Quality Testing Standards and Criteria for Comparability Claims

Version 6.17.2020

Introduction

New Meridian Corporation has developed the *Quality Testing Standards and Criteria for Comparability Claims* (QTS). The goal of the QTS is to provide guidance to states that are interested in including content from the New Meridian item bank and intend to make comparability claims with *other assessments* that include New Meridian content. Note that "other assessments" could mean New Meridian test forms administered in another state, or it could mean the state's previous assessments that include New Meridian assessment content but administered or scored by a different testing vendor going forward. (See the Potential Use Cases section for additional information.)

QTS Purpose and Organization

One of the key assumptions is that states licensing New Meridian content are interested in comparing their assessment results to those of another assessment program. For example, to maintain trendlines, a state may wish to continue reporting scores on the New Meridian Scale in its newly developed assessment that include items from the New Meridian Item Bank. This is referred to as *scale score comparability*. Or, a state may be interested in comparing the percentage of student who are on track or ready for college and careers (attain Level 4 or higher) with other states, districts or schools that administer New Meridian content and use the New Meridian Level 4 cut score to determined college and career readiness (CCR)[1]. This is called *readiness comparability*. More information about the types of comparability claims that a state might consider are described in the Comparability Claims section.

To support its intended comparability claims and provide for processes such as federal peer review, we recommend each interested state collect and submit evidence demonstrating that these types of comparisons are technically defensible. This evidence will be evaluated by independent expert reviewers to determine if the desired comparisons can be supported. If the desired comparisons cannot be supported, the reviewers will provide constructive and actionable feedback on what the state needs to do to support the comparability claims. This is referred to as the *QTS comparability review process*.

[1] "College and career readiness" (CCR) in this context refers to both being *on track* for college and careers (i.e., grade-level readiness in elementary and middle school) and *ready* for college and careers (in high school).

The criteria described in the QTS serves as the basis for the comparability review process. The QTS includes criteria for evaluating the feasibility of comparability claims in three *areas* of a testing program: design, administration, and scoring. Broadly speaking, the areas can be distinguished as follows:

- Design: "What is on the test?"
- Administration: "How is the test given?" and,
- Scoring: "How is test performance determined?"

Within each area, there are two main sections. The first section is the *Guidelines for High Quality Tests*, which describes the best practices or processes needed to support a high-quality testing program that yields valid and reliable outcomes. It serves as the foundation for the second main section, *Evaluative Criteria for Comparability Claims*. This section specifies the expectations for participating state that wish to make comparability claims based on New Meridian assessment content. It includes two subsections:

- *Supporting Evidence*. This is a list of evidence that each participating state can provide about its testing program to meet the requirements in the evaluative criteria. Across the three areas, the supporting evidence include six key aspects of the testing program:
    1. Item and Test Development (Design area)
    2. Fairness and Accessibility (Design area)
    3. Test Administration (Administration area)
    4. Item Scoring (Scoring area)
    5. Psychometrics (Scoring area)
    6. Standard Setting (Scoring area)

- *Criteria for Comparability Evaluation*. This is a table that summarizes the degree of similarity expected between a state's testing program and New Meridian test forms for two types of comparability claims: *scale score comparability* and *readiness comparability*. The table also provides the supporting evidence that will be reviewed to evaluate the degree of similarity, and ultimately to determine the type of comparability claims that a participating state can support. The working definition of the two types of comparability claims are:

- *Scale score comparability*[2]: *If a student taking the state's assessment with New Meridian content took one of the test forms offered by New Meridian, would he or she obtain the same scale score?*
- *Readiness comparability*[3]*: If a student taking the state's assessment with New Meridian content took one of the test forms offered by New Meridian, would he or she receive the same designation in terms of college and career readiness?*

The guidelines, criteria and list of supporting evidence in the QTS are based on the 2014 edition of the *Standards for Educational and Psychological Testing*. The *Standards* were developed jointly by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) and are recognized as the industry standard for assessment best practices. Any statement or requirement based on elements in the *Standards* are indicated with (blue parentheses), which includes the number reference to the specific standard. A full list of the standards referred to in the QTS are given in the Appendix.

Comparability Claims

The nature and strength of comparability claims that a state can make depends on several factors, including blueprint alignment, content coverage, alignment of performance standards, and quality and integrity of administration and scoring protocols. The QTS assumes three broad categories of comparability claims: full comparability, scale score comparability, and readiness comparability. There is also an additional category of no comparability claims.

Full Comparability

States that use New Meridian's flagship or Alternative Blueprinting Option (ABO) forms with the primary administration vendor, Pearson, can claim full comparability. This means that comparisons can be made with overall (summative) performance levels and scales scores and with major claim scores and subclaim classification across years and between states that can also claim full comparability. Such comparisons can be made at both the student and aggregate (e.g., schools, district, state) levels.

Scale Score Comparability

Scale score comparability means that a state can make comparisons with the overall (summative) scales scores across years and between states that use the New Meridian reporting scale. For example, a scale score of 725 attained on the state's assessment would mean the same as if it

[2] A more technical way to frame the scale score comparability question is: *Does the evidence submitted by the state for its assessments with New Meridian content meet the requirement of score equating between its forms and the New Meridian test forms*?

[3] A more technical way to frame the readiness comparability question is: *Does the evidence submitted by the state for its assessments with New Meridian content support an equivalent degree of rigor and interpretation in terms of college and career readiness as the New Meridian test forms*?

were achieved on the New Meridian flagship or ABO forms administered by its primary administration vendor. However, the comparisons may not be valid for major claims scores (i.e., reading or writing scores in ELA/L) or subclaim classifications. If the state adopts the New Meridian overall performance levels, including the same performance level descriptors (PLDs), then a state that can meet the requirements for scale score comparability can also make comparisons with the overall performance levels (e.g., percentage of students at Levels 1, 2, 3, etc.) across years and between states that use the New Meridian performance levels. States intending to make claims of scale score comparability can consider one of the following two approaches.

1. *States administer New Meridian-designed assessment forms*. States may elect to have the New Meridian operational flagship or ABO forms administered by its own administration vendor. Such a state should provide evidence about its administration procedures, scoring processes, and annual equating procedures to confirm that they are similar to the standard processes outlined in the QTS.

2. *States embed an anchor set of items to support cross-form equating*. States may choose to administer anchor sets comprised of New Meridian content, constructed annually through the New Meridian test construction process. An anchor set is a subset of items from a New Meridian test form that New Meridian offers for each grade level and subject area. Anchor sets are constructed to be representative of the content, depth of knowledge, and psychometric properties of the New Meridian test form. States can augment the anchor set with their own content to create a full-length test form. States intending to claim scale score comparability while using a New Meridian anchor set with augmented content must provide evidence that the augmented test still aligns with the New Meridian test blueprints, and that scoring rules and scaling and equating procedures are very similar to those outlined in the QTS.

Once the alignment and quality of these procedures are validated, states can make comparability claims about summative scale scores at both the student and aggregate levels. For example, State X using an anchor set can compare its average scale score of 728 on the Grade 8 Mathematics test to State Y using a New Meridian test form with an average scale score of 715.

Readiness Comparability

States that embed a New Meridian anchor set but build complete test forms that do not closely align with the New Meridian test blueprint can still make comparability claims about performance related to the New Meridian college and career readiness (i.e. attaining Level 4 or higher). States must provide evidence that its blueprint and administration, scoring and reporting procedures adequately align with those outlined in the QTS such that the state can still make dichotomous (yes/no) comparability claims about being ready for the next course or grade level (in elementary or middle school) or for college and careers (in high school) at both the student and aggregate levels. For example, once validated through the QTS process, State X may compare at an aggregate level its 67% of students who are college and career ready in Algebra 1 (i.e., attain Level 4 or higher) to the 55% of students who are college and career ready in State Y.

No Comparability Claims

States may be interested only in licensing content from the New Meridian item bank and not interested in making any comparability claims based on the New Meridian scale. Such states do not need to use the anchor sets, nor participate in the QTS validation process. They may, of course, calibrate and scale licensed content and perform standard setting to support reporting on their own state scale.

Summary

The table below summarizes and contrasts the different types of comparability claims assumed by the QTS framework.

Category	Full Comparability	Scale Score Comparability	Readiness Comparability	No Comparability
Content Model	New Meridian test form	Anchor set	Anchor set	Item bank only
Degree of Similarity to New Meridian Processes	Identical	High	Adequate	No requirements
Summative	Yes	Yes	Yes	No
Major Claims	Yes	No	No	No
Subclaims	Yes	No	No	No
Scale Score	Yes	Yes	No	No
College and Career Readiness (CCR)	Yes	Yes	Yes	No
Student-Level Comparisons	Yes	Yes	Yes	No
Aggregate-Level Comparisons	Yes	Yes	Yes	No

Potential Use Cases

As previously mentioned, the goal of the QTS is to provide guidance to states that are interested in including New Meridian content and intend to make comparability claims with *other assessments* that include New Meridian content. The term "other assessments" could mean something different for each state, depending on the goal for including New Meridian tasks or items on its operational test forms and the requirements for its testing program. The table below describes several potential use cases for tests that include New Meridian content.

Use Case	Description of Use Case
State-licensed "New Meridian" forms	The state licenses New Meridian content with test forms designed to match the specifications and blueprints for New Meridian test forms. The state contracts its own vendor for the other steps in the operational administration process, including delivery and scoring.
State-licensed "New Meridian" forms, supplemented with state-developed content	The state licenses New Meridian content, but also includes content from its own (state-specific) item bank. The test forms are designed to match the specifications and blueprints for the New Meridian test forms. The state contracts its own vendor for the other steps in the operational administration process, including delivery and scoring.
State developed assessments, supplemented with New Meridian content	The state develops its own test items but also licenses New Meridian content. The test forms are designed to match state-developed test specifications and blueprints. The state contracts its own vendor for test development, administration, and scoring.

QTS Supporting Documents and Usage Roadmap

To support the usage of the QTS in the comparability review process, three additional documents have also been developed. The QTS supporting documents include:

- *State QTS Starter Kit*: This document is intended for anyone supporting a state that is considering New Meridian content for its assessment program. It includes a questionnaire for collecting information about the state's goals and priorities and a checklist of potential evidence that the state can provide for the comparability review process.
- *Standard Processes*: This document describes the various processes currently implemented for the New Meridian test forms. It serves as the point of reference to which the state's assessment is compared. The intended audience is independent reviewers in the QTS comparability

review process, or anyone interested in learning more about the test development and administration processes for building the New Meridian test forms.

- *Comparability Review Guidelines*: This document provides a concrete framework for independent reviewers to follow in their evaluation of a state's submitted evidence. The intended audience is independent reviewers in the QTS comparability review process.

A roadmap of how the QTS and its supporting documents can be used in the comparability review process is illustrated in Figure 1. In this figure, the first (red) box are documents that provide information about the *standards and criteria* by which the state's comparability claims are evaluated. The second (green) box is the State *QTS Starter Kit*, which helps states understand the types of *evidence* that would help support comparability claims. The third (purple) box is the *New Meridian Comparability Review Guidelines*, which provide concrete *guidance* on how an expert reviewer compares the evidence in the second box with the standards and criteria in the first box.

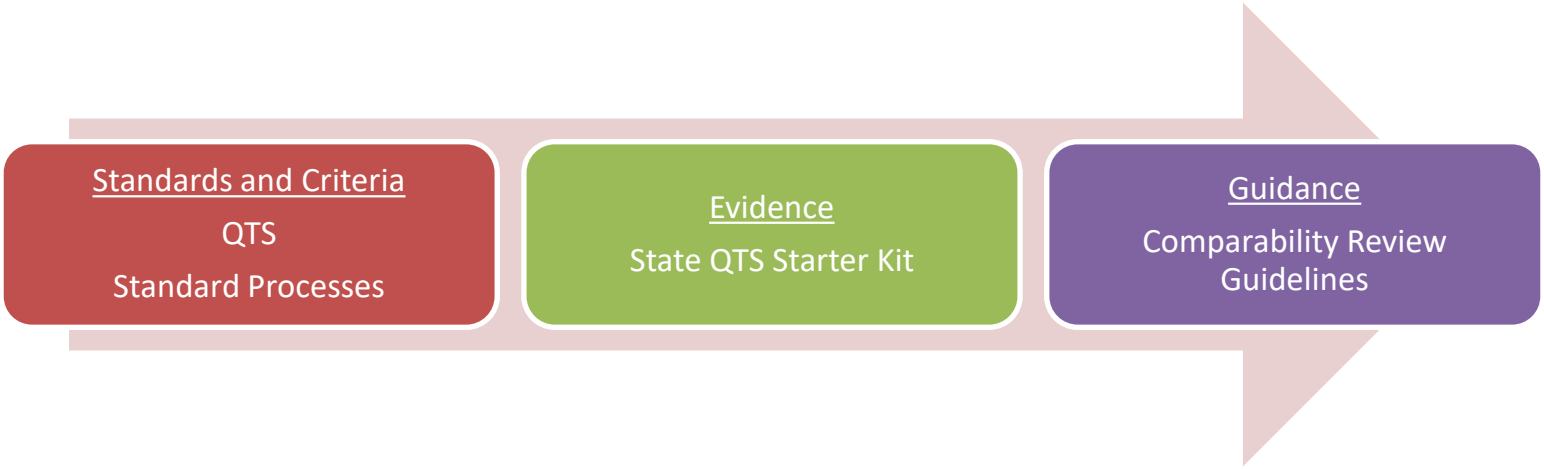

Figure 1: Suggested Roadmap for Using the QTS and its Supporting Documents

Design[4]

Guidelines for High Quality Tests

Tests should be designed and developed in a way that supports the validity of interpretations of the test scores for their intended uses. Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population. (4.0)

Evaluative Criteria for Comparability Claims

The comparability evaluation focuses on the design of the state's assessments with New Meridian assessment content (e.g., purpose, content representation, item types) and the procedures informing its development. The overarching question that the comparability review process should address is:

Are the specifications and procedures underlying the design and development of the state's assessment with New Meridian content comparable to those of the New Meridian test forms?

Supporting Evidence[5]

1. *Item and Test Development*. Documentation or materials from the state's item and test development process that describe:
    a. Test purpose, target population and intended uses (4.1, 4.12);
    b. Assessed content standards, item types, rubrics, blueprints, test formats, eligible content, and time limits, along with the rationale for the test design decisions (4.2, 4.14, 4.23);
    c. Processes for evaluating newly developed items, including subject matter expert committees, field testing and data review (4.6, 4.7, 4.8); and,
    d. Forms construction and review procedures (4.3, 4.4).
2. *Fairness and Accessibility*. Steps in the test design process that minimize construct irrelevant variance and promote valid score interpretations for the intended uses for all examinees in the intended population. This may include documentation that supports:
    a. Universal design principles (3.10); and,

[4] Any statement or requirement based on elements in the *Standards for Educational and Psychological Testing* are indicated with (blue parentheses), which includes the number reference to the specific standard. A full list of the standards referred to in the QTS are given in the Appendix.

[5] The *New Meridian Comparability Evaluation Checklist* includes specific examples of potential sources of supporting evidence from a state's testing program, along with a checklist that can used to organize the documents and materials that a state is submitting for the comparability review process.

b. Accommodations for English learners and students with disabilities, and procedures used to translate forms for students for whom English is a second language (3.2, 3.9, 3.12, 3.13, 3.14).

Criteria for Comparability Evaluation

The following table summarizes the degree of similarity expected between a state's test design and New Meridian's test forms for each type of comparability claim. The table also indicates whether the state's supporting evidence for each criterion should be considered in the evaluation of the two types of comparability claims. "Yes" means that evidence of comparability between the assessments for this criterion is required; while "Recommended" means that evidence of comparability for this criterion would help the state's case. Even if the cell for a criterion is blank, the reviewer is encouraged to consider the evidence submitted by the state as it may be informative in evaluating the state's case for making comparability claims. A detailed elaboration of the how each criterion is evaluated in the comparability review process can be found in the QTS supporting document, *New Meridian Comparability Review Guidelines*.

Criterion/ Type of Comparability Claim	Scale Score Comparability	Readiness Comparability
Overall Degree of Similarity with New Meridian's Form Design	**High**	**Adequate**
Criterion 1a Test purpose and uses	Recommended	Recommended
Criterion 1b Test and item specifications	Yes	Yes
Criterion 1c Evaluation of new items	Recommended	
Criterion 1d Forms construction	Yes	Yes
Criterion 2a Universal design	Recommended	Recommended

Criterion/ Type of Comparability Claim	Scale Score Comparability	Readiness Comparability
Criterion 2b **Accommodations and language translation**	Recommended	Recommended

Administration[6]

Guidelines for High Quality Tests

To support useful interpretations of score results, the test should have established procedures for test administration that provide clear directions and instructions for test administrators and test takers, consider all supported testing conditions, address the needs of all students, and include sound security protocols.

Evaluative Criteria for Comparability Claims

The comparability evaluation will focus on test administration processes (such as testing time, directions and instructions to administrators and test takers, accommodations allowed, etc.) for the state's assessment with New Meridian assessment content. The overarching question that the comparability review process should address is:

Are the policies and procedures for administering the state's assessment that include New Meridian content comparable to those of the New Meridian test forms?

Supporting Evidence[7]

3. *Test Administration*. Documentation or materials about the state's test administration procedures that include:
    a. Training and instructions provided to test administrators and coordinators (4.15, 6.1);
    b. Instructions given to test takers (4.16);
    c. Information about the modes of administration, such as paper-based vs. computer-based testing, and fixed-form vs. adaptive tests, including rationale for offering the test in each mode (4.3, 4.4, 4.5, 4.6);
    d. Details about test security protocols (6.6, 6.7); and,
    e. Evidence that supports accessibility of the test to all students as part of the test administration (3.4).

Criteria for Comparability Evaluation

The following table summarizes the degree of similarity expected between the test administration processes for the state's assessments and New Meridian's test forms for each type of comparability claim. The table also indicates whether the state's supporting evidence for each criterion should be considered in the evaluation of the two types of comparability claims. "Yes" means that evidence of comparability between

[6] Any statement or requirement based on elements in the *Standards for Educational and Psychological Testing* are indicated with (blue parentheses), which includes the number reference to the specific standard. A full list of the standards referred to in the QTS are given in the Appendix.

[7] The *Comparability Evaluation Checklist* includes specific examples of potential sources of supporting evidence from a state's testing program, along with a checklist that can used to organize the documents and materials that a state is submitting for the comparability review process.

the assessments for this criterion is required; while "Recommended" means that evidence of comparability for this criterion would help the state's case. Even if the cell for a criterion is blank, the reviewer is encouraged to consider the evidence submitted by the state as it may be informative in evaluating the state's case for making comparability claims. A detailed elaboration of the how each criterion is evaluated in the comparability review process can be found in the QTS supporting document, *New Meridian Comparability Review Guidelines*.

Criterion/ Type of Comparability Claim	Scale Score Comparability	Readiness Comparability
Overall Degree of Similarity with Administration Processes for New Meridian Test Forms	**High**	**High**
Criterion 3a Administrator guidance	Yes	Yes
Criterion 3b Test taker instructions	Yes	Yes
Criterion 3c Administration modes and designs	Recommended	Recommended
Criterion 3d Security protocols	Recommended	Recommended
Criterion 3e Accessibility supports	Yes	Yes

Scoring[8]

Guidelines for High Quality Tests

Test scores should be derived in a way that supports the interpretations of test scores for the proposed uses of tests. Test developers and users should document evidence of fairness, reliability, and validity of test scores for their proposed use. (5.0)

Evaluative Criteria for Comparability Claims

The comparability evaluation will focus any of the state's processes that contribute to the derivation of tests scores on the state's assessment with New Meridian content. The processes include item scoring (machine and/or human), psychometric procedures, and the performance level setting method. The overarching question that the comparability review process should address is:

Are the criteria, methodologies and procedures for scoring the state's assessment with New Meridian content comparable to those of the New Meridian test forms?

Supporting Evidence[9]

4. *Item Scoring*. Documentation or materials about the item scoring process that describe:
    a. Training and qualification procedures for human scorers (1.9, 4.18, 4.20, 6.9);
    b. Protocols for both machine and human scoring processes, and evidence that the scoring process is fair to all students (3.5, 3.8, 6.8); and,
    c. If used, validation of automated scoring processes (4.19).
5. *Psychometrics*. Technical specifications, briefs, or reports about the psychometric processes that detail:
    a. Choice of psychometric models (4.10);
    b. Scaling and equating design and procedures, including quality control processes (5.2, 5.4, 5.12, 5.13, 5.14, 5.15);
    c. Analysis of disaggregated student groups (3.6);
    d. Sampling, including purpose and methodology (1.8, 3.3, 4.9); and,
    e. Other psychometric procedures or analyses that support the reliability and validity of test scores (2.19, 5.16, 5.17, 5.18).
6. *Standard Setting*. Materials or reports about the procedures used to establish performance standards including those that describe:

[8] Any statement or requirement based on elements in the *Standards for Educational and Psychological Testing* are indicated with (blue parentheses), which includes the number reference to the specific standard. A full list of the standards referred to in the QTS are given in the Appendix.

[9] The New Meridian *Comparability Evaluation Checklist* includes specific examples of potential sources of supporting evidence from a state's testing program, along with a checklist that can used to organize the documents and materials that a state is submitting for the comparability review process.

a. Achievement or performance level descriptors (ALDs or PLDs) and how they were established (5.21);
b. Standard setting methodology and procedures (5.22); and,
c. Empirical support for the readiness (on-track or ready for college and careers) cut scores (5.23).

Criteria for Comparability Evaluation

The following table summarizes the degree of similarity expected between the scoring processes for the state's assessments and New Meridian's test forms for each type of comparability claim. The table also indicates whether the state's supporting evidence for each criterion should be considered in the evaluation of the two types of comparability claims. "Yes" means that evidence of comparability between the assessments for this criterion is required; while "Recommended" means that evidence of comparability for this criterion would help the state's case. Even if the cell for a criterion is blank, the reviewer is encouraged to consider the evidence submitted by the state as it may be informative in evaluating the state's case for making comparability claims. A detailed elaboration of the how each criterion is evaluated in the comparability review process can be found in the QTS supporting document, *New Meridian Comparability Review Guidelines*.

Criterion/ Type of Comparability Claim	Scale Score Comparability	Readiness Comparability
Degree of Similarity with Scoring Processes for New Meridian Test Forms	**High**	**Adequate**
Criterion 4a Scorer qualification	Yes, if applicable	
Criterion 4b Scoring protocols	Yes	
Criterion 4c Automated scoring	Yes, if applicable	
Criterion 5a Psychometric models	Yes	
Criterion 5b Scaling and equating	Yes	
Criterion 5c Student group analysis	Recommended	

Criterion/ Type of Comparability Claim	Scale Score Comparability	Readiness Comparability
Criterion 6a PLDs	Recommended, if applicable[10]	Yes
Criterion 6b Standard setting method	Recommended, if applicable	Yes
Criterion 6c Validity of cut scores	Recommended, if applicable	Yes

[10] In general, the criteria for standard setting (criteria 6a to 6c) apply only to the evaluation of readiness comparability. The rationale is that if a state has provided evidence to support scale score comparability, then it can make use of the New Meridian cut scores, which were established on the New Meridian scale. *If* the state, however, makes any adjustments to the existing New Meridian performance standards (e.g., modifies PLDs, changes cut scores etc.), then a standards validation process would likely be needed. Evidence from such a process should be provided and reviewed against criteria 6a to 6c.

Appendix – References from the *Standards for Educational and Psychological Testing*

Section 1 – Validity

Standard 1.8

The composition of any sample of test takers from which validity evidence is obtained should be described in as much detail as is practical and permissible, including major relevant socio-demographic and development characteristics.

Standard 1.9

When a validation rests in part on the opinions or decisions of expert judges, observers, or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described. The qualifications and experience of the judges should be presented. The description of procedures should include any training and instructions provided, should indicate whether participants reached their decisions independently, and should report the level of agreement reached. If participants interacted with one another or exchanged information, the procedures through which they may have influenced one another should be set forth.

Standard 1.10

When validity evidence includes statistical analyses of test results, either alone or together with data on other variables, the conditions under which the data were collected should be described in enough detail that users can judge the relevance of the statistical findings to the local conditions. Attention should be drawn to any features of a validation data collection that are likely to differ from typical operational testing conditions and that could plausibly influence test performance.

Standard 1.17

When validation relies on evidence that test scores are related to one or more criterion variables, information about the suitability and technical quality of the criteria should be reported.

Standard 1.18

When it is asserted that a certain level of test performance predicts adequate or inadequate criterion performance, information about the levels of criterion performance associated with given levels of test scores should be provided.

Standard 1.25

When unintended consequences result from test use, an attempt should be made to investigate whether such consequences arise from the test's sensitivity to characteristics other than those it is intended to assess or from the test's failure to fully represent the intended construct.

Section 2 – Reliability/Precision and Errors of Measurement

Standard 2.1

The range of replications over which reliability/precision is being evaluated should be clearly stated, along with a rationale for the choice of this definition, given the testing situation.

Standard 2.2

The evidence provided for the reliability/precision of the scores should be consistent with the domain of replications associated with the testing procedures, and with the intended interpretations for use of the test scores.

Standard 2.4

When a test score interpretation emphasizes differences between two observed scores of an individual or two averages of a group, reliability/ precision data, including standard errors, should be provided for such differences.

Standard 2.5

Reliability estimation procedures should be consistent with the structure of the test.

Standard 2.7

When subjective judgment enters into test scoring, evidence should be provided on both interrater consistency in scoring and within-examinee consistency over repeated measurements. A clear distinction should be made among reliability data based on (a) independent panels of raters scoring the same performance or products, (b) a single panel scoring successive performances or new products, and (c) independent panels scoring successive performances or new products.

Standard 2.9

When a test is available in both long and short versions, evidence for reliability/precision should be reported for scores on each version, preferably based on independent administration(s) of each version with independent samples of test takers.

Standard 2.10

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

Standard 2.11

Test publishers should provide estimates of reliability/precision as soon as feasible for each relevant subgroup for which the test is recommended.

Standard 2.19

Each method of quantifying the reliability/precision of scores should be described clearly and expressed in terms of statistics appropriate to the method. The sampling procedures used to select test takers for reliability/precision analyses and the descriptive statistics on the samples, subject to privacy obligations where applicable, should be reported.

Section 3 – Fairness in Testing

Standard 3.0

All steps in the testing process, including test design, validation, development, administration, and scoring procedures, should be designed in such a manner as to minimize construct-irrelevant variance and to promote valid score interpretations for the intended uses for all examinees in the intended population.

Standard 3.1

Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population.

Standard 3.2

Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, and other characteristics.

Standard 3.3

Those responsible for test development should include relevant subgroups in validity, reliability/precision, and other preliminary studies used when constructing the test.

Standard 3.4

Test takers should receive comparable treatment during the test administration and scoring process.

Standard 3.5

Test developers should specify and document provisions that have been made to test administration and scoring procedures to remove construct-irrelevant barriers for all relevant subgroups in the test-taker population.

Standard 3.6

Where credible evidence indicates that test scores may differ in meaning for relevant subgroups in the intended examinee population, test developers and/or users are responsible for examining the evidence for validity of score interpretation for intended uses for individuals from those subgroups. What constitutes a significant difference in subgroup scores and what actions are taken in response to such differences may be defined by applicable laws.

Standard 3.8

When tests require the scoring of constructed responses, test developers and/or users should collect and report evidence of the validity of score interpretations for relevant subgroups in the intended population of test takers for the intended use of the test scores.

Standard 3.9

Test developers and/or test users are responsible for developing and providing test accommodations, when appropriate and feasible, to remove construct-irrelevant barriers that otherwise would interfere with examinees' ability to demonstrate their standing on the target constructs.

Standard 3.10

When test accommodations are permitted, test developers and/or test users are responsible for documenting standard provisions for using the accommodation and for monitoring the appropriate implementation of the accommodation.

Standard 3.11

When a test is changed to remove barriers to the accessibility of the construct being measured, test developers and/or users are responsible for

obtaining and documenting evidence of the validity of score interpretations for intended uses of the changed test, when sample sizes permit.

Standard 3.12

When a test is translated and adapted from one language to another, test developers and/or test users are responsible for describing the methods used in establishing the adequacy of the adaptation and documenting empirical or logical evidence for the validity of test score interpretations for intended use.

Standard 3.13

A test should be administered in the language that is most relevant and appropriate to the test purpose.

Standard 3.14

When testing requires the use of an interpreter, the interpreter should follow standardized procedures and, to the extent feasible, be sufficiently fluent in the language and content of the test and the examinee's native language and culture to translate the test and related testing materials and to explain the examinee's test responses, as necessary.

Standard 3.15

Test developers and publishers who claim that a test can be used with examinees from specific subgroups are responsible for providing the necessary information to support appropriate test score interpretations for their intended uses for individuals from these subgroups.

Standard 3.17

When aggregate scores are publicly reported for relevant subgroups – for example, males and females, individuals of differing socioeconomic status, individuals differing by race/ethnicity, individuals with different sexual orientation, individuals with diverse linguistic and cultural backgrounds, individuals with disabilities, young children or older adults – test users are responsible for providing evidence of comparability and for including cautionary statement when credible research or theory indicates that test scores may not have comparable meaning across these subgroups.

Section 4 – Test Design and Development

Standard 4.0

Tests and testing programs should be designed and developed in a way that supports the validity of interpretations of the test scores for their

intended uses. Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population.

Standard 4.1

Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting interpretations and uses of test results for the intended purpose(s).

Standard 4.2

In addition to describing intended uses of the test, the test specifications should define the content of the test, the proposed test length, the item formats, the desired psychometric properties of the test items and the test, and the ordering of items and sections. Test specifications should also specify the amount of time allowed for testing; directions for the test takers; procedures to be used for test administration, including permissible variations; any materials to be used; and scoring and reporting procedures. Specifications for computer-based test should include a description of any hardware and software requirements.

Standard 4.3

Test developers should document the rationale and supporting evidence for the administration, scoring, and reporting rules used in computer-adaptive, multistage-adaptive, or other tests delivered using computer algorithms to select items. This documentation should include procedures used in selecting items or set of items for administration, in determining the starting point and termination conditions for the test, in scoring the test, and in controlling item exposure.

Standard 4.4

If test developers prepare different versions of a test with some change to the test specifications, they should document the content and psychometric specifications of each version. The documentation should describe the impact of differences among versions on the validity of score interpretations for intended uses and on the precision and comparability of scores.

Standard 4.5

If the test developer indicates that the conditions of administration are permitted to vary from one test taker or group to another, permissible variation in conditions for administration should be identified. A rationale for permitting the different conditions and any requirements for permitting the different conditions should be documented.

Standard 4.6

When appropriate to documenting the validity of test score interpretations for intended uses, relevant experts external to the testing program should review the test specifications to evaluate their appropriateness for intended uses of the test scores and fairness for intended test takers. The purpose of the review, the process by which the review is conducted, and the results of the review should be documented. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented.

Standard 4.7

The procedures used to develop, review, and try out items and to select items from the item pool should be documented.

Standard 4.8

The test review process should include empirical analysis and/or the use of expert judges to review items and scoring criteria. When expert judges are used, their qualifications, relevant experiences, and demographic characteristics should be documented, along with the instructions in training in the item review process that the judges receive.

Standard 4.9

When item or test form tryouts are conducted, the procedures used to select the sample(s) of test takers as well as the resulting characteristics of the sample(s) should be documented. The sample(s) should be as representative as possible after population(s) from which the test is intended.

Standard 4.10

When a test developer evaluates the psychometric properties of items, the model used for that purpose (e.g., classical test theory, item response theory, or another model) should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are screened in the data used for screening, such as item difficulty, item discrimination, or differential item functioning (DIF) for major examinee groups, should also be documented. When model-based methods (e.g., IRT) are used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented.

Standard 4.12

Test developers should document the extent to which the content domain of a test represents the domain defined in the test specifications.

Standard 4.14

For a test that has a time limit, test development research should examine the degree to which scores include a speed component and should evaluate the appropriateness of that component, given the domain the test is designed to measure.

Standard 4.15

The directions for test administration should be presented with sufficient clarity so that it is possible for others to replicate the administration conditions under which the data on reliability, validity, and (where appropriate) norms were obtained. Allowable variations in administration procedures should be clearly described. The process for reviewing requests for additional testing variations should also be documented.

Standard 4.16

The instructions presented to test takers should contain sufficient detail so that test takers can respond to a task in the manner that the test developer intended. When appropriate, sample materials, practice or sample questions, criteria for scoring, and a representative item identified with each item format or major area in the test's classification or domain should be provided to the test takers prior to the administration of the test, or should be included in the testing material as part of the standard administration instructions.

Standard 4.18

Procedures for scoring and, if relevant, scoring criteria, should be presented by the test developer with sufficient detail and clarity to maximize the accuracy of scoring. Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical for extended-response items such as performance tasks, portfolios, and essays.

Standard 4.19

When automated algorithms are to be used to score complex examinee responses, characteristics of responses at each score level should be documented along with the theoretical and empirical bases for the use of the algorithms.

Standard 4.20

The process for selecting, training, qualifying, and monitoring scorers should be specified by the test developer. The training materials, such as the scoring rubrics and examples of test-takers' responses that illustrate the levels on the rubric score scale, and the procedures for training scorers should result in a degree of accuracy and agreement among scorers that allows the scores to be interpreted as originally intended by the test developer. Specifications should also describe processes for assessing score consistency and potential drift over time in raters' scoring.

Standard 4.23

When a test score is derived from the differential weighting of items or subscores, the test developer should document the rationale and process used to develop, review, and assign item weights. When the item weights are obtained based on empirical data, the sample used for obtaining item weights should be representative of the population for which the test is intended and large enough to provide accurate estimates of optimal weights. When the item weights are obtained based on expert judgment, the qualifications of the judges should be documented.

Section 5 – Scores, Scales, Norms, Score Linking, and Cut Scores

Standard 5.0

Test scores should be derived in a way that supports the interpretations of test scores for the proposed uses of tests. Test developers and users should document evidence of fairness, reliability, and validity of test scores for their proposed use.

Standard 5.2

The procedures for constructing scales used for reporting scores and the rationale for these procedures should be described clearly.

Standard 5.4

When raw scores are intended to be directly interpretable, their meanings, intended interpretations, and limitations should be described and justified in the same manner as is done for scale scores.

Standard 5.6

Testing programs that attempt to maintain a common scale over time should conduct periodic checks of the stability of the scale on which scores are reported.

Standard 5.12

A clear rationale and supporting evidence should be provided for any claim that scale scores earned on alternate forms of a test may be used interchangeably.

Standard 5.13

When claims of form-to-form score equivalence are based on equating procedures, detailed technical information should be provided on the

method by which equating functions were established and on the accuracy of the equating functions.

Standard 5.14

In equating studies that rely on the statistical equivalence of examinee groups receiving different forms, methods of establishing such equivalence should be described in detail.

Standard 5.15

In equating studies that employ an anchor test design, the characteristics of the anchor test and its similarity to the forms being equated should be presented, including both content specifications and empirically determined relationships among test scores. If anchor items are used in the equating study, the representativeness and psychometric characteristics of the anchor items should be presented.

Standard 5.16

When test scores are based on model-based psychometric procedures, such as those used in computerize adaptive or multistage testing, documentation should be provided to indicate that the scores have comparable meeting over alternate sets of test items.

Standard 5.17

When scores on tests that cannot be equated are linked, direct evidence of score comparability should be provided, and the examinee population for which score comparability applies should be specified clearly. The specific rationale and the evidence required will depend in part on the intended uses for which score comparability is claimed.

Standard 5.18

When linking procedures are used to relate scores on tests or test forms that are not closely parallel, the construction, intended interpretation, and limitations of those linkings should be described clearly.

Standard 5.21

When proposed score interpretations involves one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly.

Standard 5.22

When cut scores defining pass-fail or proficiency levels are based on direct judgments about the adequacy of item or test performances, the

judgmental process should be designed so that the participants providing the judgments can bring their knowledge and experience to bear in a reasonable way.

Standard 5.23

When feasible and appropriate, cut scores defining categories with distinct substantive interpretations should be informed by sound empirical data concerning the relation of test performance to the relevant criteria.

Section 6 – Test Administration, Scoring, Reporting, and Interpretation

Standard 6.1

Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user.

Standard 6.6

Reasonable efforts should be made to ensure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent or deceptive means.

Standard 6.7

Test users have the responsibility of protecting the security test materials at all times.

Standard 6.8

Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented.

Standard 6.9

Those responsible for test scoring should establish and document quality control processes and criteria. Adequate training should be provided. The quality of scoring should be monitored and documented. Any systematic source of scoring errors should be documented and corrected.