KNAPP
Our experience. Your success.

"Cheat Sheet" on Psychometric Terminology*

Criterion-referenced passing score. A cut score that has been set using a process that links the performance standard for the examination (the pass point) to criteria for acceptable practice of the occupation/profession. Commonly used methods for setting criterion-referenced passing scores include the modified Angoff, Nedelsky, Ebel, and Bookmark methods.

Discrimination. The extent to which performance on an item correlates with performance on the total examination. Indicates the degree to which an item differentiates between those who have the required knowledge, skill, etc. and those who do not. A discrimination of .20-.25 or above is generally considered to be acceptable for an occupational/professional credentialing examination. The point-biserial correlation coefficient is the most commonly used classical discrimination statistic. Under IRT, the *a* parameter the measure of discrimination, with steeper slope indicating better discrimination.

Equating. A statistical procedure that identifies and compensates for variations in difficulty among different forms of the same examination. Helps to ensure that examinees will not be unfairly advantaged or disadvantaged based on the examination form they take (e.g., examinees who take an easier examination form will NOT have a greater chance of passing the examination).

Examination specifications (also test blueprint). The specifications delineate the content (e.g., knowledge, skills, competencies) to be covered by the assessment and the relative weighting of the different content areas on the examination. May also specify the types of items to be used on the examination and the proportion of the assessment that should be represented by each item type.

Item. A statement, question, exercise, or task on an examination for which the examinee is to select or construct a response, or perform a task.

Item difficulty. An estimate of the difficulty of an item. The percentage or proportion of examinees responding correctly to an item (also referred to as the *p*-value), is the classical statistic for item difficulty. Under IRT, the *b* parameter estimates item difficulty on a scale (referred to as "theta") ranging from -3.0 to +3.0. Positive values on the scale represent more difficult items.

** There are two primary approaches within measurement theory: classical theory and item response theory (IRT). The statistical analyses used under each theory to evaluate the quality of assessments are different and have been noted, where relevant. To better understand the definitions provided here, consult with your psychometrician regarding the type of statistics used for your assessment.*

Inter-rater agreement. The consistency with which two or more judges rate the work or performance of examinees (sometimes referred to as inter-rater reliability).

Job analysis (also practice analysis, job/task analysis). A general term referring to the investigation of job roles, occupations, or professions to identify job duties and tasks, responsibilities, necessary worker characteristics (e.g., knowledge, skills and abilities), working conditions, and/or other aspects of work. A job analysis provides evidence to support the validity of the examination (assuming that the examination specifications have been linked to the findings of the job analysis).

Psychometrics. The science of measuring mental processes and activities. It includes the theory underlying the development of sound measuring tools (i.e., assessments), as well as the statistical analyses used to evaluate whether these tools are accurate and reliable.

Reliability. The degree to which scores are free from errors of measurement. A reliability of 0.80 is generally considered to be acceptable for an occupational/professional credentialing examination. Various statistics are used to measure reliability, with perhaps the most common being the KR-20 and coefficient alpha.

Scaled score. A score to which raw scores are converted by a numerical transformation (e.g., conversion of raw scores to percentiles). (The raw score is the unadjusted score on a test, often determined by counting the number of correct answers.)

Scoring rubric. The established criteria, including rules, principles, and illustrations used in scoring responses to individual items and clusters of items. The term usually refers to the scoring procedure for assessment tasks that do not provide enumerated responses from which examinees must make a choice.

Validity. The extent to which a test result measures what it purports to measure.

The above definitions are adapted from a variety of sources, including: *Standards for Educational and Psychological Testing* and *Certification: A NOCA Handbook*.