**Glossary Terms from Standardized Testing Primer**

*Author:* Richard P. Phelps

*Publisher:* Peter Lang Publishing, New York, 2007

**Chapter 1**

**Content standards**- a predetermined, and sometimes legally mandated, body of subject-matter content.

**High stakes test**- used to provide results that have important, direct consequences for examinees, programs, teachers, or institutions.

**Item prompt**- the question, stimulus, or instruction that directs the test taker’s response.

**Latent trait**- a trait that is not directly measurable or observable (e.g., knowledge, problem-solving skill, empathy).

**Low stakes test**- a test used to provide results that have only minor or indirect consequences for examinees, programs, or institutions involved in the testing.

**Multiple-choice item**- a type of selected-response item that requires the test taker to select a correct response to a question (or, prompt) from among a small number of specific choices.

**Norm-referenced test (NRT) -** designed to measure a test-taker’s traits relative to a norm group.

**Performance-based test-**requires the respondent to construct a response, demonstrate a skill, or follow a procedure. Examples include answering open-ended question, conversing in a language, solving a mathematics problem while showing all calculations, writing an essay, or conducting a science experiment.

**Psychometrics**- the science of mental measurement. Experts in mental, or psychological, measurement are psychometricians.

**Representative sample**- a sample is a subgroup of a population. A sample is representative if it accurately reflects the character of the population in the aspects under study.

**Standardized test**- If any aspect of a test- format, procedures, or administration- is uniform across test takers, it can be considered a standardized test.

**Standards-based test-** achievement test designed to cover a specified content domain that is usually identified by content standards.

**Teaching to the test-** interpreted liberally, the phrase could refer to any instruction on subject matter that is covered by a test. But, teaching to the test is a genuine problem only when test security is lax and teachers have access to the exact or approximate content of an upcoming test. Otherwise, any teacher not teaching the same content domain from which a test has been developed is behaving irresponsibly.

**Chapter 2**

**Achievement (or, educational achievement) -** the accumulated knowledge and skills acquired from previous learning.

**Achievement test**- designed to measure what had been learned, rather than to predict future performance.

**Age**-equivalent score- the chronological age in a defined population for which a given score is the median (middle) score.

**Aptitude test**- designed to predict how much you might learn in the future.

**Aptitude**-potential for future learning.

**Assessment**-generally refers to large-scale, system wide measurement programs for pupil diagnosis, program evaluation, accountability. resource allocation, or teacher evaluation.

**Back translation**- a second and independent translation of a test from a second language to a first language that follows an earlier translation from the first, original language of the test to the second language; the degree to which a back translation matches the original test provides a measure of the accuracy of the original translation.

**Criterion-referenced test (CRT) -** a test that allows its users to interpret scores in relationship to a functional performance level, for example, the degree of competence attained by a particular student, without reference to the performance of others.

**Diagnostic test**- its purpose is to ascertain prior to instruction, a student’s abilities and level of achievement so that instruction might be adjusted appropriately.

**Educational attainment-** the highest grade, year, or level of school attended and completed.

**Enacted curriculum**- the subject matter content that actually is taught in the schools regardless of the original intentions.

**End-of-level-test-** test administered at the end of an educational level, such as primary of secondary.

**Intelligence Quotient (IQ) -** a unit for expressing the results of an intelligence test, based on the ratio of an individual’s mental age (MA), as determined by the test. To their chronological age (CA): IQ= MA/CA x 100.

**Intelligence Test**- test designed to measure an individual’s level of cognitive functioning not related to any particular subject matter.

**Intended curriculum**- the subject matter content that is planned for school instruction, according to legislated and/or published standards.

**Mean**- the arithmetic average of a set of scores (i.e. the sum of a set of scored divided by their number).

**Medium stakes test-** partial or conditional consequences that are defined in law or regulations to result from exceeding, or not, one of more score thresholds.

**Mental age**- a unit for expressing the results of intelligence tests based on comparing the individual’s performance on the test with the average performance of individuals in a specific chronological age group.

**Minimum competency test**- a high-stakes test that requires performance at or above a single threshold test score before certain education attainment will be recognized.

**Norms**- the results of a test administered to a “norm group”- a representative sample for a larger population of interest.

**Performance standards**- specify the level or levels of achievement expected.

**Power test**- test administered with no effective time limit.

**Raw score-** the unadjusted score on a test, perhaps the simple number of correct responses, with no accounting for variations in item difficulty or weighting.

**Scale score**- the result of the mapping of a raw test score onto a more meaningful, interpretable, or comparable scale.

**Selection**-choosing individuals based, in whole or in part, on their test score (e.g., for university admission, a scholarship, employment).

**Speeded test**- test administered within a time limit such that some test takers may not finish.

**Standard deviation**- the square root of the average squared deviation around the mean (i.e., the variance). It is used as a measure of variability in a distribution of scores.

**Summative assessment**-when achievement tests are used to summarize the accumulation of knowledge up to that point.

**System wide test**- any test that is administered to all students or to a representative sample of all students within a jurisdiction for at least on grade level.

**Chapter 3**

**End-of-course test**- test administered at the end of a course of study, such as that for algebra1 of chemistry.

**Item rotation**- for tests that are administered periodically, the process of replacing some number of already-used test items with others that have similar content and statistical characteristics.

**Lake Wobegon effect**- score inflation or artificial test score gains.

**Mastery testing-**when students are tested periodically for diagnosis as many times as necessary to prove master of the material. In mastery learning regimes, each student may progress at a different pace, but ends up in the same place, mastering the material.

**Predictive validity**- correlation between test scores and future performance.

**Score inflation**- a rise in test scores over time that is not caused by a genuine increase in the quality being measured, such as academic achievement.

**Standards for Educational and Psychological Testing- Produced by the APA, NCME, and AERA**; have become the de facto regulations governing legal uses and administration of most standardized tests in the United States.

**Test score pollution**- term coined by Haladyna, Nolen, and Haas (1991), derived from earlier work by Messick (1984), that “refers to factors affecting the truthfulness of a test score interpretation. Specifically, pollution increases or decreases test performance without connection to the construct represented by the test, producing construct-irrelevant test score variance.”

**Unintended consequences-** another way of saying that standardized testing may portend adverse effects (e.g., students who score poorly might get hurt feelings, or be held back a grade.)

**Washback effect**- when some aspect of a test given at tone grade level has an effect on school, teacher, or student behavior in earlier grades (a.k.a. backwash effect).

**Chapter 4**

**Accommodation**- a modification of test administration format and procedure devised to accommodate students with disabilities that render them unable to take a test fairly under standard format and procedure.

**Adaptive testing**- a sequential form of individual testing in which successive items, or sets of items, in the test are chosen based primarily on their psychometric properties and content, in relation to the test taker’s responses to previous items.

**Anchor test-** provides information about the equivalence (or not) of two different tests or test forms by measuring more general skills, or something else that is common to both tests. Anchor test data can be used to translate between the other two tests.

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

**Classical test (i.e., true score) theory**- developing a test this way, every test is custom-designed and relevant to a particular population.

**Computer-adaptive testing (CAT) -** test-takers are presented an item at a level of difficulty determined by their performance on the previous item. For example, correct responses may yield more difficult subsequent items, and vice versa.

**Computer-based testing (CBT) -** any testing administered by computer, with item selection and order determined either serially or by an adaptive (CAT) algorithm.

**Concurrent validity-**a form of criterion validity in which the test and the criterion are administered at the same point in time.

**Consequential validity**- a correlation between test scores and intended outcomes.

**Construct equivalence**- the extent to which the construct measured by one test is essentially the same as the construct measured by another test.

**Construct irrelevance**- the extent to which test scores are influenced by factors unrelated to the test’s intended construct.

**Construct validity**- indicates that the test scores are to be interpreted as indication the test taker’s standing on the construct measured by the test.

**Constructed**-**response item**- type of performance item for which the test taker is required to compose a response to a question or prompt. Types of constructed-response items include fill-in-the-blank, short answer, and essay.

**Construct**- the quality or concept a test is designed to measure.

**Content validity**- the extent to which the content of a test is representative of the conceptual or content domain it is designed to cover. Content validation is a process used to establish the meaning of a test through a series of studies.

**Criterion validity**- the extent to which the content of a test is representative of the conceptual or content domain it is designed to cover. Content validation is a process used to establish the meaning of a test through a series of studies.

**Curricular validity**- is strong if a test contains questions based on the content of the curriculum and weak if a test contains question not based on the content of the curriculum.

**Cut score**- a threshold, a test score that distinguishes between the scores above and below, as does the minimum score required to obtain a diploma.

**DIF (differential item functioning) -** the tendency of an item to show different statistical characteristics with different demographic groups.

**Distracters**- of the responses a test taker may choose in a selected-response (or, multiple choice) item, the ones that are incorrect.

**Equating**-the process of statistically weighting two versions of a test, or two test forms, so that they are equal in difficulty.

**Field test**- a practice, not an operational, test involving volunteers with characteristics similar to those of the population that will be tested.

Higher-order thinking- a grab-bag of cognitive processes alleged to be related to creativity, such as lateral thinking and meta-analysis. It is sometimes asserted that standardized tests cannot test higher-order thinking. More often, it is asserted that standardized tests with multiple-choice response formats cannot test higher-order thinking (but open-ended response formats can).

**Individual Education Plan (IEP)** - schools must devise an academic program uniquely appropriate to each student who has been identified with a disability that impedes learning.

**Item analysis-** the set of methods used to evaluate test items; the most common techniques involve measurement of item difficulty and discrimination.

**Item characteristic curve-** a graph (or function) of a test item with the measured trait/ability/ attitude represented by the x axis and the probability of correct responses represented by the y axis.

**Item difficulty**- the tendency of an item to receive incorrect responses.

**Item discrimination-**refers to how effectively each item differentiates between examinees who know most about the content area being tested and those who know least.

**Item exposure-** the degree to which test items are publicly revealed; conversely the number of times a test item has been used in periodical test administrations, risking exposure.

**Item pool-** the population of test items from which a testing organization may choose when assembling a test form.

**Item response** (i.e., latent trait) theory (IRT) - By comparison with classical test theory, IRT offers “item-free test person measurement” and “person-free item measurement.”

**Matrix testing**- 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. school score, state score).

**Measurement error**- the component of an observed test scores that is not the true score of the quality that one wishes to measure.

**Operational test**- a regular, genuine test administration, with the intended test population taking a test “that counts” (i.e., not a field test).

**P value**- the percentage of test takers who respond with the correct choice.

Parameter- an item characteristic curve can be identified by one (item difficulty), two (item discrimination), or three (base value for guessing) parameters. .

**Point-biserial correlation**- the correlation between test takers’ scores on the whole test and their cores on an item.

**Projective test-** presents an ambiguous stimulus and requests the test take to describe or explain. The assumption is that people respond by projecting their own inner thoughts, feelings, fears, or conflicts onto the stimuli. The Rorschach Inkblot Test and the Thematic Apperception Test are probably the most famous.

**Rasch model-** a one-parameter IRT model, named for the Danish statistician, Georg Rasch.

**Reliability**- refers to the degree to which test scores are consistent across time, conditions, and test-takers.

**Selected-response format**-requires a test taker to select a response to a question or prompt. Types of selected-response items include multiple-choice, true-false, and matching.

**Standard error**- the error of an estimate due to sampling, based on the number of observations and their distances for the mean.

**Test administration**- the act of giving a test.

**Test blueprints**- serve much the same purpose as architects’ blueprints. They provide more detail than the specs, showing how the different pieces of a test will fit together.

**Test form-** a version of a test with a particular combination of items. With large-scale tests, there can be two to many forms developed, for use in the same administration (to prevent answer copying) or in subsequent administrations (to reduce item exposure).

**Test format**- refers to the form and content of a test in its administration (e.g., paper-and-pencil, computer-based)

**Test frameworks**- are outlines of test content, detailed enough to be read meaningfully by item writers.

**Test items-** consist of questions or prompts designed to elicit an answer or response and, perhaps, a set of possible answers.

**Test specifications**- are much like the “specs” for a construction or design project. The idea is to set the parameters within which developers will work.

**True score**- in practice, an individual’s or group’s long-term average score; in theory, the score that would be obtained if measurement error were absent.

**Validity-** the extent to which a test measures the quality it purports to measure; there are several kinds (e.g., content, construct, criterion, consequential, predictive). Unlike reliability, validity can sometimes be difficult to measure precisely.

**Chapter 5**

**Gain score**- the difference between a test taker’s score, or a cohort’s average score, from one test administration and another, later administration of the same test.