Key Elements of Testing

Types of Standardized Testing

Standardized testing serves a variety of purposes. Some tests are designed to compare student scores to scores from a representative sample of students. Students are evaluated by how well they perform on the specified content (e.g., reading) as compared to the representative sample. Standardized norm-referenced tests are an example of this kind of test.

Other tests are designed to evaluate how well students demonstrate mastery of the specified content (e.g. algebra 1). Criterion-referenced tests are an example of this kind of test.

Both kinds of tests provide information about the academic knowledge and skills of the students tested. The power of standardized testing lies in the fact that students take the test under comparable conditions and that the tests are scored in a comparable manner. Therefore, the student scores can be compared either to each other or to the achievement of curricular objectives or standards.

- **Norm-referenced testing**. A standardized, national norm-referenced test compares a student's score to scores from a sample of students selected to be representative of the nation as a whole. When a student scores at the 62nd percentile, it means the student's score was equal to or better than 62 percent of the students in the norming sample who took the same test under the same standardized conditions. Norm-referenced tests are not designed to measure student knowledge of a specific curriculum or instructional program. Examples of norm-referenced tests include the Stanford Achievement Test, Ninth Edition (Stanford 9), California Achievement Test, Sixth Edition Survey (CAT/6 Survey), and the Spanish Assessment of Basic Education, Second Edition (SABE/2).

- **Criterion-referenced testing**. A criterion-referenced test is used to determine how well students have learned specific information they have been taught. To assess specific curricular or instructional objectives, items need to be developed that are aligned to specific content and/or academic standards.

  In California, the California Standards Tests (CSTs) are considered criterion-referenced tests. The California academic standards are the basis for these tests. Results are reported according to performance levels that show a student's achievement of the standards.

Key Testing Terminology

Standardized Testing Conditions

It is important that testing procedures are conducted according to certain rules and specifications. For example, each student needs to be administered the same or comparable sets of questions, given the same directions, and allowed the same time frame to complete the test. When testing conditions are standardized, the scores can reliably evaluate the extent of student academic progress.

Item Formats

Test items are the questions used to elicit student responses. Aggregating item responses produces a test score. Test items take two formats: selected response and constructed response.

- **Selected-response items**. Selected-response items require students to select an answer to an item prompt. Selected-response items include true/false, matching, and multiple-choice items. Multiple-choice items have a correct answer (i.e., from a list of alternatives) and several wrong answers (or distractors).

- **Constructed-response items**. Constructed-response items require students to write a response to a prompt. Constructed-response items range from supplying a missing word (in a sentence) to writing an extensive essay. The students are directed to demonstrate what they know in their own words.

Selected-response items (e.g., multiple-choice items) are easier and less expensive to develop and score. Students can generally answer them more quickly than constructed-response items (e.g., essays). In the same amount of testing time, multiple-choice items can cover a broader range of the curriculum than essay items. However, multiple-choice items constrain students to a single appropriate answer and are subject to guessing. Constructed-response items allow students to demonstrate more in-depth understanding of content with less likelihood of guessing. Unlike multiple-choice items, scoring constructed-response items can be expensive and time consuming.

Testing Administrations

Before any state, school district, or educational institution administers a test, the test purposes should be clear. Test purposes tend to focus around who is to be tested and the use of scores. Once the objectives have been defined, the specific administration can be determined. When considering whom to test, there are two common strategies:

- **Census testing**. Once the student population has been defined, every student in the population is tested (unless there are handicapping conditions that make testing impossible for some students).

- **Selected sample testing**. Once the student population has been defined, a sample of students is tested. The sample can be used to generate group level scores. For example, the National Assessment of Educational Progress (NAEP) tests a small percentage of California students to obtain a state level score. The sample also can be used to generate individual scores for part of the population. For example, the California Physical Fitness Test (PFT) tests all students in grades 5, 7, and 9.

When considering the use of scores, there are three methods to consider:

- **Individual student testing**. When the purpose of the test is to produce individual scores, students need to be administered all of the test items. Individual student testing includes norm-referenced testing (i.e., to compare individual student scores to scores from the norming sample), criterion-referenced testing (i.e., to evaluate student mastery of the curriculum), and diagnostic testing (i.e., to evaluate individual academic needs). These three types of tests require test scores to be compared to scores of a representative sample or to achievement of curricular goals or objectives. The easiest way to ensure comparability is to administer the same set of items to all test takers. The population tested can either be a sample or a census depending on the purpose of the test scores.

- **Matrix testing**. When the purpose of the test is to generate group level scores, students do not need to be administered the same set of items. Instead, students can be administered a sample of all items on the test. That is, a test consisting of many items is divided into a number of short tests. Each student takes one short test. Student performance on each of the short tests is aggregated to produce a group level score (e.g., a school level score). The population tested can either be a sample or a census depending on the purpose of the test scores. Matrix testing does not provide individual scores because students are given too few items to generate a reliable score and it is difficult to create comparability across different forms of the test.

- **Partial matrix testing**. To overcome the limitation in matrix testing (i.e., no individual scores), partial matrix testing offers a compromise. In this design, a set of core items is administered to all students, and other items are matrix sampled across forms. The benefit of this design is that it provides individual scores and more valid and reliable group level scores.

Measurement Principles

Validity

Validity refers to the degree to which evidence and theory support the interpretations of test scores. Interpretation is dependent on the proposed uses of test scores. It is the interpretation of the test score and how it is used that is validated, not the test.

Establishing test validity is an ongoing process and entails gathering different kinds of evidence. For example, evidence of content validity relies on subject-matter experts to review test items to ensure that items accurately measure the content. Evidence of criterion validity relies on the relationship between different test scores that measure the same content. If there is a strong, positive relationship (i.e., correlation) between scores on two different tests designed to measure Algebra I, for example, it is considered one source of evidence that both tests are valid measures of Algebra I.

Reliability

Reliability is an indicator of the extent to which scores are consistent across different administrations and/or different scorers. It is the test score that is or is not reliable, not the test. Reliability also is a general term used to describe measurement error. Error is the difference in scores from the same test (or parallel forms of the test) that has been given to the same student many times. This assumes that a student takes the same test (or parallel forms of the test) and forgets each testing occurrence many times. Practically, this cannot be done, but there are statistical methods that estimate this difference in test scores. All tests have measurement error.

Fairness

Test scores are fair when they yield score interpretations that are valid and reliable for all students taking the test. Regardless of race, national origin, gender, or disability, academic tests must measure the same knowledge of content for all students who take the test. Test scores must not systematically underestimate or overestimate the knowledge of members of a particular group.

Test Development Process

Test development is “the process through which a test is planned, constructed, evaluated, and modified.”[1] This process typically includes the following steps:

Defining the Purpose

It is important, first, to define and document the purpose of the test, the characteristics of the population to be tested, and the intended uses of test results (i.e., instructional implementation, standards achievement, etc.).

Developing Test Specifications

The test specifications define test content, including the proposed number of items, item formats (e.g., constructed response or selected response), the desired psychometric properties of test items, and other relevant information.[2] Test specifications can be presented in blueprints that specify the content and skills to be included in the test.

Developing and Field Testing Items

Item Writing

Items should be developed using clearly defined, common sense rules. Selected-response (e.g., multiple choice) items, for example, must be accurate, valid, and clear, with options that are plausible to students not well grounded in the subject matter. Constructed-response prompts also must be clear, based on subject matter students have been taught, and be fully answerable within the time frame allotted for the response.

Test Directions

The item writing process also should include development of test directions. Good test directions ensure that the items measure the students’ skills and content knowledge. Directions should include practice items, suggestions for allocating time, advice about guessing, and information about test-taking strategies.[3]

[1] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington D.C. American Educational Research Association.

[2] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington D.C. American Educational Research Association.

[3] Millman, J. & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.) *Educational measurement*. (3rd ed.). New York, National Council on Measurement in Education, American Council on Education, & McMillan Publishing Company.

Item Tryouts and Field Tests

After items are written, they must be evaluated. Item evaluation can be based on information gathered from preliminary tryouts that identify unclear wording or directions, inappropriate timing, and/or item difficulty levels. Formal field tests are conducted to provide more detailed technical data on item quality. Field testing strategies may include embedding new items in an operational test or conducting a separate tryout. All items used in California assessments have been field tested.

Item Analyses

Item analyses are conducted to evaluate the psychometric properties of items. The two properties most often examined are item difficulty and item discrimination. Item difficulty can be determined by discerning the percentage of students that answered an item correctly. If the p-value is .80 (80% of students answer an item correctly), the item is considered easy. If the p-value is .20, the item is considered difficult. Item discrimination refers to how effectively each item differentiates between students who know most about the content area being tested and those who know least. For example, students with the highest scores should generally get hard items correct while students with low scores generally will not.

Item Selection

The basic rule for item selection after field testing is to retain items that meet evaluation criteria and have adequate psychometric properties as detailed or defined in the test specifications. Content Review Panels (CRPs) consisting of experts in the content area and experts in item development conduct item evaluation. A Statewide Pupil Assessment Review (SPAR) panel reviews items for sensitivity issues.

Assembling the Test

The final step in test development is production of the test itself. Considerations in this stage include how items should be ordered and grouped, how items will look on a page, how the test should be printed, and how test security should be maintained during storage and shipping.

Attachment I

Test Measurement Principles

Questions about Appropriate Test Use

In order to determine if a test is being used appropriately, the following questions should be addressed:

- What is the purpose of the test?

- Is there adequate evidence of **validity**, demonstrating that test scores are accurate and meaningful?
  - Is there evidence that the test is accurately measuring the **desired knowledge and skills**?
  - Are the test results valid for **the stated purpose and in the particular setting** where the test is to be administered?
  - Are the test results valid for the **specific groups of students** taking the test?

- Is there **reliability** in the test scores, demonstrating that test error has been minimized, yielding the same results with repeated administrations?
  - Are test results consistent?
  - Is the level of measurement error for the test sufficiently small that misclassifying students based on the test is inconsequential?

- Are the **conclusions** drawn from test scores **fair to all students**?
  - Are the conclusions drawn from the **test results accurate for all students**, when it is possible that the reliability across sub groups may vary substantially?
  - Does the evidence indicate that the test is measuring the same content for all students?
  - Is there solid evidence that the test results do not systematically **underestimate or overestimate** the knowledge or skills of **members of any particular group**?

- Have cut scores been established for performance levels that will provide accurate and meaningful information for all students?