracting two-digit numbers with borrowing.”). For this reason, and because they are shorter than NRTs, scores from a CRT are more likely to be useful for formative, day-to-day instructional decision making than scores from an NRT would be. CRTs are described more fully in Chapter 5.

Curriculum-Based Measurements (CBMs)

Although relatively new to the regular education classroom, brief, formative, CBMs (commonly referred to as CBM probes) have been in use in special education programs since the early 1980s (Deno, Mirkin & Chang, 1982). CBM probes may be standardized or teacher made. CBM probes are designed to be sensitive to small changes in learning. This makes CBM probes suitable to monitor day-to-day progress in reading (CBM-R), math (CBM-M), writing (CBM-W), and spelling (CBM-S), which in turn facilitates day-to-day instructional decision making. Commercial CBM probes are commonly norm referenced and may include national, state, and/or local norms. However, unlike commercial NRTs, the national norm groups may not be carefully selected to match key characteristics of the population. Instead, commercial CBM norm groups typically are “convenience” samples, developed from large numbers of students tested with CBMs over time without the intent to match population characteristics.

CBM probes, or other brief formative assessment measures, are now crucial to the progress monitoring and data-based decision making components of the response to intervention (RTI) model required by the 2004 Individuals with Disabilities Education Improvement Act (IDEIA). We will discuss RTI and its significance for the regular education teacher in detail in Chapter 3. For now, suffice it to say that brief, formative assessments like CBM will be increasingly used to monitor the progress of both regular and special education students in the regular education curriculum, as required by the IDEIA.

EFFECTS ON THE CLASSROOM TEACHER

As we said earlier, we believe that a firm grounding in testing and assessment practice is vital to today’s classroom teacher. However, teachers traditionally have not been well trained in test construction and use or in assessment. Many teachers see no need for training in testing and assessment because they believe these activities are supplemental or peripheral to the instructional process. Indeed, many prospective teachers have been heard to say things like, “If I need to know about tests, I’ll go to the counselor!” Perhaps the trend toward seemingly ever-increasing “specialists” in education has fostered or reinforced such beliefs. Nevertheless, regardless of how a teacher feels about the relationship of tests and assessments to instruction, it is frequently the classroom teacher who must administer and then organize and interpret state-mandated high stakes and teacher-constructed test and assessment data, including performance and portfolio data, to curious and sometimes hostile parents and other concerned parties. Under IDEIA, regular classroom teachers will also collect and interpret these kinds of data to the parents of children with disabilities, and their advocates. Figure 1.2 depicts many sources of pressure influencing a teacher’s use of tests. It is intended to convince you of the seriousness with which you should consider your knowledge or lack of knowledge of testing and assessment. After years of consulting and teaching experience in the public schools, we have concluded that discussing test results with parents is frequently an anxiety-provoking experience for teachers. Lacking appropriate training, some teachers are tempted to try to avoid
interpretation entirely by dismissing test data as unimportant or inaccurate, not realizing that such data are often very important and thought to be quite accurate by parents. The following dialogue occurred when one of the authors had a conference with his son’s second-grade teacher.

**Teacher:** Hi, I’m Jeff’s second-grade teacher. What can I do for you so early in the year?

**Author:** Jeff says he’s in the low reading group, and I am curious about why he is. Could you explain that to me?

**Teacher:** Oh, don’t worry—we don’t label kids at this school. I think that would be a terrible injustice.

**Author:** I see, but could you explain why Jeff is in the “Walkers” instead of the “Runners”?

**Teacher:** Oh, those are just names, that’s all.

**Author:** Are both groups at the same level in reading?

**Teacher:** They’re both at the first-grade level, yes.

**Author:** I’m beginning to catch on. Are they reading in the same level books in the same reading series?
Teacher: Of course not! Some children are further along than others—that’s all—but the kids don’t know.
Author: Let me guess, the “Runners” are further ahead?
Teacher: Yes, they’re in Book 9.
Author: And the “Walkers” are in …
Teacher: Book 5. But they are only grouped for instructional purposes—I have 25 students, you know!
Author: I’m confused. Jeff’s reading scores on the California Achievement Test last May were above the 90th percentile.
Teacher: (Chuckles to herself) I can understand your confusion. Those test scores are so hard to understand. Why, even we professionals can’t understand them.
Author: A score at the 90th percentile means the score was higher than the scores of 90% of the students who took the test all across the country.
Teacher: Oh, really? (Blushing) It is very complicated. As I said, even professional educators don’t understand testing.
Author: Some do, Mrs B.

Had the teacher understood the data she was dealing with, an embarrassing situation might have been avoided. Unfortunately, many classroom teachers have similar experiences. Another often overheard statement is, “I don’t even have time to teach. How can I learn about test construction and interpretation?” We are well aware that paperwork and bureaucratic requirements increasingly burden teachers and most other professionals.

Educational testing has increased over the last few decades, primarily to meet the accountability requirements included in state and federal legislation (i.e., state high-stakes testing and the NCLB act). We discuss those and other testing developments in detail in Chapter 2. Although there has been resistance to increased testing demands, we continue to anticipate high levels of test and assessment use in the future. The public will be keeping a watchful eye on the collectors and users of test data. Teachers of the future will continue to need to be adequately trained in formative, interim, and summative testing and assessment practice. Ignoring or denying the need for proper training in this area will not make the watchful eye of the public, administrators, or other decision makers go away. Given the time constraints under which the average teacher must operate, it appears wiser to seek such training now, before you begin teaching. The alternative may be to attend what might seem like an endless series of workshops later on, while you also have to attend to your class of students and all the related responsibilities.

In short, it looks like tomorrow’s teachers will continue to be regularly exposed to student test and assessment data and will need to explain and interpret those data to parents and others. Public pressure for accountability, coupled with the formative, interim, and summative testing requirements driven by state and federal education reform initiatives, suggests that the average teacher will not be able to get ahead or even get by without a good working knowledge of test and assessment practice. We sincerely hope that this text will help you acquire such knowledge and the skills necessary to construct good tests and use test data knowledgeably and professionally.

ABOUT THE TEXT

When you talk to teachers who have completed courses in testing and assessment, you often find that they are somewhat less than inspired by their courses. In fact, talking with such veterans may have made you a little reluctant to register for this course. Nevertheless, here you are! The authors
have taught testing and assessment courses for many years and, quite frankly, are accustomed to hearing the course content described as “dry, but necessary.” Too many times, however, we have heard, “The course is dry, but necessary—I think!” Since a fair amount of straightforward, technical information is presented, we could understand calling a course in testing and assessment “dry.” Indeed, this is almost how we feel about some of the content when we teach it! Thus, hearing “dry, but necessary” from students did little more than confirm our own feelings. But, to hear “dry, but necessary—I think!” made us think and led us to evaluate our courses to determine the cause for this uncertainty. Remember, we stated earlier that we think knowledge of testing and assessment practice is vital to today’s teacher, yet somehow this was not coming across to some of our students.

We concluded that what was needed to engage readers with the material we present was a text that (1) emphasizes the application of formative, interim, and summative test and assessment practices and concepts to real-world situations and (2) provides new teachers with an understanding of the political, philosophical, pedagogical, and other factors that affect the educational testing, assessment, and measurement landscape today. With this aim in mind we wrote, and have once again recently revised, this text. Only you can tell whether we have succeeded. We invite your comments and suggestions.

**WHAT IF YOU ARE “NO GOOD IN MATH”?**

Since tests yield numerical scores and test theory is based on statistical principles, it is no surprise that many who feel weak in math have great reservations about being able to succeed in a course in testing and assessment. If you fall into this category, rest assured that many of your fears are groundless or exaggerated. In our experience, fewer than 1% of the students who have completed the testing and assessment courses we have taught have done poorly solely because of a weak math background.

Naturally, knowledge of the basic mathematical functions—addition, subtraction, multiplication, and division—is a necessity. Mix these with fractions, decimals, and a sprinkling of algebra and you are all set! If you are still doubtful, work your way through the review of math skills provided in Appendix A. That review and accompanying self-check test should also prove useful to you if you feel a little rusty in math—a fairly common experience in this age of calculators and personal computers. All operations necessary to complete the calculations in this text are covered in the review. We think you will agree that our review of math skills in Appendix A will be all that is needed.

**SUMMARY**

In this chapter, we have introduced you to issues related to educational testing and measurement and have described the orientation of the text. Its major points are as follows:

1. Tests are only tools, and like all tools, poor design, unintentional misuse, and intentional abuse can impair their usefulness.

2. Several factors can affect the usefulness of any test. These factors include the following:
   a. The technical adequacy of the test.
   b. The competency of the test user.
   c. The extent to which the purpose the test is being used for matches the purpose for which the test was developed.
d. The extent to which the population being tested matches the population the test was developed on.
e. Whether additional sources of information also are considered, especially with culturally, linguistically, and academically diverse students.

3. Summative assessments are used when the purpose of testing is to evaluate the extent of student achievement after some period of instruction. They typically are lengthy and administered at the end of the school year or semester.

4. Formative assessments are used when the purpose of testing is to inform the instructor about whether students are learning while the curriculum is being delivered, which enables ongoing instructional adjustments to be made. Formative assessments are also considered to be useful for assessing student responsiveness to instruction and tend to be brief. Interim assessments are administered two to three times per year (e.g., fall, winter, and spring) to indicate whether progress over time is occurring. When interim assessments are used to identify students at risk for failure on a summative test they are called benchmark tests.

5. Because there is no “one-size-fits-all” test, important educational decisions should not be made based entirely on test results from a single point in time (i.e., like a photo). Important decisions should also consider information from multiple sources obtained over multiple time points (i.e., like a video).

6. Tests and assessments are terms that typically refer to single measures that yield results at a single point in time.

7. Testing and assessment should be considered to be only a part of an assessment process that includes testing/assessment and the use of other measurement techniques (e.g., performance and portfolio assessments, rating scales, checklists, and observations) along with relevant background and contextual information in a critical, integrated way to make educational decisions.

8. Tests, because they have no preconceived or biased notions about test-takers, can provide objective data that can be helpful in educational decision making and can minimize bias and subjectivity that may characterize decisions based on subjective data.

9. Test use (i.e., both formative and summative) is more likely to increase than decrease in the foreseeable future.

10. There are many different types of tests and assessments, including objective, essay, and performance and portfolio tests/assessments; teacher-made and standardized tests; norm-referenced tests and criterion-referenced tests; and curriculum-based measurements (CBMs).

11. The classroom teacher who is trained in educational testing procedures will be able to use test results more efficiently and effectively and will be less likely to misuse or abuse test results.

12. This text is oriented toward the sound application of measurement principles to real-world situations.

FOR DISCUSSION AND PRACTICE

1. Thinking back to your own high school days, what were some of the factors that turned you off about tests?

2. Were these factors connected with the tests themselves, or were they the result of how the tests were used?

3. You ask an experienced fellow teacher, “How do you decide which tests to use for your students?” She responds, “That’s easy. Just pick one. They are all about the same and equally useful.” On the basis of what you’ve read in the text, how would you respond?

4. Noting that you have results from teacher-made tests, the district’s standardized high-stakes tests, formative CBM probes, last year’s grades, and a variety of other information in front of you, Mr. Quick asks “What is all that for?” You respond, “I am assigning my final grades for the semester, so I’m using all the information I have to try to be as fair as possible.” “There is no need for all that,” asserts Mr. Quick, “I just use the scores from the high-stakes test to assign my grades. It only takes a minute.” What, if anything, might be wrong with Mr. Quick’s approach to grading?
Chapter 1: An Introduction to Contemporary Educational Testing and Measurement

5. Distinguish between what the authors call the specific purposes and general purposes of educational testing in the classroom.

6. Compare and contrast formative assessments with interim and summative assessments.

7. At lunch, you mention to another teacher that you are planning to start using CBM probes to determine how responsive your students are to the new research-based instructional approach you are implementing. She says, “What on earth are CBM probes?” How do you respond?

8. Differentiate testing/assessment from the assessment process.

9. Imagine the modern U.S. high school 20 years from now. Identify several ways in which testing and measurement will have changed since the time you were attending high school.

*Answers for these questions appear in Appendix B.