Glossary of Terms:

Testing & Validation

Glossary: Testing & Validation

Adverse Impact—A substantially different rate of selection in hiring, promotion, or other employment decision that works to the disadvantage of members of a race, sex, or ethnic group.

Angoff Ratings—Ratings that are provided by SMEs on the percentage of minimally qualified applicants they expect to answer the test item correctly. These ratings are averaged into a score called the "unmodified Angoff score" (also referred to as a "Critical Score").

Critical Score—The score level of the test that was set by averaging the Angoff ratings that are provided by SMEs on the percentage of minimally qualified applicants they expect to answer the test items correctly.

Cutoff Score—The final pass/fail score set for the test (set by reducing the Critical Score by 1, 2, or 3 CSEMs).

CSEM—Conditional Standard Error of Measurement. The SEM at a *particular score level* in the score distribution (see SEM definition below).

DCR—Decision Consistency Reliability. A type of test reliability that estimates how consistently the test classifies "masters" and "non-masters" or those who pass the test versus fail.

DIF—Differential Item Functioning. A statistical analysis that identifies test items where a focal group (usually a minority group or women) scores lower than the majority group (usually whites or men), after matching the two groups on overall test score. DIF items are therefore potentially bias or unfair.

ETS—A person's true score is defined as the expected number-correct score over an infinite number of independent administrations of the test

Item Difficulty Values—The percentage of all test takers who answered the item correctly.

Job Analysis—A document created by surveying SMEs that includes *job duties* (with relevant ratings such as frequency, importance, and performance differentiating), *KSAPCs* (with ratings such as frequency, importance, performance differentiating, and duty linkages), and *other relevant information* about the job (such as supervisory characteristics, licensing and certification requirements, etc.).

Job Duties—Statements of "tasks" or "work behaviors" that describe discreet aspects of work performance. Job duties typically start with an action word (e.g., drive, collate, complete, analyze, etc.) and include relevant "work products" or outcomes.

KSAPCs—Knowledges, skills, abilities, and personal characteristics. Job knowledges refer to bodies of information applied directly to the performance of a work function; skills refer to an observable competence to perform a learned psychomotor act (e.g., keyboarding is a skill because it can be observed and requires a learned process to perform); abilities refer to a present competence to perform an observable behavior or a behavior which results in an observable

product (see the Uniform Guidelines, Definitions). Personal characteristics typically refer to traits or characteristics that may be more abstract in nature, but include "operational definitions" that specifically tie them into observable aspects of the job. For example, *dependability* is a personal characteristic (not a knowledge, skill, or ability), but can be included in a job analysis if it is defined in terms of *observable aspects of job behavior*. For example: "Dependability sufficient to show up for work on time, complete tasks in a timely manner, notify supervisory staff is delays are expected, and regularly complete critical work functions."

Outlier—A statistical term used to define a rating, score, or some other measure that is outside the normal range of other similar ratings or scores. Several techniques are available for identifying outliers.

Point Biserial—A statistical correlation between a test item (in the form of a 0 for incorrect and 1 for correct) and the overall test score (in raw points). Items with negative point biserials are inversely related to higher test scores, which indicates that they are negatively impacting test reliability; positive point biserials are contributing to test reliability in various levels.

Reliability—The consistency of the test as a whole. Tests that have high reliability are consistent internally because the items are measuring a similar trait in a way that holds together between items. Tests that have low reliability include items that are pulling away statistically from other items either because they are poor items for the trait of interest, or they are good items that are measuring a different trait.

SEM—Standard Error of Measurement. A statistic that represents the likely range of a test taker's "true score" (or speculated "real ability level") from any given score. For example, if the test's SEM is 3 and an applicant *obtained* a raw score of 60, his or her true score (with 68% likelihood) is between 57 and 63, between 54 and 66 (with 95% likelihood), and between 51 and 69 (with 99% likelihood). Because test takers have "good days" and "bad days" when taking tests, this statistic is useful for adjusting the test cutoff to account for such differences that may be unrelated to a test taker's actual ability level.

SME—Subject-matter expert. A job incumbent who has been selected for providing input on the job analysis or test validation process. SMEs should have at least one-year on-the-job experience and not be on probationary or "light/modified duty" status. Supervisors and trainers can also serve as SMEs, provided that they know how to perform the target job.