Van Blerkom Chapter 16: Characteristics of Standardized Tests

Standardized tests are useful tools for providing unbiased assessments for many students. Scores from standardized tests usually employ either norm- or criterion-referenced interpretations. Standardized tests have been developed to measure a wide range of traits, but the most common in educational situations are aptitude (how well students may learn) and achievement (how much students have learned). More detailed treatment of these two is presented below.

1. Distinction Between Aptitude and Achievement Tests

(a) Purpose

**Aptitude** tests are designed to measure one’s learning potential. Results of such test may prove useful in a variety of situations, such as with decisions regarding admissions, instructional groupings, and other instructional techniques. **Achievement** tests enable one to demonstrate current knowledge with a defined set of skills. Scores from these tests are useful for indicating one’s strengths and weaknesses relative to others and relative to defined criteria.

(b) Skills Measured

Despite their stated goals, aptitude tests are limited to measuring achievement or learned skills, just like achievement tests. However, unlike achievement tests, aptitude tests typically focus on general skills that have been acquired over a long period of time whereas achievement tests usually focus on measuring specific skills that have been acquired relatively recently. Thus, while an aptitude test may measure general problem-solving ability, an achievement test would focus more on a well-defined problem-solving skills such as the application of a specific mathematical formula. A premise of aptitude tests is that one’s prior learning predicts well one’s subsequent learning.

(c) Validity and Reliability

c1. Validity

It is important that achievement tests have a strong tie with instructional objectives since they are designed to measure one’s status regarding a specific set of skills. Thus, one must first establish that an achievement test has content-related evidence of validity. Since achievement tests are usually developed to be administered nationwide, a particular test may not be appropriate for some settings due to incongruencies between the specific setting (difference instructional objectives) and schools nationwide.

Since aptitude tests are designed to predict future learning and achievement, the most relevant evidence of validity would be criterion- and, more specifically, predictive-related. Ideally, one would administer an aptitude tests, allow an appropriate length of time to elapse so the measured learning could occur, collect scores on the measured phenomenon, and then correlate these scores with scores obtained from the aptitude test. Should a strong (positive) correlation result, then one has predictive evidence of validity for the aptitude test.

c2. Reliability

Both achievement and aptitude tests should have high internal consistency for each construct or trait measured. KR formulas or Cronbach’s alpha may be used to calculate internal consistency. Since aptitude tests are designed to predict future learning, scores from these tests should be consistent across time. Test-retest methods may be used to further establish reliability for aptitude tests. Finally, note that most all standardized achievement and aptitude tests have alternate forms. It is important that these forms provide consistent scores that agree, so alternate-forms reliability should also be established and the intraclass correlation coefficient (ICC) would be a good measure of alternate-forms reliability since it can assess agreement.

2. Aptitude Tests

Below are common aptitude tests found in educational situations.

(a) Reading readiness tests

These are primarily used to learn whether students have mastered the requisite skills for reading. Results of these tests are typically used to assign students to appropriate learning environments and remedial activities.

(b) Intelligence and scholastic tests

All intelligence and scholastic tests are limited to measuring behaviors and skills that are learned; such tests do not measure innate and static abilities. As a result, scores from such tests are influenced by an individual’s ability and opportunity to learn the skills being measured by such tests. Often scores from these tests are used to place students in developmentally appropriate learning situations. The primary role of these tests is to predict future achievement.

(c) Multiple-ability tests

Like intelligence and scholastic tests, multiple-ability tests are designed to predict future achievement. While intelligence tests typically provide a summary measure of aptitude, multiple-ability tests provide scores on several distinct traits such as spatial relations, verbal acuity, mathematical acuity, abstract reasoning, etc.

(d) College admission tests

Tests, like the SAT, ACT, and GRE, are designed to predict academic success in college. It is important that these tests demonstrate predictive validity. Note that predictions based on scores from these tests are clearly less than perfect, but it is also clear that scores from these tests do have a reasonable level of predictive validity, and this is further enhanced when combined with other predictors (e.g., high school GPA).

3. Achievement Tests

(a) Diagnostic tests

These tests are usually designed to pinpoint requisite reading or mathematics skills students lack. Sometimes these tests lack strong evidence of validity, yet when administered one-on-one, useful information can be derived from these tests.

(b) Achievement batteries

These tests typically are composed of a series of single-area tests (e.g., mathematics, vocabulary, etc.). Batteries provide norms that enable one to compare students’ performances across subjects. It is important, however, that the objectives for which the test was designed be the same objectives used by the school in which the test is administered. Without this correspondence, scores from achievement batteries will be of little use.

(c) National assessment of educational progress

This is a test and research project in which numerous schools and students participate annually. The goal of NAEP is to assess the position of US students in ten areas including science, writing, and mathematics. NAEP differs from other achievement tests in that it is seldom uses students from the same school (or the same school) twice. The goal is to obtain a representative sample of students and schools from across the country and thereby eliminate any possible testing or preparation effects in student scores.

4. Evaluating Domain and Norm Specifications of Standardized Tests

When providing test scores, it is important that a clear reference is made to the skill being measured or the population with whom the score is being compared. This issue is discussed below.

(a) Criterion-referenced interpretations: Need for Clarity and Relevance of the Domain Specification

Standardized tests that are based on criterion-referenced score reporting should make very clear the skill that is being measured. Stating, for example, that the skill being measured is mathematics is not useful. The goal is to enable one to say, based upon the student’s score, what the student can and cannot do. Thus, simply stating that the score represents performance in mathematics is not enough; a better statement would indicate precisely the areas in mathematics that were assessed, such as division, multiplication, etc. As previously noted, when standardized tests are criterion-referenced, it is important that the skills being measured match well the objectives stressed for those students tested.

(b) Norm-referenced interpretations: Need for clarity and relevance of the norm group

Clear interpretation of norm-referenced scores requires that a clearly defined group of students from which the norm is based, and an understanding of the standard score used to report the performance. To ensure a clear understanding of the norm group, publishers should include the following: (a) clear description of who participated in the norm group (e.g., 48% male eighth-graders predominately from the south with an average level of SES), (b) norms for each group to whom the test taker will be compared (e.g., all students and gifted students), and (c) how the norm group was selected, their participation rate, and the date in which they were administered the test.

Self-Test: Characteristics of Standardized Tests

Item 1 through 8 concern distinctions between achievement and aptitude tests. Indicate whether each statement pertains to:

1. achievement tests.
2. aptitude tests.
3. both achievement and aptitude tests.
4. The purpose of these tests is to certify what students can do.
5. The purpose of these tests is to estimate how students will perform in the future.
6. These tests always measure what students have learned.
7. These tests tend to measure skills taught in the classroom.
8. These tests tend to measure skills learned outside the classroom.
9. Criterion-related evidence of validity is more relevant for these tests.
10. Content-related evidence of validity is more relevant for these tests.
11. Publishers of these tests should report parallel-form reliability when more than one form of the test has been developed.

Items 9 through 13 pertain to characteristics of standardized achievement and aptitude tests. Indicate (yes or no) whether each statement is accurate.

1. Reading readiness tests measure students’ willingness to learn to read.
2. Intelligence tests are best thought of as measures of innate ability.
3. To be useful, reading diagnostic test must provide more than one score.
4. The major advantage of an achievement battery over individual tests is that test scores in different content areas can be compared to a single norm group.
5. Content validity is more relevant to single-subject tests than achievement batteries.

Items 14 through 17 pertain to the adequacy of domain and not specifications. Indicate (yes or no) whether each statement is accurate.

1. A criterion-referenced test has established a well-defined domain if the passing score is clearly stated.
2. One is more able to generalize performance on a criterion-referenced test to a broad rather than narrowly focused domain.
3. Norms are superior if based on a large, generally described sample of examinees as opposed to a smaller, concisely described sample of examinees.
4. If norms are expressed as grade equivalents, it is appropriate to administer an achievement test at a time in the school year other than when the test was normed.