STANDARDS FOR QUALITY ASSURANCE

NATIONAL BOARD OF OSTEOPATHIC MEDICAL EXAMINERS

D O

• SINCE 1934 •

June 2015

Contents

INTRODUCTION

The NBOME mission is to protect the public by providing the means to assess competencies for osteopathic medicine and related health care professions. Central to this mission, we are committed to consistently assuring that any and all products and services offered are of the highest quality and integrity. Further, we continually seek to improve our products and services to serve the needs of the osteopathic profession and the public. To this end, the NBOME Standards for Quality Assurance were developed and adopted by the NBOME Board of Directors in 2006 to provide a uniform measure of quality in all that we do. The Standards were revised in 2015, following the 2014 revision of the *Standards for Educational and Psychological Testing*, published by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education and continue to serve as a benchmark for excellence for all NBOME products and services. While the *Standards for Educational and Psychological Testing* are comparable, they differ in areas that reflect the specific nature of NBOME products and services.

The NBOME Standards for Quality Assurance are used by NBOME staff on a daily basis and successful implementation is checked through a rigorous and comprehensive audit program to ensure that the COMLEX-USA licensure examination series is a fair and reliable assessment. To the extent possible, NBOME's Standards guide the quality and integrity of all NBOME products and services.

Although no set of guidelines can encompass all parameters of quality, the NBOME Standards for Quality Assurance serve to assist the staff and the Board in rendering a judgment as to whether or not the expectations of quality are met. In the event that these standards applied to COMLEX-USA as intended are not met, a methodical analysis is required to remediate any deficiencies. When performing audits, these guidelines offer assistance in assuring the continual quality of NBOME products and services. The Standards and Assurances Committee is permitted room for professional judgment in ultimately deciding the quality of what is done at the NBOME. That judgment resides ultimately with, the NBOME Board of Directors, which provides oversight of the audit program and determines the most appropriate use of these guidelines.

The NBOME Standards for Quality Assurance are periodically reviewed and modified to meet the changing demands for quality in current and future products and services offered by the NBOME.

1.0 Assessment Design and Development

The NBOME follows test design and developmental procedures to support the validity of the interpretation(s) of test scores for their intended uses. The standards in this section refer to procedures involved in planning and designing tests and the expected test interpretations for the intended uses of test scores.

1.1 Prior to examination development and implementation:

- Describe the intended purpose of the products or services including desired major characteristics, clear statement of intended interpretation of scores and evidence of validity.
- Describe the intended uses, the intended populations and the need that the product(s) or service(s) meets.

1.2 Set forth clearly the construct(s) that the new or revised existing product(s) or service(s) is intended to assess.

- Define clearly the content domain(s) covered and justification in terms of the importance of the content for the intended purpose(s).

1.3 For new product(s) or service(s) or revising existing ones, provide a rationale and evidence to support the claim that the knowledge or skills being assessed are required and consistent with the purpose of the examination program.

- Document the rationale for a test, recommended uses of the test, support for such uses and information that assists in intended score interpretation.
- Make a clear statement of the intended interpretations of test scores for specified uses.

1.4 Develop a plan for new product(s) or service(s) or to revise existing ones that provides evidence of fairness, reliability and validity.

1.5 Document the test design and development processes used for new product(s) or service(s) or to revise existing ones that provides evidence of fairness, reliability and validity.

1.6 Summarize test development procedures including:

- Descriptions and results of the statistical analyses that were used in the development of the test.
- Evidence of the reliability, precision of scores and the validity of their recommended interpretations.
- Methods for establishing performance cut scores.

1.7 Obtain substantive advice and reviews from diverse internal and external sources, including clients and users.

1.8 When a validation rests in part on the opinion or decision of expert judges, observers or raters, fully describe procedures for selecting such experts and for eliciting judgments or ratings, including:

- The qualifications and experience of the judges are presented.
- The description of procedures including any training and instructions provided.
- Indications of whether participants reached their decisions independently and a report on the level of agreement reached.
- If participants interacted with one another or exchanged information, the procedures through which they may have influenced one another are set forth.

1.9 Evaluate new or revised product(s) or service(s) at reasonable intervals and in response to significant changes.

- Amend or revise test specifications when new research data, significant changes in the domain represented or newly recommended conditions of test use may reduce the validity of test score interpretations.
- Design subsequent validation efforts to determine how well the intended interpretations have been achieved for all relevant subgroups.
- Verify test uses periodically so that their interpretations of test data continue to be appropriate, given any significant changes in the population of test takers, the mode(s) of test administration or the purposes in testing.
- Although a test that remains useful need not be withdrawn or revised simply because of the passage of time, test developers and test publishers are responsible for monitoring changing conditions and amending, revising or withdrawing tests as indicated.

1.10 Offer NBOME product(s) or service(s) that fit the NBOME mission.

2.0 Customer Service

The NBOME strives to meet customer needs and maintain high quality in designing, developing and delivering product(s) and service(s). The NBOME identifies its customers for its assessment(s) and program(s). The term "customer" includes clients, users of scores or other assessment results, purchasers of products and services developed by the NBOME and test takers.

2.1 Identify key clients and other customers and obtain their input into the product and service design, development and operation.

2.2 Develop customer service standards and monitor the extent to which the standards are met and serve the needs of customers.

2.3 Provide convenient methods for customers to obtain information about products and services, ask questions, make comments, or register problems or concerns.

- If an answer is required to a customer inquiry, respond promptly and courteously.
- Evaluate the effectiveness of these methods.

2.4 Measure test taker and test user satisfaction periodically and use the information for continuous quality improvement.

3.0 Fairness

The NBOME ensures that product(s) and service(s) are designed, developed and administered with attention to treating people equally and fairly. Fairness is defined as treating people with impartiality regardless of personal characteristics such as age, gender, race, ethnicity or disability that are not relevant to their interaction with the NBOME. Fairness requires that construct-irrelevant personal characteristics of test takers have no appreciable effect on test results or their interpretations. The NBOME is responsible for the fairness of the product(s) or service(s) it develops and for providing evidence of fairness. NBOME product(s) and service(s) consider the diversity of the populations served. While it is not feasible to investigate fairness for all of the possible groups that may be defined, the NBOME uses experience or research to identify population groups to be included in evaluations for fairness. Groups included in the evaluations of fairness are referred to as "studied" groups in this section. Users of a product or service are responsible for assessing the relevance of the evidence provided by the NBOME for their particular needs and situations.

3.1 Develop tests that measure the intended construct and minimize the potential for tests to be affected by construct-irrelevant characteristics.

3.2 Address fairness in the design, development, administration and use of the product(s) or service(s), and document what was done.

- Provide a plan for addressing fairness for a product or service under development or facing a major revision.

3.3 Obtain and document judgmental and when possible, empirical evaluations of fairness for studied groups. As appropriate:

- Ensure that relevant subgroups are represented in the intended population.
- Ensure that symbols, language, and content that are generally regarded as sexist, racist or offensive are eliminated, except when necessary to meet the purpose of the assessment, product or service.

3.4 Provide equal access to product(s) and service(s). For assessments:

- Provide impartial registration, administration and reporting of assessment results.

3.5 When a construct is measured in different ways that are reasonably equally valid, reliable, practical and affordable, consider available evidence of group differences in assessment results in determining how to measure the construct.

3.6 Provide appropriate and reasonable accommodations for people with disabilities, in accordance with applicable laws, NBOME policies and clients' policies.

3.7 For research studies, obtain informed consent for participation of human subjects where necessary and avoid negative consequences of participation for members of all groups.

4.0 Uses and Protection of Information

The NBOME safeguards and protects critical and confidential information and provides information to the public that allows the evaluation of its products and services and promotes their proper use. The NBOME provides information that promotes public understanding of measurement and assessment as it applies to its product(s) and service(s).

4.1 Provide users of examination(s) or service(s) with the information they need to determine whether or not the examination(s) or service(s) is appropriate for them. Information relating to the test(s) includes:

- The purposes of testing.
- How tests are administered.
- Factors considered in scoring examinee responses.
- How scores are used.
- How long records are retained.
- To whom and under what conditions records are released.

4.2 Seek to use clear and jargon-free language in communications designed for a general audience.

- Medical abbreviations and acronyms are used minimally.
- When used, medical acronyms are clearly defined.

4.3 Set forth clearly how test scores are interpreted and used and clearly:

- Delimit the population(s) for which a test is intended.
- Describe the construct(s) that the test is intended to assess.

4.4 Retain the information necessary to verify assessment results for a period consistent with the intended use of the assessment.

- Advise individuals about the time that information will be retained, as is practical.

4.5 Maintain the confidentiality of personal or institutional information and inform individuals that information about them will be kept confidential unless permission is obtained or law requires disclosure.

- Keep information about individuals in accordance with applicable laws and established NBOME policy.

4.6 Maintain the security and integrity of test materials and results at all times.

- Eliminate opportunities for test takers to attain scores, test materials, specific test contents or other assessment results by fraudulent means to the extent possible.

4.7 Ensure procedures to safeguard and recover the data and information necessary for the operation of testing programs in the event of a disaster.

4.8 Allow non NBOME researchers reasonable access to NBOME-controlled, nonproprietary data and information, provided the privacy of individuals and organizations and contractual obligations are not violated.

- Discourage the release of data and information that is misleading or leads to bias or misinterpretation.
- The release of data or information may not in any way place at risk the integrity or security of any NBOME examination or service.

4.9 Protect NBOME's and clients' intellectual property rights with respect to such proprietary products as items, software, marketing studies, procedural manuals, new product development plans and the like.

5.0 Validity

Validity is one of the most important attributes of assessment quality. The NBOME collects and documents appropriate evidence for its assessments to support the intended inferences and actions based on the reported assessment results. Logical and/or empirical evidence is provided to show that each assessment is capable of meeting its intended purposes(s). Validity is a unified concept but many different types of evidence may contribute to the demonstration of an assessment's validity. Validity is not based solely on any single study or type of evidence. Though all types of evidence may contribute to the validation of an assessment, the type of evidence on which most reliance is placed varies with the purpose of the assessment. The level of evidence required may vary with the potential consequences of the decisions made on the basis of the assessment's results. Responsibility for validity is shared by the NBOME, its clients and the people who use the scores or other assessment results. NBOME provides evidence of validity at least to the extent required by the following standards. Users are responsible for evaluating the validity of scores or other assessment results used for the purposes other than those specifically stated by the NBOME.

5.1 Clearly describe the following and make appropriate information available to the public:

- The purpose(s) of each assessment.
- The construct (knowledge, skills or other characteristics) measured.

- The intended test-taking population(s).
- The intended interpretation(s) of scores or other assessment results.

5.2 Provide a rationale for the types and amounts of evidence collected to support the validity of the inferences to be made on the basis of the assessment.
  - Provide a validity plan indicating the types of evidence collected for new assessments and assessment formats.
  - Provide appropriate evidence of validity in support of each intended interpretation.

5.3 Obtain and document the logical and/or empirical evidence that the assessment meets its intended purpose(s) and supports the intended interpretation(s) of assessment results for the intended population(s).

5.4 If the use of an assessment results in unintended consequences for a group that is studied, review the validity evidence to determine whether or not consequences arise from bias.
  - If consequences arise from bias, revise the assessment to reduce, to the extent possible, the bias.

5.5 Provide evidence of validity based on test content including a thorough and explicit definition of the current content domain of interest.

5.6 When test content is a primary source of validity evidence supporting the interpretation of a test used in credentialing decisions, demonstrate a close link between test content and the associated professional and/or occupational requirements.

6.0 Assessment Development

NBOME assessments are constructed using planned, documented processes that include advice from diverse people, formative and summative evaluations and attention to fairness, reliability and validity. Test developers work from detailed specifications, obtain reviews of their work, use empirical information when it can be obtained and evaluate their finished products.

6.1 Document the desired attributes of the assessment in detailed specifications and the process used to develop the specifications.

6.2 Write test items that meet generally accepted guidelines and follow specific guidelines established by the NBOME.
  - Items meet the specifications or blueprint established by the NBOME.

6.3 Assure proper orientation of the item writers and case developers is conducted and that the expectations of the item writers and case developers are clearly understood.
  - Provide item writers with appropriate feedback on their contributions.
  - Document the item development process.

6.4 Obtain internal and/or external reviews of the assessment and related materials as deemed necessary.

- Document the qualifications, relevant experiences and demographic characteristics of the reviewers.
- Document the purpose, process by which the review(s) is conducted and the results of the review(s).

6.5 Pre-test items when feasible; if pre-testing items is not feasible, use other acceptable psychometric processes, to include:

- Review of results of administering similar items to similar populations.
- Conduct a preliminary item analysis before scores or other assessment results are reported.

6.6 Establish and document procedures to maintain the technical quality and utility of product(s) or service(s).

- Provide adequate training to those responsible for test development.
- Monitor and document the quality of test development, including documentation and correction for any systematic sources of errors.

6.7 Perform quality assurance assessments after examination administrations.

- Document the model used to evaluate the psychometric properties of items (classical test theory, item response theory, e.g.).
- Describe the sample used for estimating item properties.
- Ensure adequate sample size and diversity.
- Document the process used to screen items (item difficulty, item discrimination, differential item functioning (DIF) for major examinee groups.
- If model-based methods are used to estimate item parameters, document the item response model used, estimation procedures and evidence of the model-fit.

6.8 Periodically review active items, assessments and ancillary materials to ensure that they continue to be appropriate and current, and in compliance with applicable guidelines.

- Investigate sources of irrelevant variance when evidence indicates that irrelevant variance could affect scores from the test.
- Remove or reduce the sources of irrelevant variance, where possible.
- When tests are revised, inform test users of changes to the specifications, adjustments to the score scale and the degree of comparability of scores from original and revised tests.

6.9 Protect the security of confidential assessment materials throughout the development process.

7.0 Reliability

The NBOME ensures that scores or other reported assessment results are sufficiently reliable to meet their intended purposes and that appropriate procedures are used for determining and reporting reliability. Reliability refers to the extent to which scores or other reported results obtained on a specific form of an assessment, administered at a particular time and potentially scored by particular rater(s) can be generalized to scores obtained on other forms of the assessment, administered at other times and potentially scored by other rater(s). Reliability can also be viewed as an indicator of the extent to which assessment results are free from the effects of random variation caused by such factors as the form of an assessment, the time of administration or the scorers.

7.1 Ensure that any reported scores or other assessment results are sufficiently reliable to support their intended interpretation(s).

7.2 Estimate reliability using methods that are appropriate for the characteristics of the assessment and the intended use(s) of the results.

- Use methods that take into account important sources of possible variation in assessment results.
- Clearly state the range of replications over which the reliability/precision is evaluated and the rationale given the testing situation.

7.3 Provide estimates of relevant indices of reliability when reporting total scores, sub scores or other combinations of scores that are interpreted.

7.4 When significant variations are permitted in tests or test administration procedures, provide separate reliability analyses for scores produced under each major variation if adequate sample sizes are available.

7.5 Evaluate the reliability and standard error of measurement of reported assessment results for studied population groups, if the need for such studies is indicated and if it is feasible to obtain adequate data.

7.6 In addition to other sources of reliability evidence, provide estimates of the consistency of assessment-based credentialing decisions.

- When a test or combination of measures is used to make classification decisions, provide estimates of the percentage of test takers who are classified in the same way on two replications of the procedure.

7.7 Provide test users with the appropriate evidence of reliability/precision for the interpretation of each recommended intended score use.

8.0 Scoring, Cut Scores, Scaling and Equating

NBOME assessments use score reporting scales that are meaningful. The NBOME participates in cut score studies using rational, clearly described procedures. Assessments that are meant to be linked to other assessments have comparability that is commensurate with the use(s) made of the scores.

8.1 Establish scoring protocols and use reporting scales for scores or other assessment results that are appropriate for the intended purpose of the assessment and that discourage misinterpretations.

- Clearly describe procedures for constructing scales used for reporting scores and the rationale for these procedures.
- Test developers and test users document evidence of fairness, reliability and validity of test scores for their proposed uses.

8.2 Provide users with clear expectations of the characteristics, meaning and intended interpretations of scale scores as well as their limitations.

8.3 If results on different assessments or on different forms of an assessment are treated as though they are equivalent or comparable, provide detailed technical information on the method by which equating functions are established and on the accuracy of the equating functions.

8.4 Describe the equating or other linking studies that were done in sufficient detail to allow participants in the judgment process to evaluate and replicate the studies.

- Describe the limitations of linking studies done when the linked assessments are not alternate forms of the same assessment.

8.5 Maintain a common scale over time and conduct periodic checks of the stability of the scale on which scores are reported.

8.6 Clearly document the rationale and procedures for establishing cut scores when proposed scored interpretations involve one or more cut scores.

8.7 For a cut score study, use an appropriate data-gathering plan, choose appropriate sample(s) of raters from relevant populations and train the raters in the method(s) they will use.

- Document the study in sufficient detail to allow for evaluation and replication.

8.8 When cut scores are based on direct judgments about the adequacy of item or test performances, design the judgmental process so that the participants providing the judgments bring their knowledge and experience to bear in a reasonable way.

8.9 When standard setting panels are utilized, document the credentials of the panelists and the conduct of the panels.

8.10 Periodically assess the stability and validity of the cut scores.

- Use recognized and appropriate methods for determining the standard or cut score(s) used in examinations.
- Conduct standard setting at appropriate intervals to assure the currency of the cut score(s).

8.11 Ensure that the required level of performance is not adjusted to control the number or proportions of persons passing test(s) used to make credentialing decisions such that the level of performance required for passing a credentialing test depends on the knowledge and skills necessary for credential-worthy performance in the occupation or profession.

8.12 Establish, document and follow scoring protocols.

- Use rubrics, procedures and criteria for scoring when test scoring involves human judgment.
- Document the accuracy of scoring algorithms of examination responses.

8.13 Establish and document quality control processes and criteria used in scoring tests.

- Provide adequate training to those responsible for test scoring.
- Monitor and document the quality of scoring, including documentation and correction for any systematic sources of scoring errors.

8.14 Make rules and procedures that are used to combine scores from multiple assessments in determining the overall outcome of a credentialing test available to test takers, preferably before the test is administered.

9.0 Assessment Administration

The NBOME administers assessments in an appropriate manner to provide an accurate, comparable and fair measurement for each test taker. Administration procedures including the level of security required vary with the type and purpose of the assessment. For assessments developed for its client(s), the NBOME collaborates with its client(s) to ensure proper assessment administration.

9.1 Provide those who administer assessments with timely, clear and appropriate information about administration procedures.

9.2 Provide test takers with information needed to facilitate their registration, reservation process and the administration of their examination at the test center.

9.3 Seek to ensure a reasonably comfortable and standardized testing environment in locations reasonably accessible to the intended population(s).

9.4 Monitor administrations as appropriate to ensure relevant test administration procedures are followed.

9.5 Provide appropriate accommodations and access to training for those with impairments or disabilities as defined by the Americans with Disabilities Act (ADA) and the ADA Amendments Act.

9.6 Provide a definition and methods for addressing irregular conduct prior to, during or immediately following the administration of an examination.

9.7 Provide guidelines for the management of special circumstances directly preceding, during and following an examination administration.

- Document and report to the test user changes or disruptions to standardized test administration procedures or scoring.

10.0 Reporting Assessment Results

The NBOME provides correct, understandable scores or other assessment results and appropriate interpretive information to the recipients of score or assessment reports. These guidelines ensure that scores and other assessment results are accurate. The NBOME reports the results of its assessments and communicates in a meaningful manner to the intended recipients.

10.1 When test score information is released, provide the interpretations appropriate to the audience. Describe in simple language:

- What the test covers.
- What scores represent.
- The precision/reliability of the scores.
- How scores are intended to be used.

10.2 Provide timely reports of test results to test takers and others entitled to receive this information**.**

- Provide timely reports of test results except under circumstances that clearly require that test results are withheld.

10.3 Establish and document quality control processes and criteria to:

- Monitor and document the quality of scoring reports.
- Document and correct any systematic source of score report errors.

10.4 When a material error is found in test scores or other important information, disclose this information and issue a corrected score report as soon as practicable.

- Label the corrected report as such and document what was done to correct the report(s).
- Make clear the reason for the corrected score report to the recipients of the report.

10.5 Provide information that minimizes the possibility of misinterpretation of individual assessment results or results for groups.

- Take steps to minimize or avoid foreseeable misinterpretations and inappropriate uses of test scores.

10.6 Provide recipients with norm or criterion based information if applicable for evaluating the performance represented by test takers' scores or other assessment results.

11.0 Assessment Use

The NBOME provides information that describes and encourages the proper use of its assessments. The intended users of assessment results are advised to avoid common misuses of the assessment(s). The NBOME promotes the proper use of assessments to help score recipients use assessments fairly and appropriately, in accordance with supporting evidence.

11.1 Provide intended users with the information they need to evaluate the appropriateness of the assessments and consultative services about appropriate uses of the assessment(s).

11.2 Provide the recommended use(s) and the intended interpretations of assessments.

11.3 Alert test users to avoid common, reasonably anticipated misuses or misinterpretations of the assessment(s).

- If there is a sound reason to believe that specific misinterpretations of a score scale are likely, explicitly caution test users.
- Advise test users to take steps to minimize or avoid foreseeable misinterpretations and inappropriate uses of test scores.

11.4 Investigate and respond to allegations of assessment misuse.

- If assessment misuse is found, inform the client and the user and inform the user of the appropriate use of the assessment.
- If misuse continues, consult with the client concerning appropriate corrective actions.

12.0 Test Takers' Rights and Responsibilities

The NBOME endeavors to make test takers aware of their rights and responsibilities and protects their rights during all phases of the assessment process. The rights of test takers are included in other sections in these Standards, such as Fairness, Assessment Administration and Reporting Assessment Results. Where the rights of test takers are included elsewhere, they are not repeated in this section.

12.1 Test takers have the right to adequate information to:

- Help them properly prepare for a test so that the test results accurately reflect their standing on the construct being assessed.
- Lead to fair and accurate score interpretations.

12.2 Provide all test takers with respectful and impartial treatment, appropriate access to assessment services and information about the assessment and the administration process.

- Make available and free of charge to all test takers any information about test content and purposes that is available to any test taker.

12.3 Provide to candidates an up-to-date, accurate Bulletin of Information and Orientation Guide, where appropriate, and website address to provide:

- Access to the most current information regarding the test(s).
- Test procedures.
- Special circumstances that arise from time-to-time.

12.4 Provide information about the characteristics of each test format when a test taker is offered a choice of test format.

12.5 Obtain informed consent from test takers as necessary.

12.6 Inform test takers how to register complaints about:

- Items believed to be flawed.
- Assessment content believed to be inappropriate.
- Administration procedures believed to be faulty.
- Scores or other assessment results believed to be incorrect.

12.7 Inform test takers who are concerned about the integrity of their scores of their relevant rights.

12.8 When test scores are used to make decisions about a test taker or to make recommendations to a test taker or third party, provide to test takers timely access to a copy of any report of test scores and test interpretation. Exceptions to this are if:

- The right has been waived explicitly in the test taker's informed consent document.

- The right has been implicitly waived through the application procedure in education, credentialing or employment testing.
- The right is prohibited by law or court order.

12.9 If scores are invalidated or not reported within the normally expected reporting time, follow prescribed procedures in accordance with established NBOME policy to protect the rights and privacy of test takers whose scores are under investigation on the grounds of irregularity or misconduct until such time as a final determination is made.

12.10 Notify a test taker and give the reasons for the investigation when an individual score report is expected to be significantly delayed beyond a brief investigative period because of possible irregularities such as suspected misconduct.

12.11 At appropriate intervals, make public the pass rates and performance of test takers in aggregate, to examinees and other appropriate parties.

Glossary

- **Ability -** The knowledge, skills or other characteristics of a test taker measured by the test.

- **Accommodation –** A modification to an assessment or its administration to allow access to the intended construct for a particular test takers. Tests or assessments with

- **Assessment –** A systematic process to measure or evaluate the characteristics of performance.

- **Audit -** A systematic evaluation of a product or service with respect to documented standards to indicate whether or not the product or service is in compliance with the standards.

- **Bias -** In general usage, unfairness. In technical usage, a systematic error in estimating a parameter.

- **Biserial correlation -** A statistic to describe the relationship between performance on a single test item and on the full test. It is an estimate of the correlation between the test score and an unobservable variable assumed to determine performance on the item and assumed to have a normal distribution.

- **Calibration -** The meaning of this term depends on the context. In item response theory (IRT), "calibration" refers to the process of estimating the numbers (called "parameters") that describe the statistical characteristics of each test question. In the scoring of a constructed-response test, "calibration" refers to the process of checking to make sure that each scorer is applying the scoring standards correctly.

- **Classical test theory -** A statistical theory that forms the basis for many calculations done with test scores, especially those involving reliability. The theory is based on partitioning a test taker's score into two components: a component called the "true score" that generalizes to other occasions of testing with the same test, and a component called "error of measurement" that does not generalize. The size of the "error of measurement" component is estimated using the standard error of measurement.

- **Comparable -** Two scores are comparable if they can be meaningfully compared. Raw scores on different forms of a test are not comparable, because the questions on one form can be more difficult than the questions on another form. Scaled scores on different forms of a test are comparable if the process of computing them includes equating. Percentile scores are comparable if they refer to the same group of test takers.

- **Confidence interval -** A range of possible values for an unknown number (such as a test taker's true score), computed in such a way as to have a specified probability of including the unknown number. That specified probability is called the "confidence level" and is usually high, typically 90 or 95.

- **Construct -** The complete set of knowledge, skills, abilities, or traits an assessment is intended to measure.

- **Converted score -** A test score that has been converted into something other than a raw score. One common type of converted score is a "scaled score" — a score that has been transformed onto a different set of numbers from those of the raw scores, usually after equating to adjust for the difficulty of the test questions. Another common type of converted score is a percentile score. Instead of "converted score," the term "derived score" is often used.

- **Correlation -** A statistic that indicates how strongly two measures, such as test scores, tend to vary together. If the correlation between scores on two tests is high, test takers tend to have scores that are about equally above average (or equally below average) on both tests. The correlation can range from -1.00 to +1.00. When there is no tendency of the scores to vary together, the correlation is .00.

- **Criterion referencing -** Making test scores meaningful *without* indicating the test taker's relative position in a group. On a criterion-referenced test, each individual test taker's score is compared with a fixed standard, rather than with the performance of the other test takers. Criterion referencing is often defined in terms of proficiency levels. The test score required to attain each proficiency level is specified in advance. The percentages of test takers at the different proficiency levels are not fixed; they depend on how well the test takers perform on the test. (Compare with norm referencing.)

- **Cut score -** A point on the test score scale used for classifying the test takers into groups on the basis of their scores. Sometimes these classifications are used only to report statistics, such as the percent of students classified as proficient in a subject. More often, the classifications have consequences for individual test takers — consequences such as being granted or denied a license to practice a profession. (See also performance level descriptor.)

- **Customer -** A general term for those who sponsor, purchase, or use NBOME products or services, including clients, institutional and individual score recipients, and test takers.

- **Customer Service -**The extent to which interactions of NBOME staff with customers increase customer satisfaction.

- **Decision error -** When test takers' scores are compared with a specified cut score, two kinds of decision errors are possible: (1) a test taker whose true score is above the cut can get a score below the cut; (2) a test taker whose true score is below the cut can get a score above the cut. It is possible to modify the decision rule to make one kind of decision error occur less often, but only at the cost of making the other kind of decision error occur more often. Also called "classification error."

- **Differential item functioning (DIF) -** Differential item functioning (DIF) is the tendency of a test question to be more difficult (or easy) for certain specified groups of test takers, after controlling for the overall ability of the groups. It is possible to perform a DIF analysis for any two groups of test takers.

- **Discrimination -** Outside the testing context, this term usually means treating people differently because they are members of particular groups, e.g., male and female. In the testing context, discrimination means something quite different. It refers to the power of a test or (more often) a test question to separate high-ability test takers from low-ability test takers.

- **Distractors -** In a multiple-choice test item, the distractors are the wrong answers presented to the test taker along with the correct answer. Writers of test questions often use distractors that represent common mistakes or misinformation.

- **Equating -** Statistically adjusting scores on different forms of the same test to compensate for differences in difficulty (usually, fairly small differences). Equating makes it possible to report scaled scores that are comparable across different forms of the test.

- **Evidence-centered design -** An approach to constructing educational assessments that uses evidentiary arguments to reveal the reasoning underlying the design of the test. The test designers begin with an analysis of the types of evidence necessary to make valid claims about what test takers know or can do.

- **Fairness -** The validity of test score interpretations for intended uses for individuals from all relevant subgroups. A test that is fair minimizes the construct-irrelevant variance associated with individual characteristics and testing contexts that otherwise would compromise the validity of scores for some individuals.

- **Formative assessment -** Assessing students' skills for the purpose of planning instruction for those students. Formative assessment is done before instruction begins and/or while it is taking place. (Compare with summative assessment.)

- **Item -** A test question, including the question itself, any stimulus material provided with the question, and the answer choices (for a multiple-choice item) or the scoring rules (for a constructed-response item).

- **Item analysis -** Statistical analyses of test takers' responses to test questions, done for the purpose of gaining information about the quality of the test questions.

- **Item banking -** Creating and maintaining a data base of test questions. The record for each question includes the text of the question and statistical information computed from the responses of test takers who have taken it.

- **Item response theory (IRT) -** A statistical theory and a set of related methods in which the likelihood of achieving each possible score on a test question depends on one characteristic of the test taker (called "ability") and a small number (usually three or fewer) of characteristics of the test question. These characteristics of the test question are indicated by numbers called "parameters." They always include the difficulty of the question and often include its discrimination (the sharpness with which it separates stronger from weaker test takers).

- **Mean (of test scores) -** The average, computed by summing the test scores of a group of test takers and dividing by the number of test takers in the group.

- **Median (of test scores) -** The point on the score scale that separates the upper half of a group of test takers from the lower half. The median has a percentile rank of 50.

- **Multiple-choice item -** A test question that requires the test taker to choose the correct answer from a limited number of possibilities, usually four or five.

- **Norm referencing -** Making test scores meaningful by providing information about the performance of one or more groups of test takers (called "norm groups"). A norm-referenced score typically indicates the test taker's relative position in the norm group. One common type of norm-referenced score is a percentile score. Another type is a "standard score," which indicates the test taker's relative position in terms of the mean (average score) and standard deviation of the scores of the group. (Compare with criterion referencing.)

- **Normalization -** Transforming test scores onto a score scale so as to produce a score distribution that approximates the symmetric, bell-shaped distribution called a "normal" distribution. Normalization is a type of scaling.

- **Normal distribution -** The symmetrical, bell-shaped distribution commonly used in many statistical and measurement applications, especially in computing confidence intervals including score bands.

- **Norms -** Statistics that describe the performance of a group of test takers (called a "norm group") for the purpose of helping test takers and test users interpret the scores. Norms information is often reported in terms of percentile ranks.

- **Percentile score (percentile rank) -** A test score that indicates the test taker's relative position in a specified group. A test taker's percentile score (also called "percentile rank") is a number from 1 to 100, indicating the percent of the group with scores no higher than the test taker's score. The most common way to compute the percentile score is to compute the percentage of the group with lower scores, plus half the percentage with exactly the same score as the test taker. (Sometimes none of the test takers with exactly that score are included; sometimes all of them are.) Percentile scores are easy for most people to understand. However, many people do not realize that averages or differences of percentile scores can be very misleading. For example, the difference between percentile scores of 90 and 95 nearly always represents a *larger* difference in performance than the difference between percentile scores of 45 and 55. Comparisons of percentile scores are meaningful only if those percentile scores refer to the same group of test takers tested on the same test.

- **Performance assessment -** A test in which the test taker actually demonstrates the skills the test is intended to measure by doing real-world tasks that require those skills, rather than by answering questions asking how to do them. Typically, those tasks involve actions other than marking a space on an answer sheet or clicking a button on a computer screen. A pencil-and-paper test can be a performance assessment, but only if the skills to be measured can be exhibited, in a real-world context, with a pencil and paper.

- **Point biserial correlation -** The actual correlation between a dichotomous variable (a variable with only two possible values) and a variable with many possible values.

- **Portfolio -** A systematic collection of materials selected to demonstrate a person's level of knowledge, skill or ability in a particular area. Portfolios can include written documents (written by the person being evaluated or by others), photos, drawings, audio or video recordings, and other media. Often the types of documents and other media to be provided are specified in detail.

- **Psychometrician -** An expert in the statistical operations associated with tests of psychological characteristics, mental abilities, or educational or occupational knowledge and skills.

- **Rasch model -** A type of item response theory that assumes that a test-taker's probability of answering a test question correctly depends on only one characteristic of the test question, its difficulty. Compare to item response theory.

- **Raw score -** A test score that has not been adjusted to be comparable with scores on other forms of the test and is *not* expressed in terms of the performance of a group of test takers.

- **Reliability -** The tendency of test scores to be consistent on two or more occasions of testing, if there is no real change in the test takers' knowledge. If a set of scores has high reliability, the test takers' scores would tend to agree strongly with their scores on another occasion of testing. The type of reliability NBOME is most often concerned about is consistency across different forms of a test. For a constructed-response test, NBOME is also concerned about the consistency of the scores assigned by different scorers (called "scoring reliability" or "inter-rater reliability").

- **Reliability coefficient -** A statistic that indicates the reliability of test scores; it is an estimate of the correlation between the scores of the same test takers on two occasions of testing with the same test (typically with different forms of the test).

- **Rubric -** A set of rules for scoring the responses on a constructed-response item. Sometimes called a "scoring guide."

- **Scaling -** Statistically transforming scores from one set of numbers (called a "score scale") to another. Some types of scaling are used to make scores on different tests comparable in some way. The most common application of scaling is to make scores on different editions ("forms") of the same test comparable.

- **Score band -** An interval around a test taker's score, intended to convey the idea that an individual's score on a test is influenced by random factors. Often, the boundaries of the score band are one standard error of measurement above and below the test taker's actual score. (A score band determined in this way is a confidence interval with a confidence level, assuming a normal distribution, of 68 percent.) Score bands illustrate the limited precision of the test score as a measure of anything beyond the test taker's performance on one occasion of testing. However, score bands can be misleading in two ways. They imply that the test taker's true score cannot lie outside the band, and they imply that all values within the band are equally likely values for the test taker's true score. Neither of these implications is correct.

- **Selected-response item -** Any type of test item in which the test-taker’s task is to select the correct answer from a set of choices. Multiple-choice items, true-false items and matching items are all selected-response items. Compare with constructed-response item.

- **Standard deviation (of test scores) -** A measure of the amount of variation in the scores of a group of test takers. It is the average distance of the scores from the group mean score (but with the average distance computed by a procedure called "root-mean-square," which is a bit more complicated than the usual procedure). The standard deviation is expressed in the same units as the scores, e.g., number of correct answers, or scaled-score points. If there are many high and low scores, the standard deviation will be large. If the scores are bunched closely together, the standard deviation will be small.

- **Standard error of measurement (SEM) -** A measure of the tendency of test takers' scores to vary because of random factors, such as the particular selection of items on the form the test taker happened to take, or the particular scorers who happened to score a test taker's responses. The smaller the SEM, the smaller the influence of these factors. The SEM is expressed in the same units as the scores themselves.

- **Standard setting -** The process, often judgment based, of setting cut scores using a structured procedure that seeks to map test scores into discrete performance levels that are usually specified by performance-based descriptors.

- **Standardized test -** A test in which the content and format of the test and the conditions of testing (such as timing, directions, use of calculators) are controlled to make them the same for all test takers. (Exceptions may be made for test takers with disabilities.)

- **Subjective scoring -** Any scoring system that requires judgment on the part of the scorer. With subjective scoring, different scorers could possibly assign different scores to the same response.

- **Summative assessment -** Assessing students' skills for the purpose of determining whether instruction has been effective. Summative assessment is done after the instruction has been completed. (Compare with formative assessment.)

- **Technical manual –** A publication prepared by test developers and/or publishers to provide technical and psychometric information about a test.

- **Test** – An evaluative instrument or procedure in which a systematic sample of a test taker's behavior in a specified domain is obtained and scored using a systematic process.

- **Test developer –** The person(s) or organization responsible for the design and construction of a test and for the documentation regarding its technical quality for an intended purpose.

- **True score -** In classical test theory, a test taker's true score on a test is defined as the average of the scores the test taker would get, averaging over some very large set of theoretically possible conditions of testing — for example, all possible forms of the test, or all possible scorers that might score the responses. It is not possible to know an individual test taker's true score, but it is possible to estimate the true scores of a large group of test takers.

- **Validity -**Validity is the extent to which the scores on a test are appropriate for a particular purpose. The validity of the scores depends on the way they are being interpreted and used. Scores on a test can be highly valid for one purpose and much less valid for another. Statistics can provide evidence for the validity of a test, but the validity of a test cannot be measured by a single statistic. Evidence for validity can include:
  - statistical relationships of test scores with other information (e.g., scores on other tests of the same or related abilities, grades, ratings of performance)
  - statistical relationships between parts of the test
  - statistical indicators of the quality and fairness of the test questions
  - the qualifications of the test designers, question writers and reviewers
  - the process used to develop the test
  - experts' judgments of the extent to which the content of the test matches a curriculum or the requirements of a job

Policy Number: 2806
Original Date: 01/01/2006
Effective Date: 06/27/2015
Next Review Date: 06/27/2018
Policy Scope: ORG
Policy Approval: GOV
Executive Responsibility: President and CEO
Policy Owner: Senior Director, Quality and Examination Integrity