Accreditation - Accreditation is a voluntary process that an entity, such as a certification program, may elect to undergo. In accreditation a non-governmental agency grants recognition to the organization, after verifying that it meets published standards. Accreditation functions are performed by an independent third party, whose responsibility is to assess the applicant organization's compliance with the published accreditation standards. Accreditation is awarded for a finite period of time and requires renewal.

Audit - The process of conducting an evaluation of an entity's compliance with published standards. This is also referred to as a program audit.

Auditor - This refers to the individual(s) with the requisite skills and knowledge to conduct an evaluation of an entity's compliance with published standards. Auditors are trained to evaluate with every component of the entity under review for compliance with the standards.

Board of directors - A group of individuals appointed or elected with the authority over a credential, and assigned the duty of and responsibility for assuring its administration. This is also referred to as a Regulatory Board or a Governing Board.

Certificate test - A certificate test, like certification and licensure tests, is a criterion-referenced test (CRT). However, it is typically a low-stakes exam included as a component of a training program.

Certification program - This may refer to a few or to all components associated with awarding a certification. That is, it may refer to the certification examination or to the full set of activities related to awarding and maintaining the credential. These activities include: eligibility, examination, recertification, disciplinary, governance and policies.

Certification test - A certification test is typically a voluntary exam program designed as a criterion-referenced testing (CRT) that measures professional competence, and is sponsored by a non-governmental agency. This type of test may be targeted to measure entry level professional skills, specialty skills within the profession, or advanced skills in the profession.

Classical test theory - This is the traditional approach to measurement, which concentrates on the development of quality test forms. Classical test theory is used in item analysis,

through statistics such as the p-value and point-biserial correlation. It is also used to assemble test forms according to statistical criteria, and to evaluate the quality of those forms through reliability and validity analyses.

Classification - Classification is used in testing as the process of categorizing, or classifying, examinees into two or more discrete groups, such as pass/fail or master/non-master. The classification of examinees, into categories of competence or non-competence, is the typical goal of a criterion-referenced testing program.

Classification error - This refers to the misclassification of examinees into the pass and fail categories when a passing score is applied. Classification errors can occur in both directions. That is, a truly competent examinee might fail the test, while an incompetent examinee might pass the test. A primary goal of well-designed exam programs is to minimize classification errors.

Code of ethics - This refers to the canons or professional standards that certificate holders must agree to uphold and abide by. It is frequently an agreed upon statement of principles and expected behavior and conduct of the certificate holders. Commonly referred to as Standards of Practice or Codes of Professional Conduct, the canons are subject to enforcement and certificate holders found in violation of the code of ethics may be subject to disciplinary procedures. Codes of ethics are made a requirement for application to or awarding of certification.

Cognitive level - The cognitive level of an item refers to the type of mental processing on the part of the examinee that the item is designed to target. While Bloom's Taxonomy specifies six cognitive levels (Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation), exam programs are more likely to target a subset of these, such as the first three, in their test blueprints.

Common-item nonequivalent groups design - This is a frequently used data collection design for equating purposes. In this design two or more test forms are assembled so that they include a subset of identical, or common, items. The use of common items across test forms provides information about differences in the abilities of nonequivalent examinee groups who may be administered the different test forms. This information is then used to statistically equate the forms, in order to make examinees' scores comparable.

Computer-based testing - This refers to the mode of test administration in which items are presented to an examinee on a computer screen. Examinees typically indicate their responses by clicking with a mouse. This is contrasted with the more traditional method, paper-and-pencil testing.

Conjectural methods - This class of item-based approaches to standard setting includes the most commonly used method, the modified Angoff method.

Concurrent validity - This refers to a statistical method for estimating the validity of a test that provides evidence about the extent to which the test classifies examinees correctly. Concurrent validity estimates the relationship between an examinee's known status as either a master or a non-master and that examinee's classification as a master or a non-master as a result of the test.

Content validity - The content validity of a test is estimated through a logical process in which subject matter experts (SMEs) review the test items in terms of their match to the content areas specified in the test blueprint. For certification and licensure tests, this is typically the most important type of validity.

Contrasting groups method - This is an examinee-based approach to standard setting in which a panel of judges is used to identify a subset of examinees who would definitively be considered non-masters and another set of examinees who would definitively be considered masters. The exam is administered to these two groups of examinees, and their two resulting test score frequency distributions are then plotted. The passing score is typically set at or near the intersection between these two distributions.

Credentialing program - This refers to both voluntary and mandatory programs, including those at the registration, certification, and licensure levels of regulation.

Criterion-referenced tests - Criterion-referenced tests (CRTs) are designed to classify examinees into two or more distinct groups by comparing their performance to an established standard of competence. Certification and licensure exam programs are typically CRTs, as opposed to norm-referenced tests (NRTs).

Decision consistency - This is a measure of the reliability of the classification decision. Decision consistency estimates the extent to which, if an examinee were administered a test on two separate occasions, the same classification decision (whether pass or fail)

would be made. This measure of reliability is particularly appropriate for criterion-referenced tests.

Diagnostic score - Diagnostic scores provide detailed information about an examinee's performance on a test. In a criterion-referenced test, diagnostic scores may be particularly useful to examinees who failed the test and plan to test again.

Difficulty index - This is a measure of the proportion of examinees who responded to an item correctly. It is also referred to as the p-value.

Discrimination index - This is a measure of how well an item can distinguish between examinees who performed well on the overall test, and those who did not.

Disciplinary program - A disciplinary program is usually administered by the certification board. It addresses the investigation and disposition of complaints filed against certificate holders alleged to be in violation of a code of ethics or professional standards. A possible outcome of the disciplinary program is the imposition of sanctions against certificate holders found in violation of the organization's ethical principles.

Distractors - This refers to the incorrect options presented to the examinee in a multiple choice, or other objective item type.

Distractor analysis - This is a measure of how well each of the incorrect options in a multiple choice, or similar item type, is performing.

Eligibility - This refers to the published requirements that applicants must demonstrate and document in order to qualify for the exam. Eligibility often includes a combination of experience and education; the requirements may include references and work/job verification.

Equating - Equating is a statistical method for making the scores on one test form comparable to scores on another, parallel test form.

Equipercentile equating - This is a method of equating in which the percentile rank scores for one test form are placed on the scale of another, parallel form.

Face validity - Face validity refers to the extent to which a test appears to measure what it is supposed to measure. While it may be of value for a test to have face validity with the stakeholders of the exam program, it is not sufficient for establishing that the test actually has validity. Other, more substantive types of validity should be investigated and documented.

Fairness - This refers to whether an exam is free from any bias, or tendency to disadvantage an examinee based on characteristics such as race, religion, gender, or age. It is an important criteria for evaluating the quality of a test.

Feasibility analysis - This refers to a preliminary study that can be conducted prior to converting an exam program to computer-based testing. The feasibility analysis can help determine whether CBT is the right choice for a particular exam program, and what critical program issues may need to be addressed to ensure a successful conversion.

Field test - This refers to a phase in the test development process when items are administered to examinees, for the purpose of collecting data to evaluate the performance of the items themselves.

Frequency distribution - This refers to a tabular or graphical representation of a set of data in an orderly arrangement. In testing, one common type of frequency distribution is a test score distribution, in which test scores are arranged from lowest to highest (or highest to lowest), and the frequency with which each score occurred is displayed.

Governing Board - A group of individuals appointed or elected with the authority over a credential, and assigned the duty of and responsibility for assuring its administration. This is also referred to as a Board of Directors or a Regulatory Board.

Governance structure - This refers to the structure established to provide oversight to and administration of various aspects of a certification entity. A typical governance structure for a professional certification program may include a governing board, as well as several technical committees assigned to perform specific functions of the certification program (for example, a disciplinary committee).

Informed judgment method - This is a test-based approach to standard setting in which relevant stakeholders review the overall test in order to suggest a percent-correct score that each believes ought to be earned by a minimally competent examinee. This method

is in contrast to the conjectural methods, such as the modified Angoff, in which a panel of judges suggest percent-correct scores for each item.

Internal consistency - The internal consistency measure of reliability estimates how well the set of items on a test form correlate with one another. This method of reliability is likely to produce higher values for norm-referenced tests than for criterion-referenced tests, given that CRTs are often designed to measure a broad range of content topics.

Interrater reliability - This method of reliability is used when a test includes performance tasks, or other items that need to be scored by human raters. Interrater reliability estimates the consistency, or dependability, of the scores produced by the human raters.

Item - A test question is referred to more formally as a test item. This term is used because, while the examinee is always being asked to respond to something, the item is frequently not structured as a direct question.

Item analysis - This refers to a set of statistical procedures used to evaluate the quality of test items. The item analysis is conducted after examinees have responded to the set of items. The measures most commonly included in the item analysis are the item difficulty index, item discrimination index, and distractor analysis.

Item bank - The item bank, or item pool, is typically comprised of the entire set of items that have been written for the exam program. The bank may include items that have not yet been pretested, retired items that are no longer being used, along with items that are available for current, operational use. In most exam programs, additional information about the items, such as the content and cognitive classifications, is stored along with the item text.

Item banking software application - This refers to the database-type software program that may be used to store the exam program's items, along with additional information about the items.

Item difficulty index - This is a measure of the proportion of examinees who responded to the item correctly. It is also referred to as the p-value.

Item discrimination index - This is a measure of how well an item can distinguish between examinees who performed well on the overall test, and those who did not.

Item types - This refers to the variety of test item structures or formats that can be used to measure examinees' knowledge, skills, and abilities.

Item review - This refers to a test development phase in which items are examined by subject matter experts, professional editors, measurement experts, and others to ensure that they satisfy a variety of quality criteria.

Item specifications - Item specifications are very detailed requirements sometimes provided to the subject matter experts (SMEs) who are tasked with writing items for an exam program.

Job analysis - This refers to a study that may be conducted to identify the knowledge, skills, and abilities in a particular field necessary for professional competence. A job analysis is frequently conducted as part of a certification or licensure exam program's determination of the content and competencies that ought to be included in the test. Similar activities include task analyses, practice analyses, and role delineation studies.

Key - The key is the correct response to a multiple choice, or other objective item type.

Legal defensibility - This refers to the extent to which there is evidence regarding a test's quality that would stand up in a court challenge. For an exam program to be legally defensible, sound measurement procedures need to be both followed and documented, in the design, development, and maintenance of the test. Evidence of the test's content validity is particularly important.

Licensure test - A licensure test is typically a mandatory exam program sponsored by a governmental agency for the purpose of protecting public health, safety, and welfare. Licensure exams are criterion-referenced tests (CRTs) which individuals are legally required to pass before they may be employed in the occupation.

Linear equating - This is a method for equating in which the mean and standard deviation of one test form are placed on the scale of another, parallel form.

Minimally competent - This refers to a basic level of examinee proficiency that is determined to reflect acceptable performance for passing a given certification or licensure

exam. The concept of minimal competency is used during standard setting to determine the passing score on the test.

Misclassification - This refers to the problem of classification errors that may occur in criterion-referenced testing, when the passing score is applied. In one type of misclassification, an examinee who is not minimally competent may still pass the test; in a second type of misclassification, an examinee who actually is competent, nevertheless fails the test. A primary goal of well-designed exam programs is to reduce misclassification as much as possible.

Modified Angoff - In this frequently used approach to standard setting a panel of judges is asked to review a test, one item at a time. Each judge gives an estimate of the proportion of a hypothetical group of minimally competent examinees that would be expected to respond to each item correctly. These proportions are then averaged across items and across judges to arrive at a recommended passing score for the test.

Multiple choice item - This is the most commonly used item type in standardized testing. Each multiple choice item typically consists of a prompt or stem, which may be in the form of a question or statement, followed by a set of four or five response options. The single correct response option is called the key, while the remaining, incorrect response options are called distractors.

Norm-referenced test - A norm-referenced test (NRT) is designed to obtain individual scores for all examinees in such a way that examinees' performances can be compared to one another. While many standardized educational tests are NRTs, certification and licensure tests are designed to be criterion-referenced tests (CRTs).

Operational testing - This refers to the administration of tests under actual, scored conditions, where the results of the test are likely to have real implications for examinees and/or other stakeholders. The term operational testing is frequently used in contrast to pretesting, pilot testing, or field testing conditions.

Organizational structure - The structure established to administer a certification entity. It may include a governing board, committees, and a paid administration staff, with specific responsibilities for performing the operations and functions of the certification entity.

p-value - This is a measure of the proportion of examinees who responded to the item correctly. It is also referred to as the item difficulty index.

Paper-and-pencil testing - This refers to the traditional mode of testing in which examinees refer to paper test booklets to read the items and provide their responses using pencils. It is usually contrasted with computer-based testing.

Parallel forms reliability - This method of reliability is used when an exam program has developed multiple, parallel forms of a test. Parallel forms reliability provides an estimate of the similarity that might be expected between an examinee's scores across the separate test forms.

Parallel test forms - This refers to two or more test forms that are developed for a given exam program, according to the same test blueprint and statistical criteria. The forms should be assembled in such a way that they are as similar to one another as possible.

Pass/fail classification - This type of score information indicates whether or not an examinee demonstrated sufficient knowledge of the content and competencies measured by the test. The pass/fail classification decision is the most critical type of score information for criterion-referenced test programs.

Passing score - This refers to the minimum score an examinee must earn in order to pass a test, or to be classified as a master. The passing score for an exam program is determined through a standard setting process. It is also known as the passing point, the cutoff score, or the cut-score.

Percentile rank score - This type of score provides a comparison of an individual examinee's performance to other examinees who took the test. Specifically, an examinee's percentile rank indicates the percentage of other examinees who earned scores below that of the given examinee.

Pilot test - This refers to a preliminary item field test, in which data from a small sample of examinees who respond to the items is collected for analysis and review.

Point-biserial correlation - This is a specific statistical technique often used in testing as a means of computing an item discrimination index.

Policies - The written actions or principles that guide the administration, operation, and decision-making of the certification body. Policies are frequently set by the governing board or other entity with authority over the credentialing program. Policies are implemented by all persons associated with the functions of the certification program to assure fairness and consistency in decisions and practices.

Predictive validity - The predictive validity of a test, like the concurrent validity, is estimated statistically. Predictive validity is specifically concerned with the extent to which a test can predict examinees' future performances as masters or non-masters. It is particularly important for tests used in such applications as selection or admissions.

Pretesting - This refers to an item evaluation process. Specifically, pretesting is used to collect examinee response data for evaluating the performance of new items intended to supplement the item bank in an ongoing exam program.

Procedures - The administrative steps followed to implement and administer policy.

Professional certification - This refers to a non-governmental process for ensuring professional competency. In professional certification, standards and requirements for a profession are established to ensure that individuals awarded certification have met the requisite knowledge, skills, and abilities to perform at the pre-determined level in the profession.

Program audit - The process of conducting an evaluation of an entity, or its individual components, to determine compliance with published standards. This is also referred to as an audit.

Psychometrician - A psychometrician is a professional who works in the field of psychometrics, or measurement. Specifically, psychometrics refers to the measurement of individuals' psychological attributes, including job-related knowledge, skills, and abilities.

Recertification program - A program that defines the requirements for maintaining a credential. Recertification may be referred to as renewal, as certification is not awarded for life. Certificate holders must demonstrate continued competence at the skill and knowledge level assessed by the original certification. Frequently, recertification requirements includes continuing education and current experience.

Regulatory Board - A group of individuals appointed or elected with the authority over a credential, and assigned the duty of and responsibility for assuring its administration. This is also referred to as a Board of Directors or a Governing Board.

Reliability - The reliability of a test refers to its ability to produce consistent scores or classifications. This is one of the most important criteria for evaluating the quality of a test. There are several specific types of reliability, including decision consistency, parallel forms reliability, test-retest reliability, internal consistency, and interrater reliability.

Response options - The response options are the answer choices presented to the examinee in certain item types, such as the multiple choice. In most cases, the examinee is instructed to select the single, best option from these choices.

Scale score - This refers to a type of converted score that may be used by an exam program. Scale scores may be established for an exam program so that all of the test forms can be converted and then reported on the same scale. Over time, the scale score values of the scale score acquire meaning as stakeholders become familiar with the level of performance reflected in specific numeric values. A scale score, like a percentile rank score, has a norm-referenced type of interpretation.

Score distribution - This refers to the tabular or graphical display of a set of examinees' scores, in which the number of times each score occurred in the data set is represented.

Score report - The score report for an individual examinee provides information about that examinee's performance on the test. Depending upon the goals of the exam program, the score report might only provide an overall classification decision. Alternatively, it might include either more detailed information about the examinee's performance on test subsections, and/or additional types of information. Specific types of scores that might be provided include pass/fail classifications, diagnostic information, percentile ranks, and scale scores.

Standard setting - This refers to the process used in an exam program to determine the passing score for a test. Methods for conducting standard setting include informed judgment, conjectural, and contrasting groups methods.

Standards of practice - A published set of behavioral and professional expectations of certificate holders. Standards of practice are frequently developed with input from industry. They are updated periodically for relevance and currency.

Standardized test - This refers to an exam program in which specific procedures have been developed for many aspects of the development and administration. These procedures are intended to be consistently followed, so that test scores may be as comparable as possible. Both criterion-referenced tests and norm-referenced tests can be standardized.

Stem - In a multiple choice item, the stem is the initial question or statement that poses a problem to which the examinee must provide a response. In addition to the text statement, in some cases the stem may include a figure, chart, graphic, or other exhibit.

Subject matter expert - The term subject matter expert, or SME, is used to refer to personnel who are used at different phases of the test development process because of their extensive knowledge of the content and competencies being assessed by the exam. SMEs are used in test development activities such as item writing, item review, and standard setting.

Table of specifications - The table of specifications, also known as the test blueprint, provides the structure for the test forms that will be developed. A typical table of specifications includes a listing of the content areas important to be included on the test, along with the cognitive levels that test items are intended to target.

Test administration - This refers to the process followed to give exams to examinees, typically under standardized procedures and operational conditions.

Test assembly - This refers to the process used to construct test forms. It includes the selection of items from an item bank according to a set of statistical and content criteria. It may also involve the formatting, layout, and ordering of the items on the test, as well as the inclusion of the test directions.

Test blueprint - The test blueprint, also known as the table of specifications, provides the structure for the test forms that will be developed. A typical test blueprint includes a listing of the content areas to be included on the test, along with the cognitive levels that test items are intended to target.

Test description - This is the aspect of an exam program's test specifications which is used as a framework for designing the test. Typically, this description may include such information as the purpose of the test, the intended test audience, and other background information.

Test development - This refers to the orderly process followed in the design, development, and maintenance of an exam program. The steps followed may begin with establishing the purpose of the test, conducting a job analysis, and creating a set of test specifications. For ongoing exam programs, this will also include steps such as routine item writing and review efforts, the assembly of new test forms, regular standard setting meetings, as well as test administration and scoring activities.

Test purpose - This refers to the intended goals and uses of the results of the test. The test purpose of an exam program has many, far-reaching implications for the design, development, and maintenance of the test.

Test-retest reliability - Test-retest reliability provides an estimate of the extent to which a test is able to provide consistent, stable test scores across time.

Test security - This term is used to refer both to a set of procedures designed to prevent problems such as cheating, as well as to the need for these procedures. For a given exam program, methods to address test security may span test development, examinee registration, test administration, and post-administration statistical analyses.

Validity - The validity of a test refers to the extent to which the test actually measures what it claims to measure. This is one of the most important criteria for evaluating the quality of a test. There are several specific types of validity, including content validity, concurrent validity, and predictive validity.

Workforce - In the conduct of a certification program, this refers to the individuals who have the required knowledge and skills measured by the certification. The workforce also constitutes the total number of persons or practitioners who perform in the profession or occupation; this may or may not include all persons who could be eligible for certification.