Measuring the Power of Learning.™

Texas Educator Certification Program

Technical Manual

December 2016

Table of Contents

Preface

Purpose of this Manual

The purpose of the Texas Educator Certification Program Technical Manual is to give details about the following aspects of the Texas Educator Certification Program:

- The purpose of the tests
- How Texas uses the tests
- The approach ETS takes in developing the tests
- The validity evidence supporting score use
- The statistical processes supporting the psychometric quality of the tests
- The score reporting process
- Statistical summaries of test taker performance on all tests

Audience

This manual was written for policy makers and state educators who are interested in:

- Knowing more about the Texas Educator Certification Program
- Understanding how Texas Educator Certification Program tests are developed and scored
- The statistical characteristics of Texas Educator Certification Program tests

Purpose of the Texas Educator Certification Program Tests

Texas Administrative Code §230.5(b) requires every person seeking educator certification in Texas to perform satisfactorily on comprehensive tests. The purpose of these tests is to ensure that each educator has the prerequisite content and professional knowledge necessary for an entry-level position in Texas public schools. These programs were developed for this purpose.

TExES

The Texas Examinations of Educator Standards™ (TExES™) are criterion-referenced examinations designed to measure a candidate's knowledge in relation to an established criterion rather than to the performance of other candidates. The TExES Educator Standards, based on the Texas Essential Knowledge and Skills (TEKS), form the foundation for the TExES tests.

Developing the tests was a collaborative process involving classroom teachers and other educators from public and charter schools, university and Educator Preparation Program (EPP) faculty, representatives from professional educator organizations, content experts and members of the community. Detailed information about the test development process is available on the Texas Education Agency (TEA) website.

TExMaT

The Texas Examinations for Master Teachers™ (TExMaT™) program has its origins in legislation passed in 1999 (House Bill 2307) that required the creation of the Master Reading Teacher (MRT) Certificate, the development of standards for the certificate and the development of a Master Reading Teacher test. The MRT Certificate was implemented as part of the Texas Reading Initiative to ensure that all Texas students are reading at grade level by the end of the third grade and that their reading knowledge and skills grow throughout their public school careers. The MRT test was the first test to be offered in the TExMaT program.

In 2001, the Texas legislature passed legislation that created two additional categories of Master Teacher Certificates: the Master Mathematics Teacher (MMT) Certificates (Early Childhood–Grade 4, Grades 4–8, and Grades 8–12) and the Master Technology Teacher (MTT) Certificate. Tests for these certificates were first administered beginning June 28, 2003.

In 2002, Governor Rick Perry proposed the creation of an additional category of Master Teacher Certificate: the Master Science Teacher Certificate. In 2003, the Texas legislature created Master Science Teacher (MST) Certificates for Early Childhood–Grade 4, Grades 4–8 and Grades 8–12. Tests for these certificates were first administered beginning October 21, 2006.

TASC and TASC-ASL

The Texas Assessment of Sign Communication™ (TASC™) and Texas Assessment of Sign Communication – American Sign Language™ (TASC-ASL™) are extensions of the TExES program for certification in specific areas.

- The TASC is for candidates who plan to teach students who are deaf or hard-of-hearing. The TASC assesses sign communication proficiency within one or more of several sign communication systems used in Texas classrooms.
- The TASC-ASL is for candidates who plan to teach ASL as a Language Other Than English. The TASC-ASL assesses proficiency in American Sign Language (ASL) exclusively.

The tests use an interview format. An experienced interviewer conducts a 20-minute, one-on-one conversational interview with a candidate. The interview is videotaped, and the videotape is viewed by raters who rate the candidate's expressive and receptive sign communication proficiency. Candidates respond to signed questions that allow them to demonstrate their proficiency in signed communication. Each candidate's sign communication proficiency is measured against an established standard of competence. Candidates are not rated based on the content of their responses, but rather on how well they are able to communicate their ideas and understand the interviewer.

Assessment Development and Standard Setting

Fairness in Test Development

As the contractor responsible for the development and delivery of the Texas Educator Certification Program tests, Educational Testing Service (ETS) is committed to assuring that the Texas Educator Certification Program tests are of the highest quality and as free from bias as possible. All products and services developed by ETS — including individual test items, instructional materials and publications — are evaluated during development so that they are not offensive or controversial; do not reinforce stereotypical views of any group; are free of racial, ethnic, gender, socioeconomic or other forms of bias; and are free of content believed to be inappropriate or derogatory toward any group.

For more explicit guidelines used in item development and review, please see the *ETS Standards for Quality and Fairness* (2014).

Test Development Standards

During the test development process, the program follows the strict guidelines detailed in the *Standards for Educational and Psychological Testing* (2014), including, but not limited to:

- Define clearly the purpose of the test, the construct measured, the intended population, and the claims one wants to make about the test takers [Standard 4.1]
- Define the professional knowledge and skills necessary for the construct [Standard 11.2]
- Demonstrate a close link between the professional knowledge and skills and the test content [Standard 11.3] and validate the link using job analyses [Standard 11.7]
- Develop test specifications consistent with the purpose of the test and the domains of professional knowledge and skills defined for the construct, including the item types and numbers of items needed to adequately sample the domains of knowledge [Standard 4.2], and have these specifications reviewed by experts external to the testing program [Standard 4.6]
- Develop and review test items that provide evidence of the indicators detailed in the test specifications, using trained experts to determine relevance of content and correctness of answer(s) [Standards 4.7 & 4.8]
- Develop and revise items to be fair to population subgroups, by maximizing access and reducing construct-irrelevant barriers, including dependence on knowledge, skills or experience that could create subgroup bias [Standards 3.1, 3.2, & 4.8]

Validity

As noted by Kane (2006), validation is a process of evaluating the reasonableness of the intended interpretations and uses of test scores, and validity is the extent to which collected evidence supports or refutes the intended interpretations and uses. For initial certification tests, the main source of validity evidence comes from the alignment between what the profession defines as knowledge and/or skills important for beginning practice (e.g., TExES Educator Standards) and the content included on the test (*Standards for Educational and Psychological Testing*, 2014). The translation of the TExES Educator Standards to content specifications or test frameworks for TExES tests relies on the expert judgment of Texas educators. As test frameworks are developed for new tests or are significantly revised for existing tests, the expert judgments of Texas educators are confirmed with a wider survey of practicing teachers and teacher preparation faculty.

Job Analysis Process

The objective of a job analysis conducted for purposes of initial certification is to identify knowledge and/or skills judged to be important for effective beginning practice. There is no one prescribed way to conduct a job analysis; however, any such process should involve content (job or subject matter) experts (i.e., practitioners) and, as appropriate, others with relevant professional perspectives (e.g., educators who prepare practitioners) who are diverse with respect to characteristics such as practice setting, gender, race or ethnicity and geographic region (Kuehn, Stallings, & Holland, 1990).

The job analysis process carried out for new or significantly revised Texas certification tests includes three main steps:

1. A committee of Texas educators working with educator standards to create a test framework, which describes knowledge and/or skills to be covered by the test.

2. A survey of Texas educators to obtain independent judgments of the importance of the knowledge and/or skills represented on the test framework for beginning practice.

3. An analysis of the survey judgments to confirm the importance of the knowledge and/or skills for beginning practice.

Test Framework. The test framework describes the content that will be covered on the test and is based on the TExES Educator Standards developed for that field and grade span (SBEC/TEA, 2013). Each framework includes the domains to be measured, the competencies that define each domain and the knowledge and skills (descriptive statements) that define each competency. Draft test frameworks are not finalized until after the standards are approved by the SBEC Board.

Framework Confirmation Survey. The test framework is drafted into a web-administered survey to enable wider groups of Texas educators in the relevant certification area to provide judgments on the importance of the statements in the framework for beginning practice. The survey includes the competencies and descriptive statements. Each is accompanied by a 5-point rating scale, ranging from (1) “not at all important” to (5) “extremely important.” Educators judge the importance of each competency and descriptive statement for beginning practice. The educators are also asked to judge the relative importance of each of the framework domains.

Data Analysis. The primary goal of the survey is to differentiate between more important and less important statements. For each competency and descriptive statement, a mean importance rating of 3.50 is used as the criterion for making this distinction. Tannenbaum and Rosenfeld (1994) noted that "this criterion is consistent with a content validation strategy that appropriately reduces the probability of including unimportant . . . skills in the test content domain while not unnecessarily restricting domain comprehensiveness" (p. 204). Mean importance ratings are computed for the total group of educators (respondents). Subgroup analyses are performed if there are at least 30 respondents in a subgroup. The results of the job survey are summarized and are available to the Texas Education Agency.

Test Development Process

The Texas Educator Certification Program tests and related materials follow a rigorous development process, as outlined below and in Figure 1:

1. **Develop Test Frameworks.** Test Specialists work with Test Development Committees, comprised of Texas teachers and teacher educators, to develop test frameworks that are based on the Educator Standards. These frameworks outline the specific competencies to be measured on the new TExES tests.
2. **Conduct Job Analysis/Content Validation Surveys.** A sample of Texas educators are surveyed to confirm the relative job importance of each competency outlined in the test framework. These educators include certified practitioners in the fields related to the certification tests as well as those who prepare the practitioners in those fields.
3. **Develop and Review New Test Questions.** Texas item writers develop draft questions that are designed to measure the competencies described in the test framework. Questions undergo rigorous review by ETS Test Specialists and Texas educators to ensure that they reflect the test framework. The questions are also reviewed for accuracy and appropriateness of content, difficulty, clarity, and potential ethnic, gender, and regional bias. Additionally, constructed-response tasks are also piloted on an appropriate sample of candidates to ensure they will elicit an appropriate range of responses and perform as intended.
4. **Develop and Review Test Forms.** TExES assessments are constructed to reflect the content in the test framework. The completed test forms undergo rigorous review to ensure that they accurately reflect the test framework, that the test questions reflect an appropriate sample of the construct, and that all questions are fair, valid, and accurate. Our review processes are described in more detail in the documents referenced under "Review Processes."
5. **Set Passing Standard.** A committee of Texas educators participates in a rigorous standard-setting study to recommend a passing score for the test. TEA presents the recommendation to the SBEC Board for consideration. The SBEC Board makes the final determination regarding the passing score.

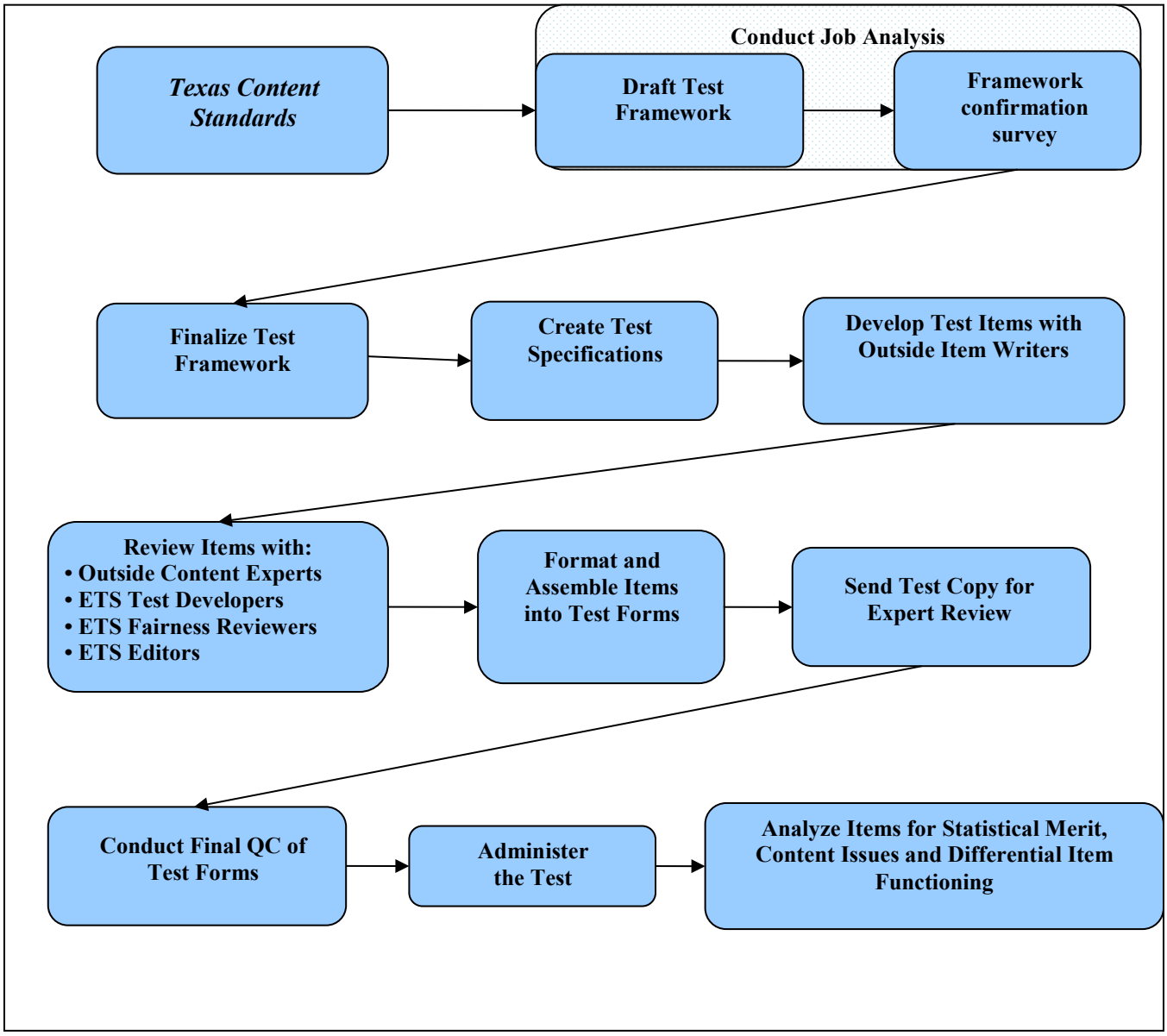

Figure 1. The Test Development Process

Review Processes

ETS follows strict, formal review processes and guidelines. All certification tests and other products developed by ETS undergo multistage, rigorous, formal reviews to verify that they adhere to ETS's fairness guidelines that are set forth in three publications.

ETS Standards for Quality and Fairness

Every test that ETS produces must meet the exacting criteria of the *ETS Standards for Quality and Fairness* (2014), based on the *Standards for Educational and Psychological Testing* (2014). These standards reflect a commitment to producing fair, valid and reliable tests. The criteria are applied to all ETS-administered programs, and compliance with them has the highest priority among the ETS officers, Board of Trustees and staff. Additionally, the ETS Office of Professional Standards Compliance audits each ETS testing program to ensure its adherence to the *ETS Standards for Quality and Fairness* (2014).

The Code of Fair Testing Practices in Education

In addition to complying with the ETS quality standards, ETS develops and administers tests that also comply with *The Code of Fair Testing Practices in Education* (2004).

ETS Fairness Review Guidelines

ETS is committed to creating tests that are as free from bias as possible. All products and services developed by ETS — including individual test items, instructional materials and publications — are evaluated during development to ensure that they are not offensive or controversial; do not reinforce stereotypical views of any group; are free of racial, ethnic, gender, socioeconomic or other forms of bias; and are free of content believed to be inappropriate or derogatory toward any group. For more explicit guidelines used in test development and review, please see the *ETS Guidelines for Fair Tests and Communications* (2015).

Standard Setting

To support the administration and use of a new or revised Texas Educator Certification Program test, ETS designs and conducts a standard-setting workshop. A separate workshop is conducted for each test. For tests consisting of two or more separately scored subtests, the standard-setting workshop results in a recommended passing score for each subtest (i.e., a composite passing score is not provided).

Most Texas Educator Certification Program tests contain selected-response items, including traditional four-option multiple-choice items. Certain tests include other dichotomously-scored item types, such as numeric entry and computer-enabled item types (e.g., drag-and-drop and text highlighting). Some tests include constructed-response items that allow for partial credit but require human scoring. The selection of the item types in a test is driven by the test framework, which in turn is based on the Texas Educator Standards with input from Texas educators. The test framework, along with psychometric considerations (test design), determines the number of items of each type on a test.

Overview

Each standard-setting workshop involves a committee of 12 to 15 Texas educators, selected from a larger pool of educators. The committee is approved by the SBEC. A two- or three-day workshop is conducted in the Austin or San Antonio area. Each workshop uses one or more variations of the Angoff method for determining a recommended passing score based on expert judgments. The Angoff method

for standard setting is the most widely used and researched approach in the area of licensure and certification testing (Cizek & Bunch, 2007).

For dichotomously scored items, the ***Modified Angoff*** method is used to gather item-by-item judgments. Committee members are asked to judge the likelihood that the target candidate would answer a question correctly. For constructed-response items, the ***Extended Angoff*** method is used to gather judgments. Committee members are asked to judge the most likely score that the target candidate would earn on the item. Each method is described in detail below.

The result of a standard-setting workshop is a single recommended passing score. If the test includes both dichotomously-scored and partial-credit items, judgments are combined to obtain the single recommended passing score. If the test includes two or more separately scored subtests, a recommended passing score is determined for each subtest.

The credibility and acceptance of a passing score is a function of the process used to recommend the passing score and the qualifications of the experts selected to recommend the passing score (Geisinger & McCormick, 2010; Hambleton & Pitoniak, 2006). All standard-setting workshops that ETS conducts on behalf of TEA are designed to incorporate core procedures that lead to credible workshop outcomes. Additionally, ETS works with the TEA to ensure that educators who serve on a standard-setting committee meet eligibility criteria (e.g., certification, years of experience) that support their participation as an expert educator qualified to make the necessary judgments.

Description of Expert Committees

Each standard-setting workshop involves a committee of 12 to 15 Texas educators[1]—classroom teachers, support professionals[2] or school leaders[3], as well as higher education faculty or program providers preparing Texas educators. Committee members are approved by TEA.

Data collected during the nomination process include:

- Current position (e.g., teacher, faculty, school leader)
- Area of expertise
- Grade-level assignment (for teachers and school leaders)
- Educator certification
- Years of experience
- Location of school
- Demographic background (e.g., gender, race/ethnicity)

The intent when selecting from the pool of potential committee members to form a standard-setting committee is to create a committee composed of approximately 70% certified EC–12 teachers (or school leaders or support professions, depending on the test) and 30% higher education faculty or program providers who prepare educators. Committee members are selected based on the match between the

[1] In some smaller incidence certification areas, the number of committee members may be less than 12; however, committees should not be less than eight committee members to support the defensibility of the results.
[2] "Support professional" would include media specialists, school counselors or psychologists and librarians.
[3] "School leaders" would include building-level principals and district-level superintendents.

committee members' area of expertise and grade-level assignment and the content of the test. Attempts are made to match the characteristics of the committee to the diversity of the Texas educator community as well as maintain a geographic diversity.

Standard Setting Methodologies and Procedures

The activities described in this section are followed in conducting standard-setting workshops for each Texas Educator Certification Program test. These activities are based on guidelines for professionally accepted and defensible standard-setting practices (AERA, APA & NCME, 2014; Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Tannenbaum & Katz, 2013; Zieky et al., 2008). In particular, these activities are designed based on the Angoff standard-setting method, which is the most frequently used approach for certification systems (Cizek & Bunch, 2006). These activities apply to all committee-based standard-setting methods and, as such, are considered essential to a high-quality standard setting.

Pre-workshop Activity

Prior to the workshop, committee members are sent a homework assignment — to review the test framework and identify the knowledge/skill sets that might differentiate a "just" qualified candidate (the target candidate) from a candidate who is "not quite" qualified. Reviewing the test framework familiarizes committee members with the content covered on the test. Additionally, this activity primes committee members for developing a target candidate description during the workshop (described below).

Reviewing the Test

The first activity for standard-setting committee members is to take and discuss the test to become familiar with its content. The purpose of the discussion is to bring the committee members to a shared understanding of what the test does and does not cover, which serves to reduce potential judgment errors later in the standard-setting process.

In taking the test, the committee members respond to the dichotomously-scored items and sketch responses to the constructed-response items. Afterward, the committee members check their responses against the answer key for the dichotomously-scored items and the scoring rubrics for the constructed-response items.

The committee members then engage in a discussion of the major content areas being addressed by the test. They are also asked to remark on any content areas they think would be particularly challenging for entry-level teachers (or school leaders or support professionals) and areas that address content that would be particularly important for entry-level teachers (or school leaders or support professionals).

Describing the Target Candidate

Following the review of the test, committee members describe the target candidate. The target candidate description plays a central role in standard setting (Perie, 2008) as it is the goal of the standard-setting process to identify the test score that aligns with this description.

The committee members use their notes from the pre-workshop assignment to help describe the target candidate. They focus on the knowledge/skills that differentiate a "just" from a "not quite" qualified candidate. The committee first breaks into small groups and develops descriptions of the

target candidate. The committee then reconvenes and, through whole-group discussion and consensus, finalizes the shared description.

The written description summarizes the committee discussion in a bulleted format. The description is not intended to describe all the knowledge and skills of the target candidate but only highlight those that differentiate a "just" qualified candidate from a "not quite" qualified candidate. The written description is distributed to committee members to use during later phases of the workshop.

Training Committee Members to Make Standard-setting Judgments

Committee members are provided ample training in how to complete the standard-setting judgments for the particular variation(s) of the Angoff method applied to the test being considered.

- **Dichotomously-scored items.** A probability-based ***Modified Angoff*** method (Brandon, 2004; Hambleton & Pitoniak, 2006) is used for dichotomously-scored items (selected-response items, numeric entry). In this approach, a committee member judges, for each item, the likelihood (probability or chance) that the target candidate would answer the item correctly. Committee members make their judgments using the following rating scale: 0, .05, .10, .20, .30, .40, .50, .60, .70, .80, .90, .95, 1. The lower the value, the less likely it is that the target candidate would answer the item correctly, because the item is difficult for the target candidate. The higher the value, the more likely it is that the target candidate would answer the item correctly.
- **Constructed-response (partial-credit) items.** An ***Extended Angoff*** method (Cizek & Bunch, 2007; Hambleton & Plake, 1995) is used for constructed-response (partial-credit) items. In this approach, a committee member decides on the assigned score value that would most likely be earned by the target candidate for each item. Committee members are asked first to review the description of the target candidate and then to review the item and its rubric. The rubric for an item defines (holistically) the quality of the evidence that would merit a response earning a particular score. During this review, each committee member independently considers the level of knowledge/skill required to respond to the item and the features of a response that would earn a particular score, as defined by the rubric. Each committee member decides on the score most likely to be earned by the target candidate from the possible values a test taker can earn.

After the training, committee members are given the opportunity to make practice judgments and to discuss those judgments and their rationales. This activity often reveals misunderstandings about the judgment process that are remediated through additional training. All committee members complete a post-training survey to confirm that they have received adequate training and feel prepared to continue; the standard-setting process continues only if all committee members confirm their readiness.

Standard-setting Judgments

Following training on the particular standard-setting method(s), committee members independently make their judgments for each item on the test (Round 1). Judgments are recorded on scannable forms. Committee members refer to the target candidate description as they review each item and make their judgments. For constructed-response items, committee members also refer to the item's scoring rubric.

Following the completion of the Round 1 judgments, forms are collected and scanned on-site. Scanned data undergo a quality review and judgments are summarized.

Adjusting Standard-setting Judgments

Following Round 1, item-level feedback is provided to the committee. The committee members' judgments are displayed for each item and summarized across committee members. Items are highlighted to show when committee members converge in their judgments (at least two-thirds of the committee members located an item in the same difficulty range) or diverge in their judgments.

Led by an ETS standard-setting facilitator, the committee members discuss their item-level judgments. These discussions help committee members maintain a shared understanding of the knowledge/skills of the target candidate and help to clarify aspects of items that might not have been clear to all committee members during the Round 1 judgments. The purpose of the discussion is not to encourage committee members to conform to another's judgment but to understand the different relevant perspectives among the committee members.

As committee members are discussing their Round 1 judgments, they are encouraged to consider their judgments in light of the rationales provided by other committee members. If a committee member wants to change his/her Round 1 judgment, she/he can record the adjusted judgment on the scannable form. Round 2 judgments are only recorded for items when a committee member wishes to change a Round 1 judgment. Otherwise, no action is required.

Following Round 2, the committee members have another opportunity to review and revise their judgments. Round 3 follows in the same process as the second round. The committee members' judgments are displayed for each item and summarized across committee members. They are encouraged to discuss their rationales. If a committee member wants to change his/her Round 2 judgment, she/he can record the adjusted judgment on the scannable form. Round 3 judgments are only recorded for items when a committee member wishes to change a Round 2 judgment. Otherwise, no action is required.

Following completion of the Round 3 judgments, forms are collected and scanned on-site. Scanned data undergo a quality review and judgments are analyzed. Each committee member's passing score is calculated and the committee member-level results are summarized across committee members to determine the committee's overall recommended passing score.

Complete Evaluation Survey

The committee members complete an evaluation survey at the conclusion of their standard-setting workshop. The evaluation survey asks the committee members to provide feedback about the quality of the standard-setting implementation and the factors that influenced their decisions. The responses to the surveys provide evidence of the validity of the standard-setting process, and, as a result, evidence of the reasonableness of the recommended passing scores.

Committee members are also shown the committee's recommended passing score. They are asked (a) how comfortable they are with the recommended passing score, and (b) if they think the score was too high, too low or about right.

Psychometric Procedures

Introduction

ETS' Statistical Analysis Center has developed procedures designed to support the development of valid and reliable test scores for the Texas Educator Certification Program. The item and test statistics are produced by software developed at ETS to provide rigorously tested routines for both classical and Item Response Theory (IRT) analyses.

The psychometric procedures explained in this section follow well-established, relevant standards in *Standards for Educational and Psychological Testing* (2014) and the *ETS Standards for Quality and Fairness* (2014). They are used extensively in the Texas Educator Certification Program and are accepted by the psychometric community at large.

As discussed in the Assessment Development section, every test in the Texas Educator Certification Program has a set of test specifications that is used to create parallel versions of each test, called test forms. Each test form has a unique combination of individual test items. The data for the psychometric procedures described below are the test taker item responses collected when the test form is administered for the first time. For small volume tests, test taker responses are accumulated for a specified period of time in order to produce reliable results.

Test Scoring Process

As of the 2015-16 testing year, all Texas Educator Certification Program tests are administered regularly via computer at computer-based test centers with the exception of the TExES Braille test; the TExMaT MRT, MMT and MST tests; and the TASC and TASC-ASL tests. The following is an overview of the test-scoring process:

- When a new form is introduced, a Preliminary Item Analysis (PIA) of the test items is completed within one week following the administration. Items are evaluated statistically to confirm that they perform as intended in measuring the desired knowledge and skills for beginning teachers.
- A Differential Item Functioning (DIF) Analysis is conducted to determine that the test questions meet ETS's standards for fairness. DIF analyses compare the performance of subgroups of test takers on each item. For example, the responses of male and female or Hispanic and White subgroups might be compared.
- Items that show very high DIF statistics are reviewed by a fairness panel of content experts, which often includes representatives of the subgroups used in the analysis. The fairness panel decides if a test takers' performance on any item is influenced by factors not related to the construct being measured by the test. Such items are then excluded from the test scoring. A more detailed account of the DIF procedures followed by the Texas Educator Certification Program are provided in "Differential Item Functioning (DIF) Analyses" in a later section, and are described at length in Holland and Wainer's (1993) manuscript.
- Based on PIA and DIF results, test developers consult with content experts or content advisory committees to determine whether any flagged item on a new test form meets ETS's standards for quality and fairness. Their consultations are completed within days after the administration of the test.

- Statistical equating and scaling is performed on each new test after PIA and DIF results are reviewed and approved for scoring.
- For tests with constructed-response (CR) items or a combination of MC and CR items, scores are reported approximately three weeks after the paper-based testing window and four weeks after the computer-based testing window. For MC tests, scores are typically reported within seven days of the computer-based test administration.
- A Final Item Analysis (FIA) report is completed six to eight weeks after the test administration. The final item-level statistical data are provided to test developers to assist them in the construction of future forms of the test.

Item Analyses

Classical Item Analysis

The term Classical Item Analysis is used by the testing industry to distinguish from other approaches to obtaining item statistics that are usually model based such as Item Response Theory. Classical item analysis uses observed item responses that test takers provide during test administration. Following the administration of a new test form, but before scores are reported, a PIA for all items is carried out to provide information to assist content experts and test developers in their review of the items. They inspect each item, using the item statistics to detect possible ambiguities in the way the items were written, keying errors or other flaws. Items that do not meet ETS's quality standards can be excluded from scoring before the test scores are reported. Because the Texas tests also use some embedded pretest items, operational items that performed poorly (as indicated by the item statistics), are replaced by pretest items that performed well. For example, if a certain item which was initially planned to be an operational item performed poorly, then a pretest item from the same domain (which performed well) is used to replace the poorly performing operational item.

In ETS' item banking system, the information from PIA is typically replaced by FIA statistics if a sufficient number of test takers have completed the test to permit accurate estimates of item characteristics. These final statistics are used for assembling new forms of the test. However, some Texas Educator Certification Program tests are taken only by a small number of test takers. For these tests, FIAs are calculated using data accumulated over several test administrations as long as a test form remains active.

Preliminary and final analyses include both graphical and numerical information to provide a comprehensive visual impression of how an item is performing. These data are subsequently sent to the test developers, who retain them for future reference. An example of an item analysis graph of an MC item is presented in Figure 2.

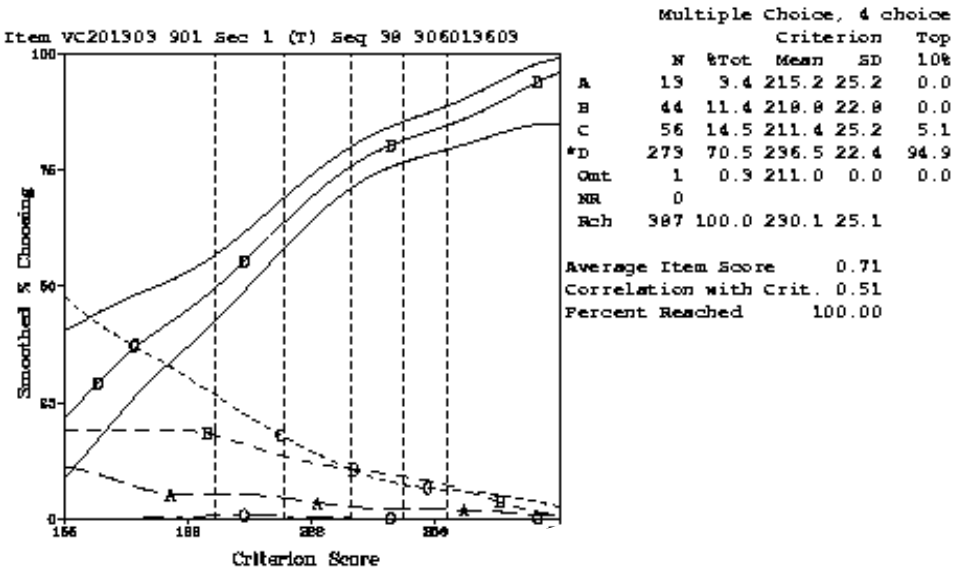

Figure 2. An example of an item analysis graph for an MC item

In this example of an MC item with four options, the percentage of test takers choosing each response choice (A–D) and omitting the item (Omt) is plotted against their performance on the criterion score of the test. In this case the criterion is the total number of correct responses. Vertical dashed lines are included to identify the 20th, 40th, 60th, and 80th percentiles of the total score distribution, and 90-percent confidence bands are plotted around the smoothed plot of the correct response (D). The small table to the right of the plot presents summary statistics for the item:

- For each response option, the table shows the count and percent of test takers who chose the option, the criterion score mean and standard deviation of respondents and the percent of respondents with scores in the top ten percent of test takers who chose the option.
- Three statistics are presented for the item as a whole: 1) The Average Item Score (the percent of correct responses to an item that has no penalty for guessing); 2) The correlation of the item score with the criterion score. For an MC item this is a biserial correlation, a measure of the relationship between a dichotomized normally distributed continuous variable assumed to underlie the dichotomous item's outcomes, and the criterion score. For CR items this is a polyserial correlation because the index summarizes a series of correlations between response categories and the criterion variable; 3) the percent of test takers who reached the test item.

For CR items, both item and rater analyses are conducted. The item analyses include distributions of scores on the item; two-way tables of rater scores before adjudication of differences between raters; the percentage of exact and adjacent agreement; the distributions of the adjudicated scores; and the correlation between the scores awarded by each of the two raters. For each rater, his/her scores on each item are compared to those of all other raters for the same set of responses.

Within one week of a new form's administration, statistical analysts deliver a PIA to test developers for each new test form. Items are flagged for reasons including but not limited to:

- Low average item scores (very difficult items)
- Low correlations with the criterion
- Possible double keys
- Possible incorrect keys

Test developers consult with content experts or content advisory committees to determine whether each MC item flagged at PIA has a single best answer and should be used in computing test taker scores. Items found to be problematic are identified by a Problem Item Notification (PIN) document. A record of the final decision on each PINned item is approved by the test developers, the statistical coordinator and a member of the Texas Educator Certification Program management staff. This process verifies that flawed items are identified and removed from scoring, as necessary.

Speededness

Occasionally, a test taker may not attempt items near the end of a test because the time limit expires before she/he can reach the final items. The extent to which this occurs on a test is called "speededness." The Texas Educator Certification Program assesses speededness using four different indices:

- The percent of test takers who complete all items
- The percent of test takers who complete 75 percent of the items
- The number of items reached by 80 percent of test takers[4]
- The variance index of speededness (i.e., the ratio of not-reached variance to total score variance).[5]

All four of these indices need not be met for a test to be considered speeded. If the statistics show that many test takers did not reach several of the items, this information can be interpreted as strong evidence that the test (or a section of a test) was speeded. However, even if all or nearly all of the test takers reached all or nearly all of the items, it would be wrong to conclude, without additional information, that the test (or section) was not speeded. Some test takers might well have answered more of the items correctly if given more time. Item statistics, such as the percent correct and the item total correlation, may help to determine whether many test takers are guessing, but the statistics could indicate that the items at the end of the test are difficult. A Texas Educator Certification Program test will be considered speeded if more than one of the speededness indices is exceeded.

[4] When a test taker has left a string of unanswered items at the end of a test, it is presumed that he/she did not have time to attempt them. These items are considered "not reached" for statistical purposes.

[5] An index less than 0.15 is considered an indication that the test is not speeded, while ratios above 0.25 show that a test is speeded. The variance index is defined as S_{NR}^2 / S_R^2 where S_{NR}^2 is the variance of the number of items not reached, and S_R^2 is the variance of the total raw scores.

Differential Item Functioning (DIF) Analyses

DIF analyses are conducted during the week after each Texas Educator Certification Program test administration, sample sizes permitting, to inform fairness reviews. DIF analysis utilizes a methodology pioneered in medical research (Mantel & Haenszel, 1959) and modified and augmented for educational applications by ETS (Dorans & Kulick, 1986; Holland & Thayer, 1988; Zwick, Donoghue, & Grima, 1993). It involves a statistical analysis of test items for evidence of differential item difficulty related to subgroup membership. The assumption underlying the DIF analysis is that groups of test takers (e.g., male/female; Hispanic/White) who score similarly overall on the test or on one of its subsections — and so are believed to have comparable overall content understanding or ability — should score similarly on individual test items.

For example, DIF analysis can be used to measure the fairness of test items at a test taker subgroup level. ETS psychometricians use well-documented DIF procedures, in which two groups are matched on a criterion (usually total test score) and then compared to see if the item is performing similarly for both groups. For tests that assess several different content areas, the more homogeneous content areas (e.g., verbal or math content) are preferred to the raw total score as the matching criterion. The DIF statistic is expressed on a scale in which negative values indicate that the item is more difficult for members of the focal group (generally African American, Asian American, Hispanic American, Native American, or female test takers) than for matched members of the reference group (generally White or male test takers). Positive values of the DIF statistic indicate that the item is more difficult for members of the reference group than for matched members of the focal group. If sample sizes are too small to permit DIF analysis before test-score equating, they are accumulated until there is sufficient volume to do so as long as a test form remains active.

DIF analyses produce statistics describing the amount of differential item functioning for each test item as well as the statistical significance of the DIF effect. ETS's decision rules use both the degree and significance of the DIF to classify items into one of three categories: A (least), B and C (most). Any items classified into category C are reviewed at a special meeting that includes staff who did not participate in the creation of the tests in question. In addition to test developers, these meetings may include at least one participant not employed by ETS and a member representing one of the ethnic minorities of the focal groups in the DIF analysis. The committee members determine if performance differences on each C item can be accounted for by item characteristics unrelated to the construct that is intended to be measured by the test. If factors unrelated to the knowledge assessed by the test are found to influence performance on an item, it is removed from the test scoring. Moreover, items with a C DIF value are not selected for subsequent test forms unless there are exceptional circumstances (e.g., the content is required to meet test specifications).

DIF Statistics

DIF analyses are based on the Mantel-Haenszel DIF index expressed on the ETS item delta scale (MH D DIF). The MH D DIF index identifies items that are differentially more difficult for one subgroup than for another, when two mutually exclusive subgroups are matched on ability (Holland & Thayer, 1985).[6] The matching process is performed twice: 1) using all items in the test, and then 2) after items classified as C DIF have been excluded from the total score computation. For most tests, comparable (matched) test takers are defined as having the same total raw score, where the total raw score has been refined to exclude items with high DIF (C items). Typical groups for whom DIF is conducted are Male/Female, White (non-Hispanic)/African American or Black (non-Hispanic), White (non-Hispanic)/Hispanic. The subgroup listed first is the reference group and the subgroup listed second is the focal group.

High positive DIF values indicate that the gender or ethnic focal group performed better than the reference group. High negative DIF values show that the gender or ethnic reference group performed better than the focal group when ability levels were controlled statistically.

Thus, an MH D DIF value of zero indicates that reference and focal groups, matched on total score, performed exactly the same. An MH D DIF value of +1.00 would indicate that the focal group (compared to the matched reference group) found the item to be one delta point easier. An MH D DIF of −1.00 indicates that the focal group (compared to the matched reference group) found the item to be one delta point more difficult.

Based on the results of the DIF analysis, each item is categorized into one of three classification levels (Dorans and Holland 1993), where statistical significance is determined using $p<.05$:

A = low DIF; absolute value of MH D DIF less than 1 or not significantly different from 0,
B = moderate DIF; MH D DIF significantly different from 0, absolute value at least 1, and either
(1) absolute value less than 1.5, or
(2) not significantly greater than 1,
C = high DIF; absolute value of MH D DIF at least 1.5 and significantly greater than 1.

C-level items are referred to fairness committees for further evaluation and possible revision or removal from the test. Test developers assembling a new test form are precluded from selecting C-level items unless absolutely necessary in rare cases for content coverage.

[6] *Delta* (Δ) is an index of item difficulty related to the proportion of test takers answering the item correctly (i.e., the ratio of the number of people who correctly answered the item to the total number who reached the item). Delta is defined as 13 - 4z, where z is the standard normal deviate for the area under the normal curve that corresponds to the proportion correct. Values of delta range from about 6 for very easy items to about 20 for very difficult items.

Test-Form Equating

Overview

Each Texas Educator Certification Program test comprises multiple test forms, with each containing a unique set of test questions, whether multiple choice, constructed response, or a combination of both. *ETS Standards for Quality and Fairness* (2014) require the use of equating methodologies when "results … on different forms of an assessment are to be treated as though they were equivalent" (page 45), as is the case for all Texas Educator Certification Program tests. Equating adjusts scores on different test forms to account for the inherent variation in difficulty among test forms, despite the rigor of the test-assembly processes. Because equating adjusts for differences in difficulty across different Texas Educator Certification Program test forms, a given scale score represents the same level of achievement for all forms of the test. Well-designed equating procedures maintain the comparability of scores for a test and thus avoid penalizing test takers who happen to encounter a selection of questions that proves to be more difficult than expected (von Davier, Holland, & Thayer, 2004; Kolen & Brennan, 2004).

Scaling

To avoid confusion between the adjusted (equated) and unadjusted (raw) scores, the Texas Educator Certification Program has typically reported the adjusted scores on a score scale that makes them clearly different from the unadjusted scores. This reported score scale is a mathematical conversion (or scaling) of the raw scores into scaled scores with predetermined cut scores and lower and upper limits. All Texas Educator Certification Program tests use a scaled score range of 100 to 300 for score reporting with the score of 240 designated as the passing score. The use of a scale common to all forms of the same test title enables test users to compare scores on test forms that may differ slightly in difficulty. In addition, the number of questions and the number of questions answered correctly are reported for each content domain to provide test takers with feedback about their areas of strength and potential improvement. Unlike scaled total scores, domain scores are not adjusted for differences in difficulty and should not be compared across different forms of the same test title.

When the first form of a Texas Educator Certification Program test consisting only of MC items is administered for the first time, a scaling relationship is established between the raw score scale and the reporting scale.

The general scaling formulas used to derive the scaling conversion are:

1. For raw scores greater than or equal to the minimum passing (cut) score obtained by a standard-setting study

   Scaled score = 240 + [60*(raw score-raw cut score) / (max raw score-cut score)]

2. For raw scores less than the minimum passing score

   Scaled score = 100 + 140*(raw score) / (raw cut score)

Equating

To maintain the comparability of the reported scores for each test, for each new form of a test, following the initial scaling of the first test form, each subsequent new form of a test, after its initial administration and before scores are reported, is equated to translate raw scores on the new form to adjusted scores on the test's reporting scale. The equating procedures take into account the difficulty of the form and the relative ability of the group of test takers who took that form.

For Texas Educator Certification Program tests, the most frequently employed equating model is the Non-Equivalent groups' Anchor Test (NEAT) design, where a set of items common to the reference form is included in the new form of the test. This design permits the isolation of differences in ability from differences in form difficulty and is widely used by practitioners because of its applicability to a variety of test settings. Often when the sample sizes are small, it may be necessary to scale the first form of a new test and then reuse it at additional administrations until accumulated volume increases sufficiently to perform equating of a new form using empirical data.

The NEAT Design

Under the NEAT or anchor test design, one set of items (e.g., Test X) is administered to one group of test takers, another set of items (e.g., Test Y) is administered to a second group of test takers, and a third set of common items (e.g., Test V) is administered to both groups (Kolen & Brennan, 2004). The common items that comprise the anchor test are chosen to be representative of the items in the total tests (Test X and Test Y) in terms of both their content and statistical properties. Anchor tests can be either internal (i.e., the common items contribute to the reported scores on the test form being equated) or external (i.e., the common items are not part of the test form being equated). Both linear (e.g., Tucker and Levine) and nonlinear (e.g., equipercentile, frequency estimation) equating methods may be used under the NEAT design. The final raw-score-to-scaled-score conversion line can be chosen based on the relationship between the reference form scores and the new form scores, including characteristics of the anchor and total test score distributions, the reliability of the tests and the sizes of the samples used in the analysis.

The NEAT design can be used for tests composed of MC items only, CR items only or a combination of MC and CR items:

1. Tests containing MC items only are equated using an internal anchor test. In these cases, the anchor test includes approximately 25 percent of the items in the total test.

2. Tests containing sufficient numbers of both MC and CR items are equated using a combination of MC and CR items as an internal anchor test.

3. Tests containing MC items and a small number of CR items are equated using only the MC items in an internal anchor test.

The Equivalent Groups Design

For tests that have a large number of test takers per administration, an equivalent group's equating design may be employed. Two different forms are administered at the same administration: an old test form with an established raw-to-scaled score conversion and a new test form. The two forms are spiraled; that is, the bundles of booklets sent to testing centers are assembled so that the two forms alternate. Because a large number of test takers are in effect randomly assigned to take one or the other of the spiraled test forms involved, it may be assumed that the average test taker's ability in each group is equivalent and that any differences observed can be attributed to differences in form difficulty. Both linear and nonlinear (e.g., direct equipercentile) equating methods may be used with this design.

The Single Group Design

In certain circumstances, such as when an item with significant DIF is found, a new raw-to-scaled score conversion is required to score the form without the flawed item. In these cases, a single group of test takers that has completed all the items is selected for analysis. Two sets of test statistics are calculated: one includes all items and the other where a good pretest item(s) replaces the flawed item(s). The raw means and standard deviations of the two are set equal, establishing an estimate of the original test score for each possible raw score on the new version of the test. The original raw-to-scaled score conversion is then applied to the estimates, yielding a new conversion for the new version of test form.

The SiGNET Design

When a new test form is introduced and the number of test takers is too low to permit an accurate estimation of item characteristics, the Texas Educator Certification Program uses the Single Group Nearly Equivalent Test design. This test design allows items in certain portions of the test to be pretested to determine their quality before they are used operationally.

The basis of this equating design for small samples is that a group of test takers takes two largely overlapping forms of a test as a single combined form (during a single administration). The scored items for the operational test form are divided into several testlets of an equal number of items (an example is presented in Figure 3). Each testlet matches the total test specifications as closely as possible (i.e., like a mini test). An additional testlet is created, also matching the specifications as closely as possible. This additional testlet is not scored and is only used for trying out the items before being used as operational items in a future new form. If the scored testlets are designated as Testlets 1–6 and the additional unscored testlet is designated as Pretest Testlet 7, then Form 1 is composed of Testlets 1–6 and Form 2 is composed of Testlets 2–7. The seven testlets are administered as a single test form in the first administration. When a sufficient number of test takers have taken the administered form for a single group equating, the second form (Testlets 2–7) is equated to the first form (Testlets 1–6). New pretest items are added to the newly developed form. A strong feature of the SiGNET equating design is that it allows new forms to be equated using a Single Group (SG) equating method, which has been shown to have much less random equating error than other equating designs such as the nonequivalent anchor test (NEAT) design.

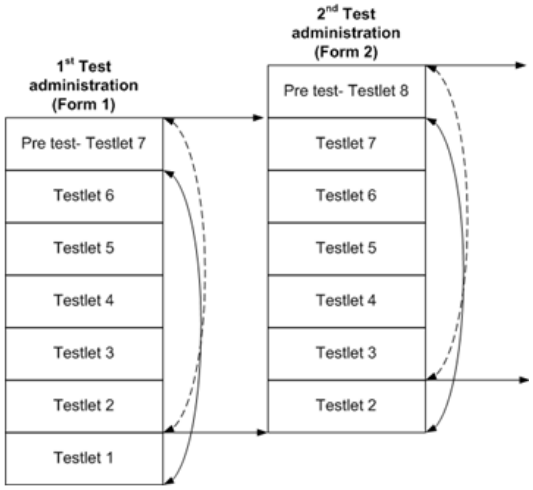

Figure 3. A graphical illustration of the SiGNET equating design

The ISD and Modified SiGNET Designs

For computer-based tests without constructed-response items, where scores are reported within seven days after the test administration, a previously used form that does not require equating is administered, or when appropriate, pre-equated test forms are administered to eliminate the need of equating following an administration.

Two solutions may be used for introducing newly developed test forms in a computer administered format: Interchangeable Section Design (ISD) and modified SiGNET design.

1. The first solution is a newly conceptualized ISD with item response theory (IRT) pre-equating proposed for tests with moderate to high volumes to move to continuous testing. With this design, tests are separated into sections, called testlets, either according to content domains with each testlet containing one or more content categories — or as mini-tests, with each testlet mimicking the full test. Multiple versions of each testlet are created, which are considered interchangeable, with the same content specification and statistical characteristics. By randomly selecting a version for each testlet and combining the testlets into forms during computer delivery, an exponential number of form combinations are generated to reduce security concerns and to accumulate data for IRT calibration. See Figure 4 for an illustration. With the use of interchangeable testlets to accumulate data and to reduce the need for a substantial item pool, this solution addresses challenges regarding tests with smaller candidate volumes. Implementation of this solution requires accumulation of data on existing forms to establish an adequate item pool to assemble the necessary testlets and to use the IRT model for pre-equating.

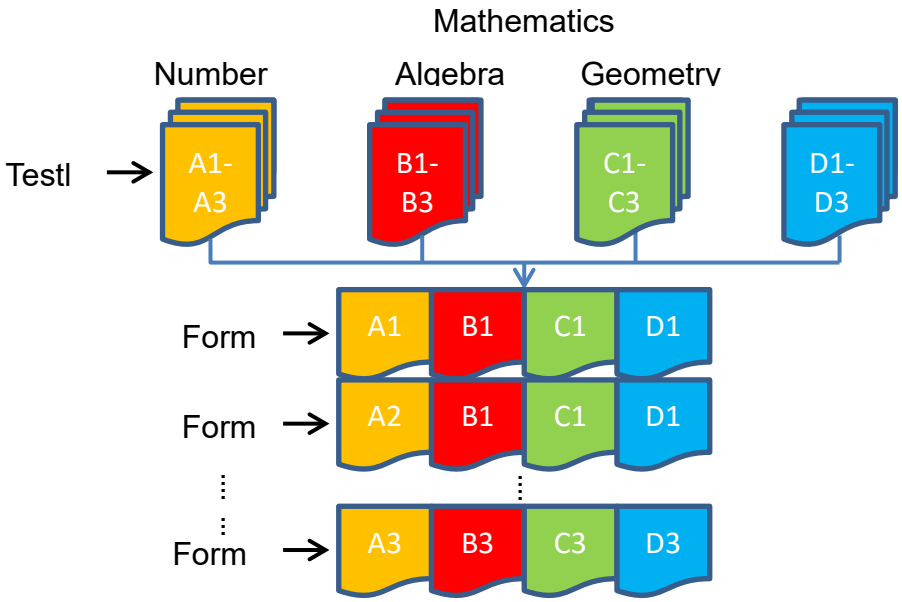

Figure 4. An example of ISD Solution

2. The second solution is a generalization of the SiGNET design that was described earlier. Several test forms are assembled using SiGNET design to make them available for continuous administrations. This design does not require random assembly of test forms at the time of administration, because each intact form have already been assembled and equated prior to the administration so that we can report scores within the seven-day scoring window. This solution requires greater overlap between test forms, but it is only used for lower volume test titles. New forms may contain approximately 80 percent of overlap with existing forms. This option allows for score reporting within seven days, but increases item exposure.

Test Statistics

Reliability

The reliability of a test refers to the degree of consistency or stability of test scores. An index of reliability enables testing practitioners to generalize beyond the specific collection of items in a particular form of a test to a larger universe consisting of all possible items that could be posed to the test taker. Because tests consist of only a sample of all possible items, any estimate of a test taker's actual capabilities will contain some amount of error. Psychometrically, reliability may be defined as the proportion of the test score variance that is due to the "true" (i.e., stable or non-random) abilities of the test takers. A person's actual (or "observed") test score may thus be thought of as having a "true" component and an "error" component. Here, "error" is defined as the difference between the observed and true scores. Because true scores can never be known, the reliability of a set of test scores cannot be assessed directly, but only estimated.

Reliability estimates for MC total, category and equating scores are computed using the Kuder and Richardson (1937) formula 20 (KR 20). Reliability may be thought of as the proportion of test score variance that is due to true differences among the test takers with respect to the ability being measured:

$$reliability = 1 - \frac{error\ variance}{total\ variance},$$

If the test is not highly speeded, Cronbach's alpha estimate of internal consistency reliability will be an adequate estimate of alternate-form reliability (Cronbach, 1951). However, because Texas Educator Certification Program tests are used to make pass/fail decisions, information about the reliability of classification is also relevant in estimating test reliability. Statistical methodology used to estimate reliability of classification is described in more detail later in this section.

Standard Error of Measurement

The standard error of measurement (SEM) is an estimate of the standard deviation of the distribution of observed scores around a theoretical true score. The SEM can be interpreted as an index of expected variation if the same test taker could be tested repeatedly on different forms of the same test without benefiting from practice or being hampered by fatigue. The SEM of a raw score is computed from the reliability estimate (rx) and the standard deviation (SDx) of the scores by the formula:

$$SEM_x = SD_x \sqrt{1 - r_x}\ .$$

The Conditional Standard Error of Measurement (CSEM) is specific to each score level and reflects the errors of measurement associated with low-scoring test takers or high-scoring test takers. CSEMs for Texas Educator Certification Program tests are computed using Lord's (1984) Method IV, and are included in the Texas Educator Certification Program Test Analysis Reports.

Reliability of Classification

Because Texas Educator Certification Program tests are intended for certification, assessing the consistency and accuracy of pass/fail decisions is very important. The Livingston and Lewis method (1995) is used to estimate decision accuracy and consistency at each cut-score level. Classification accuracy is the extent to which the decisions made on the basis of a test would agree with the decisions made from all possible forms of the test (i.e., an estimate of the test taker true score). Classification consistency is the extent to which decisions made on the basis of one form of a test would agree with the decisions made on the basis of a parallel, alternate form of the test.

The estimated percentages of test takers correctly (classification accuracy) and consistently classified (classification consistency) tend to increase in value as the score reliability increases and as the absolute value of the standardized difference (SSD) between the mean total score and the qualifying score increases. When the mean score of test takers is well above or below the qualifying score, the number of test takers scoring at or near the qualifying score is relatively small. Therefore, with fewer test takers in the region of the qualifying score, the number of test takers that could easily be misclassified decreases and the decision reliability statistics reflect that fact by increasing in value.

Reliability of Scoring Constructed-Response Items

The reliability of the scoring process for Texas Educator Certification Program constructed-response (CR) items is determined by a multi-step process.

- The inter-rater correlations for each item are obtained from the two independent ratings, and the inter-rater reliabilities are computed from them using the Spearman-Brown formula (Haertel, 2006).
- Variance errors of scoring for each item are calculated by multiplying the item's variance by $(1 - r_{cis})$, where r_{cis} is the item's inter-rater reliability.
- The variance errors of scoring for all of the items are added together to form the variance of errors of scoring for the entire test.
- The standard error of scoring is defined as the square root of the variance errors of scoring for the sum obtained in step 3.

Please note that the standard errors of scoring for MC tests are zero, as the recording of item responses for these tests is performed mechanically, not by human judgment.

Scoring Methodology

Scoring

For tests consisting only of MC items, a raw score is the number of correct answers on the test. There is no penalty imposed for guessing incorrect responses to MC items.

For tests that include CR items, raw item scores are weighted composites of scores on the individual CR items. For each item, the responses are reviewed and scored by qualified raters according to pre-specified scoring rules (rubric)[7] detailed in the Preparation Manuals, at www.texes.ets.org/prepmaterials. The ratings are based on a rubric developed by educators who are specialists in the subject area. The score on any single CR item is the sum of the scores assigned by two independent raters or is the score determined through consensus by a team of three or more raters and scoring leaders.

For tests that include both MC and CR questions, raw scores are a weighted composite of the raw MC score and the scores on the individual CR items. A test taker's score in the MC portion of the test is the sum of the number of items answered correctly.

Scoring Methodology for Constructed-Response Items

A CR item is one for which the test taker must produce a written, typed, brailled, spoken or signed response. Such items are designed to probe a test taker's depth of understanding of a content area and/or communicative abilities that cannot be assessed solely through MC items. The time suggested for a response can vary from 20 seconds to 60 minutes. Scoring can be:

- Analytic by focusing on specific traits or features
- Holistic by focusing on the response as a whole
- Focused holistic by blending analytic and holistic

Test developers in conjunction with stakeholder committees are responsible for the creation of scoring guides. Test developers are responsible for the selection of samples for training purposes, and the training of scoring leadership in test content and scoring standards and procedures.

Every test that contains CR items has a General Scoring Guide (GSG), which is written to verify that well-trained, calibrated raters will be able to consistently evaluate responses according to clearly specified indicators. Question-Specific Scoring Guides (QSSG) and Scoring Notes also are developed to inform raters of some of the item-specific features that a response might contain. Final ratings are assigned to a response after a careful reading to find the evidence that the item has been answered. That evidence then is evaluated by selecting the set of descriptors in the scoring guide that best fits the evidence. This rating can be on various scales, such as 0–3 or 0–6, depending on how much evidence an item is designed to elicit from test takers.

[7] For many tests, if there is a discrepancy of more than one point between the scores assigned by the two raters, a third person scores the response. For some tests, "back readings," or third readings, are carried out on a subsample of a certain percentage of papers.

Scoring guides for new items are developed as the prompt is developed and are finalized at the "sample pulling" before the first scoring of a prompt. Sample pulling is the process during which the chief reader and/or content scoring leader and/or scoring leaders and/or test developers for a given test:

- Read through the test takers' responses
- Find responses at each score point on the score scale for the test
- Agree on how to score the selected papers
- Document the rationales for the agreed-upon scores
- Arrange the selected papers into training and calibrating sets for each question on a test

After a scoring guide is finalized during its first use, it can be changed only under very narrowly defined conditions and with approval from the statistical coordinator of the test.
The goals of scoring a response according to a GSG for a test as well as a QSSG are that:

- A candidate receives a fair and appropriate score.
- All candidates are rated in the same manner using the same criteria.
- Scoring is conducted consistently throughout a scoring session and from one scoring session to another.

To verify the standardization of the scoring process, the following materials must be developed for every CR item:

- Benchmark papers: exemplars of responses to each score point on the score scale, usually located at the mid-range of a score point
- Training papers: responses used to train raters in the variety of responses that can be expected across the range of each of the points of the scoring guide, often presenting unique scoring issues
- Annotations for the responses (rationales): supplemental information used to explain why sample papers received the given score, providing consistency in what is said during training
- Prompt notes: supplemental information used to explain specific prompts. Raters access these notes during scoring to help them understand what to expect from responses for that specific prompt. This is especially important when responses are scored based on language skills and content information.
- Training manuals: an outline of the process that a scoring leader should follow in training raters

Scoring leaders are responsible for direct training of raters as well as overseeing the quality of scoring. Their responsibilities include:

- Assisting in selecting training materials
- Conducting scorer training and, if necessary, retraining
- Monitoring scoring through backreading and rater feedback
- Verifying that all scoring procedures are followed
- Recommending raters for scoring leadership

Raters are responsible for reading at a sustained rate and giving appropriate scores based on established criteria. They are practicing educators and higher education faculty who are familiar and knowledgeable with the test content.

Consistency in the scoring of a form is verified by:

- Prompt notes that clearly indicate how an item should be interpreted in addition to what to expect from responses. Prompt notes may also provide content-related information for raters.
- Annotations (rationales) that clearly indicate how individual papers should be scored
- Training procedures that are outlined and scripted
- Bias training to minimize the possible impact of bias that raters may bring to the scoring session
- Calibration of raters to ensure that they apply the scoring guides consistently from administration to administration

Content Category Information

On many Texas Educator Certification Program tests, items are grouped into content categories known as domains and competencies. To help test takers in further study or prepare to retake the test, the test taker score report shows how many "raw points" have been earned in each domain or competency. For domains and competencies consisting of MC items, "raw points" means the number of items answered correctly. For domains and competencies consisting of CR items, "raw points" are based on the sum of the ratings awarded to the answer.

ETS provides Educator Preparation Programs (EPPs) with the same level of individual student category information that is provided to test takers because of EPPs' desire to assist test takers in developing study plans and to have information about the effectiveness of their preparation. Although this information is currently being supplied, ETS cautions that category scores are less reliable than total test scores, given the reduced number of items measuring a category. Furthermore, raw domain scores are not comparable across test forms because test forms may vary in difficulty. ETS encourages EPPs to consider other information about a student's understanding in addition to domain or competency scores when making instructional decisions for students.

Score Reporting

Score reporting is the process by which test results are calculated and made available to test takers, Educator Preparation Programs (EPPs) and Texas Education Agency (TEA).

Quality Assurance Measures

Responses to MC items on computer-delivered tests are automatically verified before scores are reported. For paper-based tests, although the machine scanning of MC answer sheets allows virtually no opportunity for scoring error, test takers who believe that their tests might have been scored incorrectly may request Score Review.

For those tests that include CR items, all responses are scored by human raters who have been carefully screened, provided with extensive training materials and other support, and monitored throughout the scoring process. Although all constructed responses are reviewed and scored by a minimum of two raters, test takers who believe that their constructed responses might have been scored incorrectly may request Score Review.

Score Reports

Each test taker receives access to a (printable) score report that includes the test taker's overall score, passing status and, where applicable, information regarding performance on specific areas of the test.

Also accessible (and printable) is a document entitled *Understanding Your Texas Educator Certification Program Test Scores* that provides extensive information to help test takers better understand and interpret the test results presented on their score reports.

Scores and other test results are also reported to TEA and made accessible to the test taker's EPP.

Statewide Summary Reports

Statewide Summary Reports provide EPPs with summaries of aggregated test scores for their test takers. The Computer-Administered Testing (CAT) Statewide Summary Report, which is based on an administration month, is accessible to EPPs one month following the end of the administration month. The Paper-Based Testing (PBT) Statewide Summary Report, which is based on an administration date, is accessible to EPPs one day after each administration's CR scores are reported to candidates. Both reports are available through the ETS Data Manager (EDM) tool. Annual (CAT and PBT) Statewide Summary Reports contain the same information as the monthly and administration-based reports but summarize the data for an entire testing year.

Title II Reporting

ETS assists the Texas Education Agency and Educator Preparation Programs (EPPs) in preparing reports to comply with federal reporting requirements on the quality of their teacher preparation programs. These requirements are commonly known as Title II. In October 1968, Congress enacted Title II of the Higher Education Act (HEA) and reauthorized it as part of the Higher Education Opportunity Act (HEOA) that amended the HEA.

Section 207 of Title II requires the annual preparation and submission of three reports on teacher preparation and licensing: one from educator preparation programs to states, a second from states to the U.S. Secretary of Education, and a third from the Secretary of Education to Congress and the public.

By law, these reports must be submitted annually.

Appendix – Statistical Characteristics of Texas Educator Certification Program Tests

Table 1 in this section provides important scoring and statistical information for many of the Texas Educator Certification Program tests. Notes at the end of the table provide more information about the data included.

- Number of Test Takers — Represents the annual volume for the 2015–16 testing year. If a test taker took a test more than once within this period, that person is only counted at the first attempt.
- Average Reported Score — Mean reported score of test takers who tested during the 2015–16 testing year. If a test taker took a test more than once within this period, only the first attempt was used in this calculation.
- Standard Deviation — Standard deviation of the reported score of test takers who tested during the 2015–16 testing year. If a test taker took a test more than once within this period, only the first attempt was used in this calculation.
- Pass Rate — Average passing rate of test takers who tested during the 2015–16 testing year. If a test taker took a test more than once within this period, only the first attempt was used in this calculation.
- Reliability — The tendency of individual scores to be consistent from one version of the test to another. For mixed-format tests (i.e., multiple-choice and constructed-response) with fewer than two constructed-response questions, reliability is calculated for only the multiple-choice portion of the test. For tests with insufficient data, reliability is not calculated.
- Standard Error of Measurement — A statistic that is often used to describe the expected variation in a test score if an individual is retested many times with parallel forms of a test. A test taker's score on a single version of a test will differ somewhat from the score the test taker would get on a different version of the test. The more consistent the scores from one version of the test to another, the smaller the standard error of measurement. If a large number of test takers take a test for which the standard error of measurement is 3 points, about two-thirds of the test takers will receive scores within 3 points of the scores that they would get by averaging over many versions of the test. On some tests, the standard error of measurement could not be estimated because there was no version of the test that had been taken by a sufficient number of test takers. On other tests, the standard error of measurement could not be adequately estimated because the test consists of a very small number of questions or tasks, each measuring a different type of knowledge or skill. Finally, for tests containing both multiple-choice and constructed-response questions where the number of constructed-response questions is less than two, the standard error of measurement for the reported score could not be estimated.

- Standard Error of Scoring — For tests with constructed-response components, where the scoring involves human judgment, this statistic describes the reliability of the process of scoring the test takers' responses. It is an estimate of the correlation between the scores resulting from two independent replications of the scoring process. It includes as measurement error only the independent replications of the scoring process. (Because it does not take into account the adjudication of discrepancies between the first and second ratings, the standard error is a slight underestimate of the correlation of two complete scorings). If a large number of test takers take a test for which the standard error of scoring is 1 point, about two-thirds of the test takers will receive scores within 1 point of the scores that they would get if their responses were scored by all possible scorers. On some constructed-response tests, the standard error of scoring could not be estimated because there was no version of the test that had been taken by a sufficient number of test takers. On some constructed-response tests, the standard error of scoring could not be estimated because the responses were not all scored independently by two different scorers. The standard error of scoring for a multiple-choice test, or a domain or competency score consisting of only multiple-choice questions, is not applicable because multiple-choice scoring is a purely mechanical process with no possibility of disagreement between scorers.

Table 1 — Statistical Summary Statistics for Total Scaled Scores

The table below gives the Number of Test Takers, Average Reported Score, Standard Deviation, Pass Rate, Reliability, Standard Error of Measurement, and Standard Error of Scoring for many of the Texas tests.

Test Code	Test Name	Number of Test Takers	Average Reported Score	Standard Deviation	Pass Rate	Reliability	Standard Error of Measurement	Standard Error of Scoring
068	Principal	4329	248	14.93	72	0.78	7.79	n/a
072[a]	Texas Assessment of Sign Communication (TASC)	33	3.36	0.88	88	n/a	n/a	n/a
073[a]	Texas Assessment of Sign Communication (TASC–ASL)	52	3	1.29	60	n/a	n/a	n/a
085	Master Reading Teacher	55	260.42	14.72	91	0.79	n/a	3.88
086	Master Technology Teacher	4	250.25	13.59	75	n/a	n/a	n/a
087	Master Mathematics Teacher EC–4	9	256.78	22.54	78	0.86	n/a	n/a
113	English Language Arts and Reading/Social Studies 4–8	433	254.2	20.65	79	0.88	6.65	n/a
114	Mathematics/Science 4–8	329	252.21	22.93	75	0.89	7.15	n/a
115	Mathematics 4–8	1920	246.59	29.4	64	0.89	9.17	n/a
116	Science 4–8	1149	243.28	24.98	59	0.85	8.7	n/a
117	English Language Arts and Reading 4–8	1939	255.16	21.53	78	0.88	8.16	n/a
118	Social Studies 4–8	1050	244.92	27.35	64	0.88	8.81	n/a
129	Speech 7–12	527	249.54	23.11	68	0.87	8.05	n/a
139	Technology Applications 8–12	96	239.98	21.61	60	0.86	7.32	n/a
141	Computer Science 8–12	368	245.48	21.56	67	0.92	6.13	n/a
142	Technology Applications EC–12	567	257.12	19.49	83	0.87	7.33	n/a
150	School Librarian	294	253	16.77	78	0.7	8.87	n/a

Test Code	Test Name	Number of Test Takers	Average Reported Score	Standard Deviation	Pass Rate	Reliability	Standard Error of Measurement	Standard Error of Scoring
151	Reading Specialist	214	272.74	11.51	99	n/a	n/a	n/a
152	School Counselor	1504	261.16	13	94	0.75	7.63	n/a
153	Educational Diagnostician	493	256.47	16.14	86	0.8	7.96	n/a
154	English as a Second Language Supplemental (ESL)	15339	252.93	18.96	78	0.71	10.52	n/a
157	Health EC–12	849	260.29	16.25	89	0.8	7.42	n/a
158	Physical Education EC–12	3063	254.72	18.63	81	0.81	9.5	n/a
160	Pedagogy and Professional Responsibilities EC–12	26899	265.24	15.97	93	0.86	8.51	n/a
161	Special Education EC–12	6772	253.4	18.79	80	0.89	6.83	n/a
162	Gifted and Talented Supplemental	469	257.17	12.7	92	0.74	7	n/a
163	Special Education Supplemental	693	252.67	14.5	85	0.8	6.84	n/a
164	Bilingual Education Supplemental	2568	246.06	18.04	66	0.73	8.71	n/a
171	Technology Education 6–12	448	267.14	15.33	94	0.91	4.96	n/a
172	Agricultural Science and Technology 6–12	325	260.08	14.82	93	0.83	6.19	n/a
173	Health Science Technology Education 8–12	145	276.66	10	100	0.8	4.88	n/a
175	Marketing Education 8–12	32	248.94	10.99	84	0.82	7.11	n/a
176	Business Education 6–12	773	246.99	16.84	73	0.83	6.84	n/a
177	Music EC–12	1203	251.64	17.4	79	0.84	7.08	n/a
178	Art EC–12	1099	263.78	14.9	94	0.84	6.69	n/a
179	Dance 8–12	229	248.89	19.76	76	0.78	8.28	n/a
180	Theatre EC–12	424	251.96	18.71	75	0.84	6.97	n/a
181	Deaf and Hard of Hearing	93	256.19	17.65	83	0.77	8.34	n/a
182	Visually Impaired	50	258.46	12.36	90	0.76	7.02	n/a
183	Braille	55	265.27	14.89	96	0.77	9.54	n/a
184	American Sign Language (ASL)	67	266.9	19.18	88	0.89	8.8	n/a
190	Bilingual Target Language Proficiency Test (BTLPT) Spanish	2750	245.17	28.5	63	0.89	8.49	5.25
195	Superintendent	468	254.94	10.78	91	0.69	6.51	n/a
231	English Language Arts and Reading 7–12	3000	243.01	25.21	64	0.85	10	4.17
232	Social Studies 7–12	3413	234.27	27.08	46	0.9	7.4	n/a
233	History 7–12	1099	242.82	24.59	61	0.86	8.44	n/a
235	Mathematics 7–12	2446	239.89	32.82	58	0.93	8.67	n/a
236	Science 7–12	1772	240.75	27.57	58	0.92	7.32	n/a
237	Physical Science 6–12	92	225.14	35.55	42	0.92	9.22	n/a
238	Life Science 7–12	987	234.48	29.33	45	0.88	9.12	n/a
240	Chemistry 7–12	127	241.11	34.25	60	0.9	9.1	n/a
243	Physics/Mathematics 7–12	105	244.83	28.4	61	0.91	7.96	n/a
256	Journalism 7–12	204	251.44	15.95	85	0.79	7.9	n/a
270	Pedagogy and Professional Responsibilities for Trade and Industrial Education 6–12	273	255.27	17.47	84	0.86	8.03	n/a

Test Code	Test Name	Number of Test Takers	Average Reported Score	Standard Deviation	Pass Rate	Reliability	Standard Error of Measurement	Standard Error of Scoring
272[b]	Agriculture, Food, and Natural Resources 6–12	56	258.11	16.26	89	n/a	n/a	n/a
273[b]	Health Science 6–12	94	254.99	18.79	83	n/a	n/a	n/a
274	Mathematics/Physical Science/Engineering 6–12	81	255.16	26.86	77	n/a	n/a	n/a
275[b]	Marketing 6–12	49	252.12	15.2	82	n/a	n/a	n/a
276[b]	Business and Finance 6–12	156	236.72	20.85	49	n/a	n/a	n/a
610	Languages Other Than English – French EC–12	100	229.07	26.84	44	0.9	7.96	2.67
611	Languages Other Than English – German EC–12	26	249.73	23.58	65	0.94	7.1	2.77
612	Languages Other Than English – Latin EC–12	31	258.9	22.43	81	n/a	n/a	n/a
613	Languages Other Than English – Spanish EC–12	1224	238.23	24.08	50	0.88	7.85	2.78
801	Core Subjects EC–6 ELAR and STR	17749	255.87	20.02	85	0.81	9.12	n/a
802	Core Subjects EC–6 Mathematics	17749	252.82	25.74	76	0.79	11.76	n/a
803	Core Subjects EC–6 Social Studies	17749	246.7	26.42	71	0.74	13.24	n/a
804	Core Subjects EC–6 Science	17749	248.09	22.77	73	0.76	11.1	n/a
805	Core Subjects EC–6 Fine Arts, Health & Physical Education	17749	257.17	18.23	90	0.72	9.92	n/a
806	Core Subjects 4–8 English Language Arts and Reading	3747	244.99	24.88	66	0.83	10.37	n/a
807	Core Subjects 4–8 Mathematics	3747	245.75	29.42	72	0.81	12.61	n/a
808	Core Subjects 4–8 Social Studies	3747	245.05	25.83	71	0.74	13.43	n/a
809	Core Subjects 4–8 Science	3747	247.56	29.06	71	0.8	13.39	n/a

[a] For test codes 072 and 073, the summary statistics were calculated by converting alphabetic scores reported to candidates to numeric scores
(A = 5, B = 4, C = 3, D = 2, E = 1).

[b] These tests were new during the 2015–16 testing year and were taken by too few test takers to estimate Reliability and Standard Error of Measurement.

Bibliography

American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME). (2014). *Standards for Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education*, 17, 59–88.

Cizek, G. J., & Bunch, M. B. (2007). Standard setting: A guide to establishing and evaluating performance standards on tests. Thousand Oaks, Calif.: Sage.

Code of Fair Testing Practices in Education (2004). Washington, D.C.: Joint Committee on Testing Practices. (Mailing address: Joint Committee on Testing Practices, Science Directorate, American Psychogical Association, 750 First Street, NE, Washington, D.C. 20002-4242).

Cronbach, L. J. (1951). Coefficient Alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.

Dorans, N. and Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing differential item functioning on the Scholastic Aptitude Test. *Journal of Educational Measurement*, 23, 355–368.

ETS Standards for Quality and Fairness (2014). Princeton, N.J.: Educational Testing Service.

Geisinger, K. F., & McCormick, C. M. (2010). Adopting cut scores: Post-standard-setting panel considerations for decision makers. *Educational Measurement: Issues and Practice, 29*, 38–44.

Haertel, E.H. (2006). Reliability. In B.L Brennan (Ed*.), Educational Measurement* (4th ed., pp. 65–110). Washington, D.C.: American Council on Education.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational Measurement,* Fourth Edition, pp. 433–470. Westport, Conn.: American Council on Education/Praeger.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education,* 8, 41–55.

Holland, P.W. and Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H.I. Braun (Eds.), *Test validity*, pp. 129–145. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Holland, P. W. and Wainer, H. (1993). *Differential item functioning.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4 ed., pp. 17–64). Westport, Conn.: American Council on Education/Praeger.

Kolen, M. J. and Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and Practices* (2nd Ed.). New York, N.Y.: Springer-Verlag.

Kuder, G. F. and Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2, 151–160.

Kuehn, P. A., Stallings, W. M., and Holland, C. L. (1990). Court-defined job analysis requirements of teacher certification tests. *Educational Measurement: Issues and Practice*, 9, 21–24.

Livingston, S.A. and Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32, 179–197.

Lord, F. M. (1984). Standard errors of measurement at different ability levels. *Journal of Educational Measurement*, 21, 239–243.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from the retrospective analysis of disease. *Journal of the National Cancer Institute,* 22 (4), 719–748.

Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice,* 27, 15–29.

Puhan, G., Moses, T., Grant, M., and McHale, F. (2009). Small Sample Equating Using a Single Group Nearly Equivalent Test (SiGNET) Design. *Journal of Educational Measurement, 46 (3),* 344–362.

State Board for Educator Certification/Texas Education Agency. (2015). *Texas Educator Certification Program: TExES™ Faculty Manual.* Retrieved from: http://cms.texes-ets.org/texes/downloadlibrary/.

Tannenbaum, R. J., & Katz, I. R. (2013). Standard setting. In K. F. Geisinger (Ed.), *APA Handbook of Testing and Assessment in Psychology*. Washington, D.C.: American Psychological Association.

Tannenbaum, R. J. and Rosenfeld, M. (1994). Job analysis for teacher competency testing: Identification of basic skills important for all entry-level teachers. *Educational and Psychological Measurement*, 54, 199–211.

Von Davier, A. A., Holland, P. W., and Thayer, D. T. (2004). *The kernel method of equating*. New York, N.Y.: Springer.

Zwick, R., Donoghue, J. R, and Grima, A. (1993). Assessment of Differential Item Functioning for Performance Tasks. *Journal of Educational Measurement*, 30, 233–251.