The Layman's Guide to Testing Terminology

As SMT interfaces with clients and credentialing organizations throughout the country, we have found that a common understanding of basic testing and measurement terms greatly enhances the communication process about various aspects of the examination program being discussed. For this issue of *DIMENSIONS,* SMT thought it would be useful to prepare a quick-reference glossary of common test development terms as a general reference.

Answer Key. The computer file containing the correct responses for each item on the examination.

Bad Pairs. Two or more items identified as similar in content or where the relationship between the items leads the candidate to the correct responses of one or more items. Identifying bad pairs alerts the test developer to select only one of the items for the examination.

Bankfit Report. A Bankfit report outlines the item bank by content areas (test specifications) indicating the number of items in each area, the subcontent categories, and status of the items. The bankfit report also displays the content area's percentage of items required on an examination.

Biserial Correlation. As a measure of correlation, the point-biserial coefficient is an index of discrimination between higher and lower scoring candidates and estimates the degree of association between two variables: a single test item and a total test score. Item writers will use the data yielded by the point-biserial correlation in conjunction with p-values to examine the quality of particular items.

Cognitive Level. The level of complexity whereby an individual attains, retains, and develops knowledge and intellect.

- *Knowledge* - The knowledge level requires a candidate to answer questions solely by memory and involves the recall of definitions, facts, rules, sequences, procedures, principles, and generalizations.

- *Application* - The application level involves the use of abstracts in concrete situations. The abstractions may be in the form of general ideas,

procedures, or methods. They may also be in the form of technical principles, ideas, and theories that must be remembered or applied.

- *Analysis* - The analysis level requires students to break down information into its constituent parts. Finding assumptions, distinguishing facts from opinion, discovering casual relationships, and finding fallacies in stories or arguments.

Camera -Ready Copy. The final draft of the examination that is used for printing examination booklets.

Cut Score. See Standard Setting.

Distractors - those question options that are incorrect responses. The candidate selecting a distractor to answer a problem or question presented in the stem will not receive credit.

Equating. In the construction of parallel test forms, the forms are assembled to be approximately equivalent in content, format, and difficulty. Since parallel test forms are not exact replicas and are only approximately equivalent, they can be made statistically equivalent by using equating procedures. Equating procedures are used to adjust for differences in test form difficulty, not test form content.

Equating is a process where scores on one form of an examination are converted to scores on another form of the same examination. This process allows comparable comparisons to be made across examinee groups regardless of the examination form administered. An unfair comparison would occur if the raw score of an examinee who by chance took a more difficult form of a test were compared to that of an examinee who by chance took an easier form of the same examination.

Estimated Mean Test Statistic Report. This report is available in the more advanced item banking systems and combines the statistical data associated with each item and calculates the data reflecting the overall expected test statistics (e.g., test mean, reliability).

Field Testing. Field testing is a method for gathering performance statistics on new items prior to use for scoring purposes. Field testing is conducted in a controlled setting where volunteer groups of practicing professionals and students currently enrolled in training programs, take an actual examination where performance statistics are gathered on trial items.

Form File. The computer file containing item bank identification numbers from a completed pullsheet. The form file is then used to assemble the items into an examination and is also used as an answer key.

© 1997, Schroeder Measurement Technologies, Inc., All Rights Reserved

Items - the individual questions that collectively make up the licensure or certification examination.

Item Analysis. An item analysis is the calculation of certain statistics for each item on an examination for the purpose of assessing performance. The statistics calculation typically include:

- Number and percent of candidates selecting each response
- Difficulty index (p-value) for each item and the proportion of examinees selecting the correct answer
- Discrimination index for each item (biserial correlation)
- Standard deviation
- Standard error of measurement (SEM)
- Mean score for all candidates
- KR-20 estimate of reliability
- Distribution of item test discrimination indices
- Distribution of item difficulty indices

- Item comments from field test participants are reviewed along with the item statistics. Poorly performing items are either salvaged through a revision process or discarded. Items that perform well are placed in an active status in the item banks.

Item Bank. The test support system used to store, retrieve and maintain descriptive information about test items, such as where the items fit in the test specifications, performance statistics, and developing and printing examinations.

Item Bank Identification Number. A unique identifier applied to each item in the item bank that allows a test developer to track various data associated with an item (e.g., statistical data, historical data, specification levels, etc.).

Item Review and Revision. The purpose of an item review is to decide whether an item is relevant to the practice of the profession, content valid, fits the appropriate specification level, is unbiased, and correctly keyed and referenced. Items are reviewed and revised to erify that they meet appropriate item writing criteria.

Item Writing. Item writing is usually performed in item writing workshops with subject matter experts participating in the process. A trained workshop coordinator instructs the subject matter experts in item writing techniques and provides definitions of item writing terminology.

© 1997, Schroeder Measurement Technologies, Inc., All Rights Reserved

Job Analysis. A job analysis refers to the study of the elements of knowledge, skill and ability necessary for an individual to practice. Job analysis also refers to the determination of those tasks which job incumbents typically perform that are important to competent performance.

Keyed Response - the correct answer to the problem or question presented in the stem. The candidate receives credit when selecting the keyed response.

Mean. The mean is the arithmetic average of a set of scores found by summing the scores and dividing that sum by the number of scores.

Multiple-Choice Item - an item containing a stem and several options from which a candidate selects a response (see example below).

What is the capital of Florida? ***(Stem)***

(A)	Tallahassee	***(Keyed Response)***
(B)	Jacksonville	***(Distractor)***
(C)	St. Petersburg	***(Distractor)***
(D)	St. Augustine	***(Distractor)***

Non-Doc. A term that is used to describe the initial printing of an assembled examination that is reviewed by subject matter experts for content and editors for grammar and punctuation. The non-doc contains all the data associated with each examination item. The non-doc is considered a first-draft form of the examination.

Options - the choices given as possible answers to the stem [e.g., (A), (B), (C), (D)] of a test question. Only one option is the keyed (correct) response.

P-Value. The p-value is a shorthand method for expressing the proportion of examinees who responded correctly to a test question. The p-value represents a difficulty index relative to the particular group of examinees to whom the item was administered. The p-value provides the test developer with valuable item performance data.

Pretesting. Pretesting is another method for gathering performance statistics on new items prior to use for scoring purposes. Tryout items are interspersed throughout an examination without being identified as such to the testing candidates. The embedding of items allows for a higher rate of response. This is the best available method for prediction of an item's statistical characteristics because pretest items are answered only by entry-level candidates.

© 1997, Schroeder Measurement Technologies, Inc., All Rights Reserved

Pretested or field tested items undergo a full item analysis where the p-value and biserial correlation can be computed for each item. In addition, a distractor analysis can be performed to ascertain the attractiveness of incorrect responses to the uninformed candidate. Items with poorly performing distractors, low p-values, and negative biserials are identified through this process. In the pretesting process, items are also screened for accuracy, brevity, clarity, and balance coverage of the specified categories.

Pullsheet. The worksheet created by a test developer listing the item bank identification numbers of the items that will be assembled into an examination. The pullsheet is used to create a form file.

Reliability. A general term denoting consistency of measurements derived from repeated observations on the same subject under the same circumstances. High reliability increases the dependability of an examination, since the scores are less subject to chance variability. The most commonly used statistic for measuring reliability is the KR-20 (Kuder-Richardson Formula) Index.

Specifications Fit Report. A Specifications Fit report is run against the assembled examination. The spec fit report compares the examination items against the test specification percentages, and indicates whether or not all of the test specification percentages have been met.

Standard Deviation. A measure of the variability or dispersion in a set of scores that provides an indication of the average amount by which the scores deviate from the mean of the distribution.

Standard Error of Measurement. The standard error of measurement (the standard deviation of error) provides an absolute rather than a relative measure of the extent to which raw and true scores are equivalent.

Standard Setting. Once test specifications have been developed from approved task statements, items have been written, and an examination form has been developed, a performance standard (cut score) must be established. The sound development of an examination form is the determination of a standard, or minimum passing score, used to compare and interpret the test scores. The standard, or minimum passing score is most typically determined by a subject matter expert panel procedure (Angoff). Each subject matter expert estimates the percentage of minimally competent candidates who will answer each question on the examination correctly. When the proportions established by the panel are summed across all questions, the result is the recommended minimum passing score or standard. The cut score is generally expressed as the number of items an examinee must answer correctly in order to pass.

© 1997, Schroeder Measurement Technologies, Inc., All Rights Reserved

Stem - the part of the item that presents a problem or question that requires a response from the candidate (see Item).

Test Production. The actual examination form production (camera copy) is typically includes the following steps.

- Obtain copy of test specifications
- Run a Bankfit report
- Create a pullsheet
- Create a form file
- Produce a non-doc copy
- SME review of proposed form
- Run an electronic grammar check
- Run a Spec Fit report
- Run an Estimated Test Stat Mean report
- Check for bad pairs
- Review psychometric characteristics of the item: (item performance)
- Review item comments
- Review of grammar, punctuation, and readability by test editor
- Produce camera copy of examination
- Perform camera-ready proofing
- Produce an answer key

Test Specifications. Test specifications, sometimes called test blueprints, for a licensure/certification examination program, are drafted based on the information collected from the job analysis. The purpose of test specifications is to guide test developers in constructing examinations which are consistent with the job analysis as well as ensuring that each form of the examination tests the same basic concepts. The job specifications outline the content of the examination to be developed, the relative emphasis to be placed on each content area, the appropriate cognitive level at which to test the examination content, and the number of questions on the examination.

Test Specification Classifications. The content/subcontent areas (examination outline or blueprint) by which a test developer constructs an examination. Sometimes referred to as content areas or content domains.

© 1997, Schroeder Measurement Technologies, Inc., All Rights Reserved