1

A Review of the Nature of Standardized Tests

This chapter presents an overview of standardized tests and their applications in education. The first section presents a brief historical review of standardized testing in schools. This is followed by a discussion of the nature of these tests and their uses in education. Several technical terms (e.g., reliability) are introduced and discussed.

A Brief History of Standardized Testing in America

Large-scale standardized testing in the United States can be traced to the First World War. At the beginning of U.S. involvement in the war, the military was overwhelmed with volunteers. At the time, much of the country, including the

military, was deeply immersed in the efficiency movement. The idea, borrowed from the industrial workplace, was to use input, including human capital, in such a manner as to maximize output and minimize waste. Subscribing to this idea, the U.S. military was committed to finding scientific ways to maximize the efficiency with which it used human capital for its war machine. A solution was offered by leaders of the American Psychological Association (APA).

Headed by Robert Yerkes, the APA proposed developing an objective and scientific way for planners to allocate men to positions in the military hierarchy. Yerkes and his colleagues proposed and developed two tests designed to measure the mental ages of recruits and volunteers. The Army Alpha test was developed for examinees who could read and the Army Beta test was developed for those who could not. These examinations were administered to nearly two million young men. The military used the test results to classify examinees for various posts, ranging from those selected for officer training to those who were labeled "morons" or "imbeciles" and dismissed.

The results of the Army testing project were widely considered to have been a phenomenal success. In fact, within a few decades after the war, the number and variety of standardized tests had increased exponentially and there was almost no sector of U.S. society untouched by the standardized testing movement (e.g., see Haney, 1984). As Popham (2000) noted, "Almost anyone who could crank out multiple-choice items and bundle them together, or so it appeared, began publishing group aptitude or achievement tests" (p. 19). As some observers have noted, were it not for the success of the military testing project, it is possible that the standardized testing movement in this country would have remained largely an academic pursuit and not have taken center stage in educational policies and reforms (see Haney, 1984).

The Growth of Educational Testing

Shortly after World War I, the Scholastic Aptitude Test (SAT) was developed and adopted by many colleges for

admissions purposes. Achievement tests developed by E. L. Thorndike at Columbia University began to find their way into schools and districts across the country. Driven by legislation which mandated schooling of immigrants, standardized tests played an increasing role in the educational process. Standardized intelligence and achievement tests were seen by proponents as tools which could bring efficiency to schooling by (a) providing a means of allocating a diverse population of students to educational experiences which were best suited to their "native" abilities, and (b) providing policymakers and the public with an objective and fair assessment of actual achievement (see Cronbach, 1975).

The rate of growth of standardized testing increased with the arrival of high-speed computing. Computers made it possible to score multiple-choice exams electronically and in a fraction of the time required to score them by hand. The result was to make testing of millions both practical and economically feasible. As a result, large-scale standardized testing grew rapidly and by the middle of the 20th century almost all school systems were involved in some form. The federal government gave momentum to this movement in the 1960s by requiring that standardized achievement tests be used to gauge the success of the massive Title I programs funded under the Elementary and Secondary Education Act of 1965 (ESEA; see Brooks & Pakes, 1993).

Accompanying the growth of standardized testing was a new professional, the psychometrician or measurement specialist. These individuals were trained in the traditions of quantitative psychologists and statisticians. The specialists provided the technical foundations for the testing movement. Their jargon and technical sophistication made their discipline virtually inaccessible to most educational practitioners. Backed by large testing companies and corporations, testing specialists defined standards for the development and use of educational and psychological assessments so much so that in time assessments for high-stakes educational decisions were seen to be beyond the ability of educational practitioners (see Haney, 1984).

Trends and Challenges In Standardized Testing

The testing boom which began in the 1920s on the heels of the Army Alpha and Beta project drew criticism and controversy, but nothing like what occurred in the 1960s. With the growth of psychometrics as a distinct discipline and the tendency of political leaders to turn to standardized testing as a policy tool, the number of tests administered in schools increased and the types of decisions based on test scores also increased. By the late 1960s, most states had extensive standardized testing programs. These programs linked test results to such a growing variety of high-stakes decisions that even proponents of standardized testing often questioned some of their uses (see Haney, 1984). The inevitable result was backlash. The 1960s saw a variety of widely read and influential criticisms of standardized testing, perhaps culminating in the call by several influential educational organizations for a moratorium on all standardized testing in schools (e.g., see Bandesh, 1962; Gould, 1981).

The criticisms of the 1960s focused on many aspects of testing programs, but a key problem was the relevance of standardized testing for classroom decisions. Attracted by the air of scientific objectivity, policymakers often ignored the fact that most standardized tests only partially matched local curriculum and provided little information about the skills and abilities students actually had. The results of these criticisms and general social concerns about the quality of education provided to children led to the criterion-referenced and minimum competency testing movements (see Mehrens & Lehmann, 1991). Unlike traditional standardized tests, these new tests were designed to provide information about what students could and could not do, and to certify whether students had met minimum performance standards.

If the criterion-referenced testing (CRT) and minimum competency testing movements changed the nature of standardized tests and promoted mandatory testing, *A Nation at*

Risk and related publications raised the stakes to a level not seen before (National Commission on Excellence in Education, 1983). The famous conclusion of this report, which pointed to poorly performing schools as a threat to national security, resonated with the public and policymakers in the early 1980s and served to place state mandated high-stakes testing at the front of the educational agenda. Not long after this report was released, mandated standardized testing existed in nearly every state. These tests were linked to student promotions, teacher evaluations, school evaluations, and so on. The pressure, however, of high-stakes testing and accountability often led to questionable educational practices. While the normal expectation was that these problems primarily involved teacher activities such as providing inappropriate aid for students, an article published in the late 1980s posed the prospect that even large testing companies were somehow responsible for polluting the quality of the educational process. A West Virginia physician reported that while he treated children who could barely read, he was astounded to find that they had high scores on standardized reading tests (Cannell, 1988). More important, he observed that nearly all states reported the impossible results that they were above the national average. The resulting question was whether testing companies were in collusion with state departments of education. These criticisms served to draw attention to test use, test preparation practices, and, most important, the impact of testing on teaching and learning (see Haladyna, 2002).

Questions about how tests scores affected the teaching and learning process in schools reached a crescendo in the mid-1990s and led to what has been termed the authentic testing movement. The multiple-choice item, which had been the hallmark of standardized tests since the Army Alpha and Beta tests, was seen to be limited in the types of skills it assessed and was believed to promote and reflect a behaviorist/mastery perspective on learning (Shepard, 1991). In the 1990s, much of the educational community had begun to feel that learning and thinking were more complicated than the simple

mastery learning, input–output model implied by the multiple-choice item. Perhaps best captured in a classic paper by Grant Wiggins was the call for a new form of test, one in which students are encouraged to think, for which there may not be just one right answer (Wiggins, 1989). The impact of this movement was to once again change the nature of standardized tests. Traditional multiple-choice items were, if not completely replaced, given a far less prominent role in testing programs, replaced by open-ended items, performances, and other tasks thought to be more consistent with the complexity of the ways in which children think and learn.

The current trend in standardized testing is toward what has been referred to as the standards movement. The idea is that not only should tests be more consistent with the ways in which people think and learn, but the content of the test and the criterion for performance should both reflect the highest standards with respect to national and international goals and norms. It should be noted that there is considerable debate on the quality and usefulness of state standards and their appropriate role in education (Falk, 2000).

No Child Left Behind

The Elementary and Secondary Education Act of 1965 was designed to redress discrepancies in educational outcomes among students, which seemed to be linked to differences in socioeconomic background. The Title I provision of the act provided funds for schools serving large percentages of low-income students, but added the caveat that schools needed to demonstrate their effectiveness using standardized tests. The impact was a dramatic increase in the use of standardized testing in schools. The current reauthorization of ESEA, known as No Child Left Behind (NCLB), promises to expand high-stakes testing as never before (for more information, see www.ed.gov/legislation/ESEA02/). NCLB redirects federal support for education to local school systems and specifically calls for the following:

- Mandatory testing of all students in Grades 3 through 8
- Use of test results to evaluate the performance of schools
- Reporting of test results to parents and other stakeholders

More than previous versions of this law, standardized tests are at the heart of NCLB. This bill also explicitly links students' outcomes on standardized tests with significant consequences for schools and educators. These consequences include recognition and rewards for schools that meet growth targets and interventions and the prospect of closure for those that do not.

What Do School Administrators Need to Know About Standardized Tests?

Because of the obvious growth of standardized testing and the increase in consequences of student outcomes on these tests, it is important for school administrators to be well versed in the use of standardized tests and knowledgeable of their basic character. The following section presents information on the basic structure of standardized tests.

Basic Characteristics of a Standardized Test

Standardized Testing

What Is a Standardized Test?

The most direct answer to this question is that a standardized test is an examination administered under strictly uniform conditions and interpreted in a consistent manner. The essence of this definition is that all the key aspects of testing are uniform. The same test is administered to all examinees, the conditions under which the test is administered are standardized (time, resources, etc.), and the ways in which scores are interpreted are likewise standardized.

What Is the Advantage or Motivation for Standardization?

To answer this question, it is insightful to once again consider history. While some form of testing has been around as long as schools have existed, for most of that history tests were either oral or essay-based. Examinees were given questions by an instructor or committee and responses were evaluated. The questions were not necessarily of the same difficulty and responses were not necessarily given the same degree of scrutiny. Allegations of cheating, favoritism, and a host of other problems led to many challenges and disputes. These issues reached something of a crisis when widely reported research studies demonstrated that the grades teachers assigned to essays and math exams varied dramatically from one teacher to another and from one occasion to another for the same teacher (see Haney, 1984).

Modern ideas about testing developed from the work of experimental psychologists interested in measuring human capacities which could not be easily quantified. These included intelligence, hearing, and sensory acuity. A guiding principle of this movement was the concept of experimental control, which proposed that measurements of human subjects should be made under strict laboratory conditions. This meant that scientists and their assistants would make their observations and measurements under specified environmental conditions, in a specified way, and adhere to professional guidelines in interpreting their results—the goal being to gain a precise and accurate measurement of the subject. These principles were transferred from the laboratory and adapted to the measurement of attributes such as intelligence, achievement, and beliefs, as well as a host of others. It is this adaptation that gives the standardized tests we encounter in schools their distinct characteristics. Whether the focus is on intelligence, achievement, attitudes, or some other attribute, calling a measure a standardized assessment implies that (a) it has been carefully constructed to measure the construct of interest, (b) the conditions under which the examination should be administered are specified and carefully controlled, (c) the

way in which responses are scored is specified, and (d) the way in which scores are interpreted, that is, their meaning, follows precise rules. It is hoped that these tools will lead to accurate measurements. In the next sections we focus more specifically on the construction of standardized tests.

The Construction of Standardized Tests

It can be said that one of the best strategies for confronting a challenge is a good understanding of its nature. With that in mind, we focus on the development and construction of standardized tests. Although there are many different types of standardized measures, as discussed below, for the moment we focus on measures of academic achievement. Typically, constructing a standardized achievement test involves the following tasks: (a) specifying the construct the test is intended to measure, (b) developing items or tasks to measure the desired constructs, and (c) administering the test and analyzing results for evidence of its quality and for purposes of constructing interpretive aids.

Specifying the Construct

In a typical classroom, a teacher might identify the objectives covered over the past three weeks. The teacher might then determine the importance of each objective and develop a corresponding number of related items. The process is similar with respect to the construction of a standardized measure; however, the scope is different. Because a standardized test is a major undertaking, test companies strive to develop measures that are as widely applicable as possible. With regard to achievement, to increase the relevance of their test, developers select a specific content area and grade level and devise a test which measures those things that are commonly taught or expected of students across the country. This might mean obtaining copies of curriculum guides from hundreds of school districts and dozens of state education agencies. Widely used textbooks, instructional materials, and guidelines

published by professional associations might also be collected for purposes of review. These documents would all be carefully screened and analyzed by measurement experts, content experts, and practitioners in an effort to identify the most commonly addressed instructional outcomes. These outcomes would form the content domain of the test. It is important to note that this domain is an abstraction and might not actually exist in any school district or any school in the population for which the test was designed. It therefore becomes important for prospective users of a test to correlate their local instructional objectives with the content domain represented in the test.

The development process applies to tests targeted for a national audience. However, the same issues exist for tests tailored for specific states or even school districts. Since the early 1970s, most states have developed statewide assessment programs which include a standardized test developed specifically to reflect students' mastery of the state's curriculum. In the past, these were largely criterion-referenced tests designed to reflect whether students had met minimum performance levels. More recently, state curriculum frameworks specify not only content standards (what students ought to know), but also performance standards (what students ought to be able to do). Standards-based assessments tend to report performance in terms of the proficiency levels of students, with many states making use of some variant of the reporting scheme for the National Assessment of Educational Progress (NAEP): Advanced, Proficient, Basic, and Unsatisfactory (see the NAEP Web site at http://nces.ed.gov/nationsreportcard/).

The controversy associated with standards-based state assessment programs stems from the fact that the implementation of the state curriculum can vary dramatically from one setting to another. These variations have to do with the competencies of teachers, especially their ability to implement classroom innovations; the resource differentials which exist between and within school systems; and the external supports which may vary dramatically from one community to another. The impact of these factors is such that the printed

curriculum and the taught curriculum may differ, and these differences may not be simply due to random chance. It is therefore important for school-based personnel to exercise caution in drawing inferences from student performances on national or state standardized tests.

Items and Tasks

Turning again to our discussion of the development of standardized tests, once the content domain and emphasis of the test have been defined, the next step is to construct or select the test items. A typical classroom teacher might simply choose items from the instructor's manual which accompanied the text or she might develop her own. In either event, if the items appeared to pose problems for students, they might be discarded and not used in future tests. Contrast this with the item development process of a large testing company. First, there is likely to be an entire staff of professional item writers. These are typically content experts that have been trained in matching item formats to instructional objectives and in writing items that avoid many of the pitfalls (e.g., ambiguity) of the typical classroom tests. The items are drafted, undergo an extensive internal review for technical quality and content appropriateness, are likely to be reviewed by a bias review committee, and will then undergo extensive field tests in the target population. The field-test results contain information on the item's difficulty, discrimination, and a host of other characteristics. As typically used, item difficulty refers to the percentage of respondents who respond correctly to an item, and discrimination is a measure of an item's ability to distinguish between more and less knowledgeable examinees. These statistics, as well as others, determine if an item will be included in an actual test.

Technical Characteristics of Standardized Tests

Once the items have met whatever criteria are set, the test is assembled and the question of the quality of the test is

raised. The question of quality among measurement specialists is usually answered by examining the reliability and validity of the measure.

Reliability

In popular use, reliability refers to the extent to which one obtains consistent results with some thing or process. For example, a reliable automobile is one that consistently starts when the ignition is turned, a reliable employee consistently shows up for work when scheduled, and a reliable performer consistently yields good (or bad) performances. The point is that the question of quality, for most people, is intricately linked to the notion of consistency. Few people, for instance, would be satisfied with an automobile that could only be expected to start 50% of the time. Similarly, when it comes to measurements of human attributes, a quality measure should yield consistent scores when an individual's standing on the attribute has not changed. For example, if I were to stand on a scale and record my weight, step off the scale, and a moment later repeat the process, by all accounts, the numbers I record should be the same. If they are not, then there is something problematic about this measurement procedure. Perhaps I was not particularly careful in recording the measurements, or perhaps the scale suffers from some mechanical malady which leads it to yield varying weights for a given object. In either event, the lack of consistency in results is clearly a problem, one which would threaten my confidence in the quality of the resulting scores and the usefulness of the procedure.

In the context of educational achievement, the notion of consistency is still intricately linked to assessments of the quality of an examination. In fact, estimates of the reliability of a test can be thought of as indexes of the extent to which the test yields consistent scores for examinees. There are many different ways of estimating the reliability of a test. These techniques differ largely with respect to the types of inconsistency they detect.

Test–Retest Estimates of Reliability. The test–retest technique for estimating reliability entails administering the same test to a group of examinees on two distinct occasions. Of course, this is only reasonable if the attribute measured by the exam is expected to be stable over time. As such, these indexes are typically only reported for measures of intelligence, aptitude, and other characteristics expected to remain stable or change very slowly. If the scores of examinees are not similar on the two occasions, then the inference is that scores on the exam are subject to random errors.

Reliability Estimates Based on Equivalent Forms. It is not uncommon for test developers to have multiple forms of the same test. This is necessary for purposes of makeup exams, educational research using a pre–post design, and so on. Because the two forms of the exam will not consist of the same items, administering both forms to the same examinees provides valuable information. One interpretation is that the scores should be the same unless the responses of examinees are affected by extraneous factors.

Internal Consistency Estimates of Reliability. Among the most commonly reported forms of reliability are measures of internal consistency. These indexes reflect the extent to which the items in a test yield consistent outcomes for examinees. If the items measure dramatically different constructs, this type of evidence could be misleading. However, most tests have multiple items designed to measure the same or very similar skills. Comparing examinee performance across the items yields valuable information about consistency. Some of the more popular measures of internal consistency are the split-half estimates, Kuder–Richardson formulas, and Cronbach's alpha.

Interrater Estimates of Reliability. Because many of the current commercially developed tests have open-ended items or give respondents latitude in constructing a response, judges must

be used to evaluate the appropriateness of a given response. If the system of scoring responses is operating as one would hope, the scores assigned by different judges to a given examinee should be consistent. Interrater estimates of reliability reflect this aspect of consistency.

Reliability Estimates for Classifications of Examinees. The minimum-competency and CRT movements of the early 1970s emerged in response to the widespread use of standardized tests designed to facilitate comparisons among students. These latter tests were not particularly informative about what examinees could or could not do. CRT was designed to accomplish this end. Accordingly, based on performance on CRT, examinees are typically classified into one of two categories: those who demonstrated mastery of the content of the test and those who did not. In this instance, reliability is primarily focused on the consistency of these classifications. A popular measure of the extent to which examinees are consistently classified is the kappa index.

Standard Error of Measurement

The reliability statistics mentioned above are group-based statistics. A measure of the amount of error associated with a specific examinee is given by the standard error of measurement. This index can be interpreted as the typical amount of error associated with the scores of individuals. It is typically used to construct a range or interval for the scores of individual examinees.

Validity

The issue of validity is the most basic of all measurement concepts. Simply put, it poses the question of whether the instrument measures that which the user intends. This concept is so fundamental that it precedes the question of reliability in importance. If a test is not valid, then its reliability is of no importance to the user. On the other hand, a test

cannot be valid if it reflects only random error. Reliability is a necessary, but not sufficient, condition for validity.

Validity is a question of test use. Any given test can be expected to be differentially valid, given the various uses to which it is put. A test designed to measure arithmetic achievement for third graders may do an excellent job for this group, but when given to students who speak a language different from that assumed by the developers of the test, it may provide little information about arithmetic achievement. The issue of validating a test is really a matter of gathering evidence to justify the use of a test for a given purpose with a particular population. This is perhaps one of the most critical issues for educational administrators.

There are several ways in which developers attempt to provide evidence of the validity of their measures. These include evidence related to (a) the content of the test, (b) the accuracy with which a test measures some underlying construct, and (c) the relationship of test scores to an external criterion measure.

Construct Validity. There are many psychological constructs which are of concern to educators. These include intelligence, motivation, self-concept, and anxiety. Commercially developed measures of these constructs are not as common in schools as achievement tests, yet they are important. Measures of psychological constructs are usually based on specific theories. Intelligence, for example, may be thought of as a global or multifaceted entity. Intelligence tests based on these two distinctly different theoretical perspectives can be expected to differ considerably. Evidence of construct validity, as usually reported in test manuals, includes a focus on the internal structure of the measure, studies of the relationship of the measure with others, and studies of groups with known characteristics.

Content Validity. Evidence that a test has content validity is based, as one might expect, on the items in the test. The question of content validity concerns the extent to which the

items on a test reflect the type of content and cognitive skills expected by the user. Within the context of content validity, the term *instructional validity* is often used. This is a key idea that poses the question of consistency between what was taught and what is represented in a given test. To the extent that there is a mismatch, a test has limited utility for drawing inferences about student achievement.

Predictive Validity. This type of validity is usually reported for tests designed to predict some future event. Many college admissions tests, for example, report validity coefficients related to their ability to predict freshman grade point averages. Similarly, many aptitude and intelligence tests report validity coefficients, which are indexes of the extent to which the measure can predict valued outcomes such as grade point averages and class rank.

Consequential Validity. A final issue discussed in the context of validity is the notion of consequential validity. Again, building on the dependence of validity on test use, this type of validity is concerned with the consequences of test use. For example, if a student was administered an interest inventory and interpreted the results in terms of what he could or could not do, the results could be said to have a negative impact on the student's aspirations. The general idea is that the use of any measure should be evaluated in terms of its potential impact on those involved.

Test Scores and Norms

In the definition of a standardized test presented above, it was noted that not only was test administration standardized, but the interpretation of results was also standardized. This observation is based on the fact that a point of reference is needed to give a test score meaning, and a uniform point of reference is needed to ensure standardization. For example, it is true that a score of 5 has no meaning in and of itself. To give this score meaning, we could say either that it represented a

perfect performance on a test or that it was higher than the score received by, say, 90% of the other examinees. This distinction is between an absolute and a relative interpretation of test performance.

Criterion, Domain, and Standards-Based Interpretations of Test Scores. A true criterion-referenced interpretation of a test score would involve comparing the score to some fixed performance standard. If passing an exam meant that a student should answer 75% of the items correctly, then the criterion would be 75%. Using this criterion, students could be classified as either "masters" or "nonmasters." During the minimum competency testing movement, most states reported mastery classifications for examinees.

A domain-referenced interpretation of a test score is focused on interpreting the scores relative to the proportion of the domain a student has mastered. If the domain of interest were simple addition, a domain-referenced interpretation of a test score would involve estimating the percentage of items in the domain the student could answer correctly. For example, if a domain was narrowly defined as all the sets of two single-digit numbers (e.g., 1 + 1 = ??), then a sample of items drawn from this domain could be used as a test and the percentage of the items a student could answer correctly on the test could be used to estimate the percentage he or she could answer correctly in the domain.

A standards-based interpretation of a test score is a type of criterion-referenced interpretation, but instead of focusing on, say, the percentage of items in a domain an examinee can answer correctly, the focus is on the level of proficiency examinees are able to demonstrate. Typically, several levels of proficiency are established and, based on examination results, examinees are classified accordingly.

Norm-Referenced Interpretations of Test Scores. Relative comparisons or interpretations of test scores are based on a direct comparison of an individual's performance with some norm

group. Test users must consider the age and appropriateness of the norms before informed use can occur. Outdated norms have been the basis of severe criticisms of standardized achievement tests (see Cannell, 1988). Similarly, inappropriate norms have been used as a basis for charges of discrimination or test bias. Norms can be based on the scores of individuals or groups such as schools. Group norms should be the basis for interpreting group-level statistics (e.g., school means), and individual norms should be the basis for interpreting the performance of individuals.

Relative comparisons are usually accomplished with three broad categories of scores: percentiles, normalized scores, and expectancy scores. A percentile expresses an individual's score relative to the position of other scores in a group. The question answered with a percentile rank is the following: What percentage of the persons in the norm group had this score or lower? The higher the percentile rank of a score, the better the performance.

Normalized scores come under a variety of names: norm-curve equivalents, stanines, T scores, and so on. All share the characteristic that the raw score or number correct is transformed to a score with a given mean and standard deviation. For example, stanines have a mean of 5.0 and a standard deviation of 2.0. They facilitate comparison of an individual's performance across content areas, assuming the norm group is the same.

The final type of scores considered in this chapter is expectancy scores. Broadly speaking, these scores summarize performance relative to the expected standard. Age- and grade-equivalent scores are expectancy scores. A grade equivalent provides information about an examinee's performance compared to others in neighboring grades and provides answers to questions such as whether a student is reading at grade level.

Different Types of Standardized Tests

In the past, when educators discussed standardized testing in schools, they were almost always concerned with a

group-administered examination which consisted of multiple-choice, matching, or true-false items. The chief purpose of the exam was to compare the performance of students, or some aggregate such as a grade or school, to performance trends locally and nationally. Today, what is meant by a standardized test has evolved to include portfolios of student work, examinations based on open-ended items with more than one right answer, samples of products produced by examinees, and even oral presentations and performances. Tests are designed not only to facilitate the comparison of students with others, but to identify specifically what students can and cannot do, and, more important, to foster teaching and learning of valued skills. In this book, I consider two broad categories of tests: intelligence and aptitude tests and achievement tests.

Intelligence and Aptitude Measures

Measures of intelligence or aptitude are typically focused on what examinees have the potential to do. Although there is a tendency to equate performance on these measures with innate, fixed capacities, most in the measurement community readily argue that these measures reflect learned skills which are modifiable. The types of skills measured with intelligence and aptitude tests are generally broader than those reflected in achievement tests and are usually not linked to a specific instructional experience, but instead reflect general skills useful in learning in a variety of contexts and content areas. These skills include (a) memory, (b) pattern analysis, and (c) abstract reasoning. Intelligence tests can be group-administered or individually administered. Group-administered measures are more common in schools because they are relatively inexpensive and do not require a lot of time to administer. In contrast, individually administered measures usually require a trained specialist, can be prohibitively expensive, and may take days to administer. These are usually reserved for questions concerning students with special needs.

Intelligence and aptitude tests are intended to provide additional information about the learning potential of individual

students. Some are designed to yield a single overall measure of ability, whereas others yield multiple and more specific indicators. Measures which produce a single global measure of ability are usually based on a specific theory of intelligence and are useful for predicting achievement in a variety of different learning situations. However, for any given setting (e.g., content area or course) they are likely to have less predictive power than measures which yield scores for multiple abilities.

Intelligence and aptitude tests have many uses in schools. They are used to group students, identify students with special needs, select students for special programs, and aid in counseling and vocational guidance. These applications are problematic when (a) decisions are based solely on the basis of ability measures, and (b) ability measures are treated as fixed and innate rather than as learned behaviors which are affected by environment, motivation, and a host of other factors. Largely because of inappropriate uses, ability measures have been at the center of a continuing controversy in education. The relationship between performance on intelligence and aptitude measures and a student's social background, race, or gender has led to questions of bias and prompted the development of culture-fair tests (for an extended discussion, see Wigdor & Garner, 1982).

Achievement Tests

Unlike intelligence and aptitude tests, achievement tests are designed to reflect a student's performance over a defined content domain. Achievement tests may be multi-battery measures which yield scores on a variety of different content areas, single-battery measures which are focused on a specific content area, or diagnostic measures. Multi-battery achievement measures are common in state assessment programs. Some of the more popular ones include the Iowa Tests of Basic Skills, the Cognitive Test of Basic Skills, and the Stanford Achievement Tests. In addition to producing scores in a variety of content areas (e.g., reading, mathematics, social studies), most have multiple levels which facilitate following a student's academic growth from elementary school to high school

graduation. Multi-battery achievement measures are ideal for drawing conclusions about a student's relative strengths and weaknesses in various content areas.

Multi-battery achievement tests sample a variety of content areas, but to make the exams manageable, each content area is sampled only sparingly. As a consequence, the more specificity that is required about a student's performance, the less satisfactory these measures are. Single-subject measures provide the type of detailed information about a student's performance in a particular subject area that can be used for instructional planning. Diagnostic measures not only provide a level of specificity not found in multi-battery measures, but they also provide information about the enabling skills required to perform certain tasks.

Conclusion

Standardized tests have a long and controversial history of use in schools (for a thorough review, see Walker, 2000). In recent decades, they have been criticized for their perceived negative impact on teaching and learning. Regardless of this controversy, however, they are likely to play an increasing role in education. Because of this fact, it is important that administrators be aware of the basic characteristics of standardized tests. This chapter presented a basic overview of standardized tests. Their construction, technical foundations, and uses in education were discussed. Later chapters provide more detail on issues of test preparation activities.

Sources of Information and Additional Readings

Standards for Evaluating Standardized Tests

American Educational Research Association, American Psychological Association, and National Council on Measurement

in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: APA.

This book presents guidelines formulated by a joint committee of the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. It presents explicit guidelines on test construction, fairness issues, and test application. This document serves as a guide for test developers and those that publish reviews of published tests. It is an essential reference for anyone who may need to evaluate a test.

Critiques of Tests

The *Mental Measurements Yearbook*, published by the Buros Institute, is an annual publication of critiques of some of the most popular tests used in education and psychology. The SilverPlatter service produces a CD version of the yearbook (www.silverplatter.com/catalog/mmyb.htm).

Information About Standardized Tests

Perhaps the best source of information on tests in the information age is the Education Resource Information Center Clearing House on Assessment and Evaluation (www.ericae.net). This site provides links to most large publishers of tests (e.g., Educational Testing Service, Psychological Corporation), reviews of published tests, ratings of state testing programs, a test locator service, and even test selection tips.

National Council for Measurement in Education (www.ncme.org)

This is a primary site for measurement specialists in education. Here you will find discussions of key issues and a wealth of information on specific topics.

Educational Testing Service (www.ets.org)

This is the site for the largest testing company in the world. ETS is responsible for the Scholastic Assessment Test

(SAT) and a host of other examinations. A visit to this site is a must.

Bias and Fairness Issues in Standardized Testing

The National Center for Fair and Open Testing focuses on the fair and equitable use of tests in education, employment, and other settings. Their Web site (www.fairtest.org) has a wealth of information for anyone with concerns about test use.

Additional Readings

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching*. (8th ed.). Columbus, OH: Merrill. This is a shorter work than the others, but very informative.

Rudner, L. M., Conoley, J. C., & Plake, B. S. (1989). *Understanding achievement tests: A guide for school administrators*. Washington, DC: The ERIC Clearinghouse on Tests, Measurements, and Evaluation. This is a brief and concise guide which can be a good start.

Thorndike, R. M. (1997). *Measurement and evaluation in psychology and education* (6th ed.). Columbus, OH: Merrill. This is an excellent source or follow-up to the material presented in this text. The coverage is comprehensive and user friendly.

Worthen, B. R., White, K. R., Fan, X., & Sudweeks, R. R. (1999). *Measurement and assessment in schools* (2nd ed.). New York: Longman. This is an excellent guide for the administrator, with a good outline for a school testing program.