TEST VALIDITY

As described in the AERA, APA, and NCME *Standards for Educational and Psychological Testing* (2014), "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests.... The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations" (p. 11). Various types of evidence may be considered in establishing the validity of test score interpretations, and a number of methods are typically used to gather such evidence.

The validation process used by Massachusetts Department of Elementary and Secondary Education (the Department) and the Evaluation Systems group of Pearson followed professionally accepted procedures for developing a validity argument to support the interpretation of the Massachusetts Tests for Educator Licensure (MTEL) test scores. The validation process focused primarily on establishing that the content of the tests is appropriate for the purposes of the testing program. In addition, the Department and Evaluation Systems provide guidance to test takers, educator preparation programs, and statewide stakeholders regarding the appropriate interpretation and use of program test scores.

Throughout the test development process, the Department and Evaluation Systems aim to enhance the validity evidence supporting the interpretation of test scores as recommended by the *Standards for Educational and Psychological Testing* (2014). The steps taken by the Department and Evaluation Systems to provide validity evidence are outlined below. Refer to the MTEL Technical Manual for complete information about the test development steps and test data for the MTEL.

The MTEL program is custom-developed with the extensive involvement of Massachusetts educators. The collaborative MTEL test development process involves the combined expertise of Massachusetts classroom educators and administrators, educator preparation faculty, psychometric experts, researchers engaged in the study of teaching and learning, and the Department's policy and program personnel. Public school educators and educator preparation program faculty play critical roles in the test development process. This document outlines the major test development milestones and the role of educators in the process.

- **Establishing the basis for the test.** The purpose of the testing program—to support state educator licensure decisions—and the test areas to be assessed are established by state rules and regulations.
- **Defining the test objectives.** AERA, APA, and NCME Standard 11.2 states that "evidence of validity based on test content requires a thorough and explicit definition of the content domain of interest" (p. 160). The test objectives describe the content knowledge that the practitioner must possess to practice appropriately and, therefore, define eligible test content. These test objectives are reviewed, revised, and approved by practicing educators and faculty at educator preparation institutions.
- **Conducting content validation of the test objectives.** "When test content is a primary source of validity evidence in support of the interpretation for the use of a test for employment decisions or credentialing, a close link between test content and the job or professional/occupational requirements should be demonstrated." *Standards for Educational and Psychological Testing (*Standard 11.3 , *AERA, APA & NCME, 2014)* Content validation of the test objectives occurs through correlation with documentation of content requirements as well as through a survey of job incumbents.

    1. Correlation with Documentation of Content Requirements. Test objectives are aligned with relevant laws and regulations, and with student and national standards, where available, to provide documentation of the basis of the test objectives. Thus, the content of the tests is verified as being relevant.
    2. Content Validation Survey. A Content Validation Survey of the proposed test objectives is conducted to verify that the test objectives reflect current educational practice in Massachusetts. Massachusetts public school educators and college and university faculty are surveyed, *results are analyzed, and only those objectives found to be important to the job of a* Massachusetts educator are eligible to be measured by the tests.

Copyright © 2016 Pearson Education, Inc. or its affiliate(s). All rights reserved.

Pearson and its logo are trademarks, in the U.S. and/or other countries, of Pearson Education, Inc. or its affiliate(s).

- **Developing test items.** Test items are developed and reviewed with specific reference to licensing and job requirements. Licensed practitioners and educators who serve on the various advisory committees conduct reviews of test items with reference to these requirements. The Content Review Committee reviews the test questions and discusses recommendations for revisions, replacement or deletion of questions according to the consensus judgements of the committee members. During test item review meetings, committees of educators are asked to review each item and consider its alignment to the test objectives and requirements as well as its accuracy, freedom from bias, and job-relatedness.
- **Preventing bias.** Guarding against bias in the test materials involves the collaboration of educators and reviewers focused on excluding language, content, or perspectives that might disadvantage examinees based on background characteristics irrelevant to the purpose of the test, and on including content and perspectives that reflect the diversity of a state's population. The Bias Review Committee (BRC) reviews test materials for sensitivity and fairness to help ensure that the test materials are free from bias. The committee confirms that the test questions do not include language or content that might disadvantage or offend an examinee because of her or his gender, race, nationality, national origin, ethnicity, religion, age, sexual orientation, disability, or cultural, economic, or geographic background. In addition, educators from diverse backgrounds are invited to participate in the test development process. They serve as members of Content Advisory Committees (CACs), reviewing the test objectives and draft test items for each test field.
- **Pilot testing of items.** Educator licensure candidates participate in the pilot testing of questions proposed for the tests subsequent to review by the item review committees (CAC and BRC). Based on the pilot test results, items undergo a psychometric review to determine their appropriateness for inclusion on operational test forms. Acceptable item statistics based on pilot testing serve as another source of evidence regarding the importance and relevance of the test content for educator licensure candidates.
- **Setting passing standards.** Another committee of educators is convened to help establish the passing standards for every test. These committees meet to discuss performance standards and, using their professional judgment regarding the content required of the entry-level educator, to provide recommendations of the level of performance (passing score) deemed acceptable for entry-level educators. These judgments are then presented to the Department for consideration in establishing passing scores at a level appropriate to the profession.
- **Communicating appropriate interpretations with test users.** It is important that test scores are understood and used appropriately by the various potential users of the test results. Evaluation Systems includes an explanatory page of text with every examinee score report describing the included information. This information is also posted on the testing program website. Reports to educator preparation institutions include appropriate interpretive cautions. In addition, Evaluation Systems has worked closely with the state to provide guidance regarding the appropriate and psychometrically sound uses of the test scores.

Copyright © 2016 Pearson Education, Inc. or its affiliate(s). All rights reserved.

Pearson and its logo are trademarks, in the U.S. and/or other countries, of Pearson Education, Inc. or its affiliate(s).

TEST RELIABILITY

AERA, APA, and NCME (2014) define test reliability, as a general concept, as "the consistency of scores across replications of a testing procedure" (p. 33). There are a number of statistics that may be used to estimate test reliability. In general, reported reliability values range from 0.00 to 1.00, with higher values indicating greater reliability of test scores. In a licensing context, reliability measures for the MTEL* may be influenced by many factors, such as:

- **Number of examinees.** In general, reliability estimates based on larger numbers of examinees are more stable than estimates based on smaller numbers. For this reason, reliability estimates are calculated for tests that are taken by one hundred or more examinees.
- **Test length.** Reliability estimates tend to be higher for tests with greater numbers of questions.
- **Test content.** Reliability estimates are typically higher for tests that cover narrow, homogeneous content than for tests (such as many used for educator licensure) that cover a broad range of content.
- **Examinees' knowledge.** Reliability estimates tend to be higher if examinees in the group have widely varying levels of knowledge and lower if they tend to have similar levels of knowledge.

Total Test Decision Consistency*. In a licensing context, the most important testing outcome of the MTEL is the pass/fail decision. Total test decision consistency is a reliability statistic that describes the consistency of the pass/fail decision on the total test. A single-test estimate of total test decision consistency (Breyer and Lewis, 1994) is provided for test forms taken by 100 or more examinees. Each test form is carefully divided to create two halves that are parallel in terms of item content and item statistics. Performance on the two test halves is then compared to provide a decision consistency statistic. This statistic is reported in the range of 0.00 to 1.00; the closer the estimate is to 1.00, the more consistent (reliable) the decision is considered to be.

Assessments with component subtests. This program includes assessments that consist of two or more subtests. The subtest model is used for two reasons. First, many educator licenses require candidates to demonstrate proficiency across a variety of domain content. With a single test model in which the total test score is based on performance on all test items, outstanding performance on one component (e.g., reading/language arts) may compensate for poorer performance on another (e.g., mathematics). In a subtest model, candidates must pass each subtest separately, thus providing evidence of acceptable proficiency on each component. Second, the subtest model allows candidates who pass some subtests and fail others to retake only the failed components. Both of these characteristics are considered advantages by many policy agencies. One consequence of the subtest model, however, is that the pass/fail decisions are based on a decreased number of test items, when compared to a total test model in which all test items contribute to the pass/fail decision. As a result, it is not rare that the reliability estimates from licensure testing achieve lower magnitudes than those from large-scale achievement testing, which relies on longer tests.

*Refer to the MTEL Technical Manual for test data.

Copyright © 2016 Pearson Education, Inc. or its affiliate(s). All rights reserved.

Evaluation Systems, Pearson, P.O. Box 226, Amherst, MA 01004

Pearson and its logo are trademarks, in the U.S. and/or other countries, of Pearson Education, Inc. or its affiliate(s).