Chapter 3 Standard Technical Processes

Overview
- Standards
- Scaling
- Equating
- Reliability
- Validity
- Measures of Student Progress
- Sampling

Technical Details and Procedures
- Standards
- Scaling
- Equating
- Reliability
- Validity
- Measures of Annual Improvement
- Sampling

Overview

The Texas assessment program strives to create tests that are fair, accurate, valid, and reliable. The Standards for Educational and Psychological Testing (American Psychological Association, 1999) by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education provides a set of guidelines for evaluating the quality of testing practices. By using these standards to guide test development, the Texas Education Agency (TEA) can make certain that Texas assessments are technically adequate and appropriate for their purposes.

The purpose of this chapter is to provide a general description of the procedures TEA follows to promote fairness, accuracy, validity, and reliability in the Texas assessment program. The specific procedures used within each component of the assessment program are provided in subsequent chapters. This chapter is divided into two sections: an overview and technical details. Each section describes six concepts in varying degrees of detail. The overview section explains each of the six concepts without requiring the reader to have technical expertise to understand the information. The technical details section of the chapter elaborates the six concepts as procedures TEA uses to create tests; some technical background knowledge may be necessary to understand information presented in the final section. The six concepts that are described in this chapter are listed below, with a brief definition of each.

Standards relate levels of test performance directly to what students are expected to learn as expressed in the statewide curriculum by establishing cut scores that define performance categories like Met Standard and Commended Performance.

Scaling is a procedure that is used to transform test scores systematically so that they are easier to interpret and can be compared across test administrations.

Equating is used in conjunction with scaling to put different tests on a common scale. Equating also makes test scores easier to interpret and compare across test administrations.

Reliability refers to the extent to which the test results are consistent across testing conditions.

Validity refers to the extent to which the test measures what it is intended to measure.

Measures of Annual Improvement track student performance across time and can be broadly classified as either growth or projection measures. Improvement measures track student performance from grade to grade, whereas projection measures use current student performance to predict future performance.

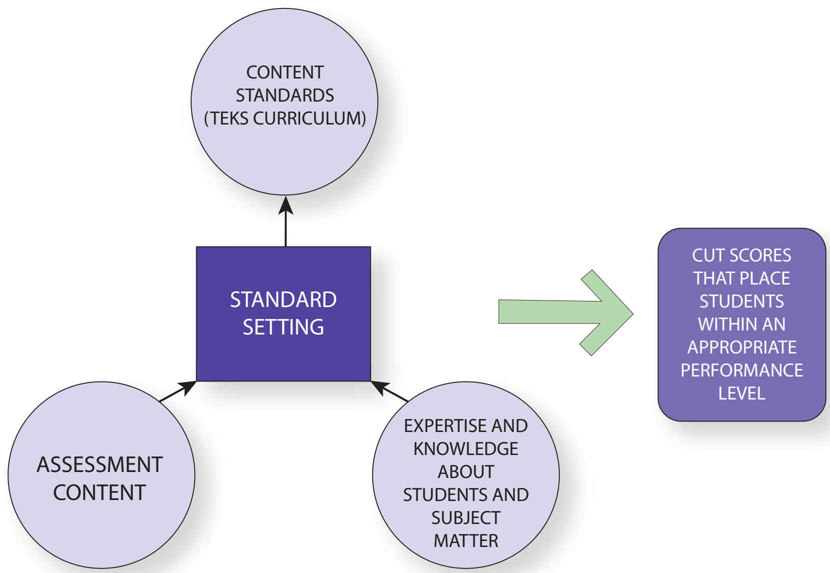
Sampling is a procedure to select a smaller number of observations (in this case, Texas students) that are representative of the entire body of Texas students. The results from well drawn samples allow TEA to estimate characteristics of the larger population of Texas.

Standards

One of the most critical aspects of a statewide testing program is the establishment of performance levels that provide a frame of reference for interpreting the test scores. Once an assessment is given, students, parents, educators, administrators, and policymakers want to know, in clear language, how students performed on that assessment. Performance standards relate test performance directly to the student expectations expressed in the state curriculum, the Texas Essential Knowledge and Skills (TEKS), in terms of what students are expected to learn by the completion of each grade level. Performance standards describe the level of competence students are expected to exhibit as they progress through the educational system.

Through the process of standard setting, cut scores are determined to reflect the performance standards. Performance standards provide information about what a test score means and facilitate the translation of student scores into decisions about individual students. In Texas, for example, the Texas Assessment of Knowledge and Skills (TAKS) standard-setting process results in two cut scores that partition three performance levels: Commended Performance, Met Standard, and Did Not Meet Standard. The standard-setting process is the process of linking content standards (i.e., the TEKS), educator knowledge about student performance, and the assessment's content to set cut scores that place students within one of several performance levels as depicted in Figure 2.

Figure 2. The Standard-Setting Process



The general process of setting performance standards in Texas includes the following steps:

1. Texas educators use the TEKS curriculum to establish content standards.
2. Policymakers set general performance level descriptors based on the content standards. In Texas, for example, the State Board of Education (SBOE) determined that there would be three descriptors for TAKS—Commended Performance, Met Standard, and Did Not Meet Standard.
3. Standard-setting panelists take the general descriptors and make them specific by elaborating on what they mean for students in a particular content area and grade level. The content-specific performance level descriptors describe what students should know and be able to do in each of the performance categories on the test.

4. Using the content-specific performance level descriptors, standard-setting panelists complete the standard-setting process and produce a recommendation for cut scores that indicate how the general performance level descriptors map onto the test scores.
5. The performance standards recommendation from the standard-setting committee is then submitted to either the Commissioner of Education or the SBOE depending on which is responsible for determining the final performance standards.

Scaling

Scaling is the process of associating a system of numbers to some characteristic of interest. We use different types of scales every day to indicate information about temperature, time, speed, etc. For example temperature is frequently expressed using the Fahrenheit scale—it is 102 degrees Fahrenheit. However, temperature can also be expressed on a different scale such as the Celsius scale—102 degrees Fahrenheit is equivalent to 39 degrees Celsius. Both numbers refer to the same temperature, but express it on different scales. Test scores work in a similar fashion.

The number of items that a student answers correctly on a given test is known as the raw score, which can be interpreted only in terms of a specific set of test questions. Raw scores from different test forms should not be compared unless the set of questions on both forms is identical. A hypothetical example can help illustrate the reasons why. If 75% of students earn at or above a raw score of 34 out of 50 on one test, and 80% of students earn at or above a raw score of 34 out of 50 on a separate test with a different set of questions, there are two possible explanations for the differing group performance. It is possible that the second group of students was more proficient than the first. However, it is also possible that the second test was easier than the first. When subsequent test administrations use new forms, the questions on the new forms are likely to be different from those on the old forms. Although the different forms may target the same knowledge and skills, it is likely that some forms will be slightly easier or more difficult than others. Thus, differences in student performance cannot necessarily be directly compared using the raw scores across testing administrations unless the tests are placed onto a common scale.

When different tests are placed onto a common scale, the resulting scores are called scale scores. Unlike raw scores, scale scores do allow direct comparisons of student performance across separate test forms and different test administrations. A scale score is a conversion of the raw score onto a “scale” that is common to all test forms for that assessment. The scale score takes into account the difficulty level of the specific set of questions. The scale score informs test users about students' performance relative to each other and relative to the performance standards across separate assessment forms. Scaling is the process of creating the scale scores.

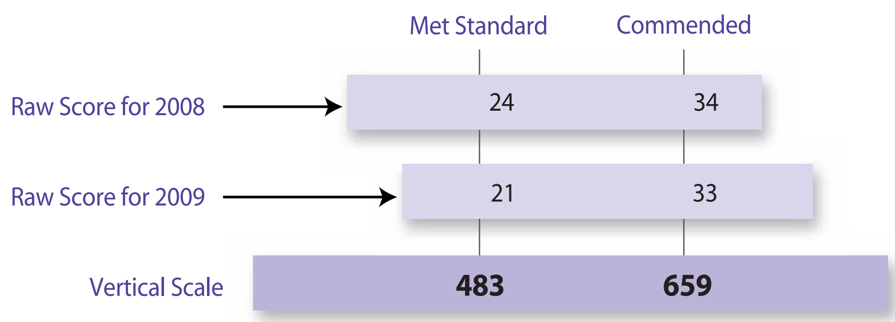
There are two types of scale scores generated for Texas tests: horizontal scale scores and vertical scale scores. Vertical scale scores were provided as supplementary information for reading and mathematics tests of TAKS English grades 3–8 and for TAKS Spanish grades 3–6, and horizontal scale scores were used in most tests in Texas in 2008–2009. A horizontal scale remains constant across years. For example, the TAKS horizontal scale ranges from approximately 1300 to 3300 for each grade, and the score necessary to achieve Met Standard is always 2100. Horizontal scores can be compared within the same grades and subjects, but cannot be used to evaluate a student's progress across grades or subjects. By contrast, a vertical scale places scores of tests that measure similar domains at different grades onto a common scale, hence making those scores comparable with one another. The vertical scale for TAKS reading and mathematics for grades 3–8 ranges from 0 to 1000. Under the TAKS vertical scale, the score necessary to achieve Met Standard increases each year, beginning, for example, in mathematics at 500 in grade 3 and increasing to 700 in grade 8.

Equating

Used in conjunction with the scaling process, equating is the process that "balances" the slight difficulty differences across test forms and administrations to place the scores onto a common scale. By using sophisticated statistical methods, TEA "equates" the results of different tests, enabling the comparison of scale scores across test forms and testing administrations. The concrete example below can help to explain the reasoning and purpose behind equating.

Figure 2 illustrates test raw score cut scores and scale scores from two different test forms which vary slightly in difficulty. The TAKS scale score required for Met Standard and Commended Performance remains the same across both forms—483 is the cut score for Met Standard, and 659 is the cut score for Commended Performance. Note, however, that the raw score required to achieve Met Standard in 2008 was 24, whereas the raw score required to achieve Met Standard in 2009 was 21. Likewise, the raw score required for Commended Performance was 34 in 2008, but only 33 was required in 2009. At first glance it may appear that less is expected of students in 2009 than 2008, but that would be a misinterpretation. Rather, because the 2009 test was slightly more difficult than the 2008 test, students of equal ability who scored a 24 on the 2008 test would be expected to achieve a score of 21 on the 2009 test.

Figure 3. Relationship between Raw Scores and Scale Scores at the Performance Standards



Equating is done to ensure equitability. By balancing the slight differences across test versions and administrations, equating enables fair comparisons of results when tests are not equally difficult. The technical aspects of equating are discussed in more detail in the "Technical Details and Procedures" section of this document.

Reliability

The concept of reliability is based on the idea that repeated administrations of the same test should generate consistent results about student performance. Reliability is a critical technical characteristic of any measurement instrument, because unreliable instruments cannot be interpreted in a valid way. Thus, the reliability of test scores should be demonstrated before issues such as validity, fairness, and interpretability can be discussed. There are many different methods for estimating test reliability. Some methods of estimating reliability require multiple assessments to be administered to the same sample of students; however, these types of reliability estimates are burdensome on the schools and students. Therefore, reliability estimation methods that require only one test administration have been developed. The specific methods used in Texas to estimate assessment reliability based on one test administration are detailed later in the chapter.

Validity

The results of TAKS, TAKS–Modified (TAKS–M), TAKS–Alternate (TAKS–Alt), and end of course (EOC) tests are used to make inferences about students' knowledge and understanding of the TEKS. Similarly, the Texas English Language Proficiency Assessment System (TELPAS) test results are used to make inferences regarding English language acquisition in alignment with the English language proficiency standards (ELPS). When test scores are used to make inferences about student achievement it is important that the assessment supports those inferences. In other words the assessment should measure what it was intended to measure in order for any inferences about test results to be valid. Therefore, test makers are responsible for accumulating evidence that supports the intended interpretations and uses of the scores (Kane, 2006). Evidence to support the validity of interpretations and uses of test scores can be categorized into one of the following categories:

1. Evidence based on test content
2. Evidence based on response processes
3. Evidence based on internal structure
4. Evidence based on relations to other variables
5. Evidence based on consequences of testing

The "Technical Details and Procedures" section of this chapter provides information about the specific processes used to collect evidence that supports interpretations and inferences based on the Texas assessment program. The accumulation of validity evidence for test scores is an ongoing process.

Measures of Student Progress

Progress measures track student performance from grade to grade as opposed to comparing, for example, current year grade 4 students with last year's grade 4 students. By tracking individual student performance across time, improvement measures enable the performance of students who have different initial proficiency levels to be compared. In contrast with improvement measures, projection measures do not evaluate student score changes across past years; rather, they project future student performance from current performance. The projections are based on the average performance of previous cohorts of students. Thus, projection measures predict the future performance of current students based on the growth of previous students.

To measure growth, a vertical scale was developed for TAKS English grades 3–8 and for TAKS Spanish grades 3–6. The vertical scale enables student growth to be tracked by analyzing grade to grade score changes. The vertical scale was reported for informational purposes in the 2008–2009 school year and will be implemented beginning with the 2009–2010 school year. As for projection measures, Texas implemented the Texas Projection Measure (TPM) during the 2008–2009 school year for the TAKS test. More information about the vertical scale is presented in the "Technical Details and Procedures" section of this chapter. Further information about the TPM can be found in chapter 2.

Sampling

Sampling plays a critical role in the research and annual development activities necessary to support the Texas assessment program. The assessment program affects all students (or the population of students) in Texas. A sample is a group of students smaller than the population that can be used to represent the overall population.

Through the careful selection of student samples, TEA is able to gather reliable information about student performance on its tests while minimizing campus and district participation. In particular, sampling is used in the Texas assessment program for testing that is part of a research study and stand-alone field tests.

Research studies in general involve assessing a subset of students from the state population, sampled and assigned to various testing conditions, to support the reliability and validity of the assessment program. Stand-alone field-test results are used to evaluate statistical properties of newly-developed test items prior to their use on a live test form. Because the results will be generalized to the overall student population, the way in which a sample of students is selected is critical. Samples for stand-alone field tests are carefully selected to mirror important characteristics of the state population such as ethnicity and campus size. The results can then be generalized and used to make recommendations on item selections for use on future Texas assessments.

Technical Details and Procedures

Standards

A variety of research-based methods are available for setting performance standards to relate test performance to the level of expected attainment. This section describes some of the standard-setting methods used in Texas. For a summary of some commonly used standard-setting methods, refer to *Standard Setting* by Cizek and Bunch (2007).

Modified Item-Mapping Method

One method used to set standards in Texas is a modified item-mapping method (Lewis, Green, Mitzel, Baum, & Patz, 1998). The item-mapping method has been chosen for several reasons. First, it is a method that has been used in many other statewide testing programs to set high-stakes educational standards. Second, it is a procedure well suited for assessments that contain both selected-response (multiple-choice) and constructed-response (short-answer) items, which some TAKS tests contain.

Item-mapping is a method for setting standards where panelists make judgments about where minimally proficient (or advanced) students would be able to correctly answer items in a specially constructed test booklet of test items ordered from easiest to most difficult (Ordered Item Booklet). The panelists indicate the last item in the Ordered Item Booklet that the minimally proficient (or advanced) would be expected to answer correctly. A thorough description of the components of the process used to set performance standards on the TAKS can be found in chapter 11 of the *2003–2004 Technical Digest*. A brief description of the modified item-mapping method used to set standards on the TAKS test is detailed on the following page:

- Round 1—Panelists are asked to imagine a group of students who just barely met the standard (borderline students) and a group of students who just barely achieved Commended Performance. Panelists then receive an Ordered Item Booklet and rating sheets. Prior to each round, panelists read each item, identify the knowledge and skills needed for a correct response, and review the Performance Level Descriptors. Working independently, panelists then place markers at locations in the Ordered Item Booklets to separate Met Standard from Did Not Meet Standard and Commended Performance from Met Standard. In doing so, panelists are reminded to reconsider items in the vicinity of the locations they selected.
- Round 2—Panelists are shown anonymous information about their fellow panelists' judgments. The information that is shown includes the page number on which each panelist placed his/her marker for each of the two cuts and the median item number for each cut. In table groups, panelists discuss their reasons for placing items above or below their cut points. Facilitators provide item p-values (the percentage of students who answered the items correctly) for all the ordered items in the test booklet. Panelists discuss these item-level data with their table groups and then individually reevaluate their cut point decisions and provide Round 2 judgments.
- Round 3—Panelists are again shown a summary of their fellow panelists' anonymous judgments (again as page number placements) and discuss as a whole group the skill requirements of items near the group median cut points. Additional information is provided about the median page number for each cut score as well as a corresponding raw score cut on an operational test. Panelists are then given impact data which shows the percentages of students expected to be in the three performance category levels based on estimates from the first operational test administration or field-test administration. The impact data includes overall percentages and percentages by group (ethnicity, gender, and economically disadvantaged status). Following the presentation and discussion of impact data, the panelists will individually reevaluate their cut point decisions and provide Round 3 judgments.

Item-mapping has also been used to set standards for the TAKS–M assessments, which include many of the same design features of the TAKS.

Other Methods Used in Texas

In addition to the modified item-mapping method described above, Texas has used other methods to set standards. When selecting a standard-setting method, the method should match the assessment format. As mentioned above, item-mapping has been used for TAKS and TAKS–M because it is a method that works well with assessments that use both selected-response and constructed-response items. The Texas assessments for English language learner (ELL) and special education populations include design features tailored towards the testing population. For this reason, different standard-setting methodologies have been used.

For example, the TAKS–Alt used a combination of two different procedures to set standards: a modified extended Angoff procedure (Angoff, 1971; Hambleton & Plake, 1995) was used in combination with a modified performance profile procedure (Morgan, 2003). For a more detailed description of the procedures used to set standards on the TAKS–Alt, see chapter 6.

For the TELPAS grades 2–12 reading tests, a two phase approach was used to set standards (or proficiency levels). During the first phase, an internal work group reviewed item-level data, test-level data, and impact data to recommend a set of cut points for each grade/grade cluster assessment. Then, during the second phase, an external review group of state educators reviewed the cut point recommendations from the first phase, the test forms on which the cut points were based, and impact data. For further information on the methods used to set the proficiency levels for TELPAS reading tests, see chapter 7.

Performance Level Descriptors

The product of the standard-setting process is a set of cut scores that classify students into performance levels. The performance levels for TAKS include: Commended Performance, Met Standard, and Did Not Meet Standard. Brief descriptions of each of the three performance levels are listed below.

Commended Performance represents high academic achievement. Students in this category performed at a level that was considerably above the state passing standard, demonstrating a thorough understanding of the grade-level knowledge and skills.

Met Standard represents satisfactory academic achievement. Students in this category performed at a level that was at or somewhat above the state passing standard, demonstrating a sufficient understanding of the grade-level knowledge and skills.

Did Not Meet Standard represents unsatisfactory academic achievement. Students in this category performed at a level that was below the state passing standard, demonstrating insufficient understanding of the grade-level knowledge and skills.

The performance standards recommendation from the standard-setting committee is submitted to either the Commissioner of Education or the SBOE depending on which is responsible for determining the final performance standards. The standard-setting methods described above for each of the assessment programs were approved by the Texas Technical Advisory Committee (TTAC), which includes national standard-setting experts.

Scaling

The Scaling Process

The scaling process places test score data from different tests onto a common scale. There are three broad approaches to scaling: subject-centered, stimulus-centered, and response-centered (Crocker & Algina, 2006; Torgerson, 1958). Subject-centered approaches aim to locate students on a scale according to the amount of knowledge each student possesses. By contrast, stimulus-centered approaches attempt to place the test items or stimuli on a scale according to the amount of knowledge required to answer each item correctly. Response-centered approaches can be thought of as a combination of subject-centered and stimulus-centered approaches. They are the most complex approaches, seeking to simultaneously locate students and items on a scale based on the how students respond to the items and the difficulty of the items to which they respond.

Texas uses a response-centered approach to scale its tests, which involves specialized statistical methods that estimate both student proficiency and the difficulty of a particular set of test items. Specifically, Texas tests use a statistical model known as the Rasch Partial-Credit Model (RPCM) to place test items and measures of student proficiency on the same scale across assessments. This initial scale is then transformed to a more user-friendly metric to facilitate interpretation of the test scores.

Rasch Partial-Credit Model

Test items (multiple-choice, gridded response, short-answer, and essay) for most Texas assessments are scaled and equated using the RPCM. The RPCM is an extension of the Rasch one-parameter Item-Response Theory (IRT) model attributed to Georg Rasch (1980), as extended by Wright and Stone (1979), Masters (1982), Wright and Masters (1982), and Linacre (2001). The RPCM was selected because of its flexibility in accommodating multiple-choice data as well as multiple response category data (e.g., essay response items worth 5 points). The RPCM maintains a one-to-one relationship between scale scores and raw scores, meaning each raw score is associated with a unique scale score. An advantage to the underlying Rasch scale is it allows for comparisons of student performance across years. Additionally, the underlying Rasch scale enables the maintenance of equivalent performance standards across test forms.

The RPCM is defined by the following mathematical measurement model where, for a given item involving $m + 1$ score categories, the probability of person n scoring x on prompt i is given by:

$$P_{xni} = \frac{\exp \sum_{j=0}^{x} (\theta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\theta_n - \delta_{ij})} \qquad x = 0, 1, \ldots, m_i$$

The RPCM provides the probability of a student scoring x on the m steps of question/prompt i as a function of the student's proficiency level, θ_n (sometimes referred to as "ability"), and the step difficulties, δ_{ij}, of the m steps in prompt i. (See Masters, 1982, for an example.) Note that for multiple-choice and gridded-response questions, there are only two score categories: (a) 0 for an incorrect response and (b) 1 for a correct response, in which case the RPCM reduces to the standard Rasch one-parameter IRT model, and the resulting single-step difficulty is more properly referred to as an item difficulty.

The advantages of an RPCM scaling include the following:

- All items, regardless of type, are placed on the same common score scale.
- Students' achievement results are placed onto the same scale. Hence, direct comparisons can be made with respect to the types of items students of differing achievement levels can answer. This facet of the RPCM is helpful in describing test results to students, parents, and teachers.
- Field-test items can be placed on the same scale as items on the live—or operational—tests. This enables student performance on the field-test items to be linked to all items in the test bank, which is useful in the construction of future test forms.
- The RPCM allows for the pre-equating of future test forms, which can help test builders evaluate test forms during the test construction process.
- The RPCM also supports post-equating of the test, which establishes a link between the current form and previous forms. Linking the current form to previous forms enables comparisons of test difficulties and passing rates across forms. Because both pre-equated and post-equated item difficulty estimates are available, any drift in scale or difficulty can be quantified.

Test scores in Texas are transformed from the Rasch scale to a more user-friendly scale using a linear transformation. After the transformation on some TAKS tests, for example, the final horizontal scale ranges from approximately 1200 to 3300. A common scaling allows for the direct comparison of performance-level standards established by the Commissioner of Education or the SBOE for current forms to future test forms.

Vertical Scaling

A vertical scale is a scale score system that allows us to directly compare student test scores across grade levels within a subject. Vertical scaling refers to the process of placing scores of tests that measure similar domains but at different educational levels onto a common scale. When monitoring students' growth from one year to the next, a vertical scale may be preferable to a within-grade horizontal scale, because a horizontal scale does not facilitate cross-grade comparisons whereas a vertical scale makes such comparisons more accessible.

In order to implement a vertical scale, a research study needs to be conducted to determine differences in difficulty across grade levels or grade clusters. Vertical scale studies were conducted in spring 2008 for TELPAS reading and TAKS reading and mathematics grades 3–8 that used some of the embedded field-test (see the "Field-Test Equating" section of this chapter) positions to include vertical scaling items. Several of the regular field-test forms were designated for use in the study. The studies assume a common item design (see the "Equating" section of this chapter) in which items from different grade levels appear together on consecutive grade-level tests, thus allowing direct comparison of item difficulties across grade levels. By embedding field-test items across grade levels, the average difference in item difficulty is calibrated across grades, and an equating constant is applied to all of the items on the grade-level test of interest (see the "Post-equating" section of this chapter).

For detailed information on these studies, see the Vertical Scaling Studies section of the Technical Report series.

For additional information about interpreting the TAKS vertical scale, see the Vertical Scale resources page on TEA's Student Assessment Division website.

Equating

As mentioned earlier, Texas uses the Rasch IRT model to scale its assessments. Kolen and Brennan (2004) describe three data collection designs that facilitate the equating of IRT scales:

1. Two or more tests can be administered to randomly equivalent groups of examinees.
2. Two or more tests can be administered to the same group of examinees, with the test administration order counterbalanced to compensate for time and/or fatigue effects.
3. A common set of test items can be administered to nonequivalent groups.

Texas uses the third data collection design, administering common items to nonequivalent groups, to equate most of its tests because of its relative ease of implementation, and, more importantly, because it is less burdensome on students and schools. Under the common-items/nonequivalent groups design, each sample of students takes a different form of the test with a set of items that are common across tests. The common items, sometimes referred to as anchor items, can be embedded within the test or stand-alone as a separate anchor test. The specific data collection designs and equating methods used in Texas are described below. Interested readers can see Kolen and Brennan (2004) or Petersen, Kolen, and Hoover (1989) for a more detailed explanation of equating designs and methods.

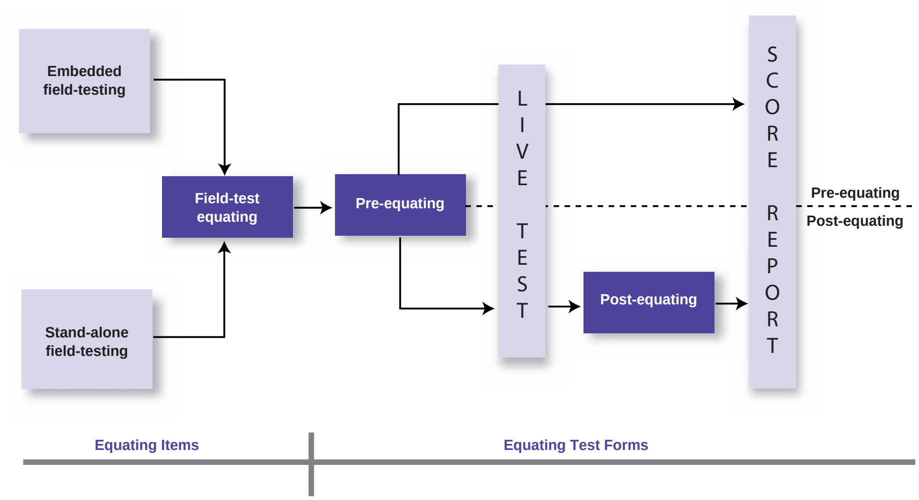
Rationale

There are essentially three stages in the item and test development process where equating takes place:

1. Pre-equating test forms under construction
2. Post-equating operational test forms after administration
3. Equating field-test items after administration

The three stages listed above allow the originally established performance standards for the baseline assessments to be maintained on all subsequent test forms. For example, the TAKS performance standards were set by the SBOE in November 2002, and the tests were administered for the first time in spring 2003. Thus, the base scale for the TAKS test was established with the spring 2003 administration. All subsequent TAKS test forms are equated to this scale, with the exception that new TAKS tests (e.g., grade 8 science) would have base scales established in their first year of implementation. TAKS, TAKS–M, TELPAS reading, and EOC are ongoing programs that require annual equating.

Figure 4. Field-Test Equating Process



The specific equating procedures involve the steps that follow.

Pre-equating

The pre-equating process is one in which a newly developed test is linked, before it is administered, to a set of items that appeared previously on one or more test forms. The goal of pre-equating is to produce a table that establishes the link between raw scores and scale scores before the test is administered. Using pre-calibrated items available in the item bank, TEA staff and Pearson psychometricians construct new test forms by selecting items that meet both the content specifications of the test under construction and the targeted difficulty level for the overall test. For example, for TAKS assessments in grades 5, 8, and at exit level, pre-equated score tables are used for retest forms assembled to give students who have not previously demonstrated a Met Standard level of proficiency additional testing opportunities.

Because the difficulty of the items is established in advance, the difficulty level of newly developed tests can be determined, and the anticipated connection between the raw scores, scale scores, and performance level standards can be identified. Once the anticipated connection between raw scores, scale scores, and performance levels has been established, a raw score to scale score conversion table is produced which maps each raw score to a scale score, and establishes the performance level cut scores.

Post-equating

Wright (1977) outlines the procedure performed on the common-item set to calculate an equating constant to transform the difficulty metric obtained from the current calibration to the same difficulty established by the original test form. This post-equating constant is defined as follows:

$$t_{a,b} = \frac{\sum_{i=1}^{k}\left(d_{i,a} - d_{i,b}\right)}{k}$$

where $t_{a,b}$ is the equating constant, $d_{i,a}$ is the Rasch difficulty of item *i* on current test a, $d_{i,b}$ is the Rasch difficulty of item *i* on previous test *b*, and *k* is the number of common items. Once the equating constant is obtained, it is applied to all item difficulties, transforming them so that they are on the same difficulty scale as the items from the original form. After this transformation, the item difficulties from the current administration of the test are directly comparable with the item difficulties from the original form and with the item difficulties from all past administrations of the test (because such equating was also performed on those items). Since, under the Rasch model, both item difficulty and person proficiency are on the same scale, the resulting scale scores are also comparable from year to year.

The post-equating of TAKS assessments uses conventional common-item/non-equivalent groups equating procedures, whereby the base/live test Rasch item difficulties (live scale) are compared with their previously field-tested values (baseline scale) to derive a post-equating constant. The post-equating constant is calculated as the difference in mean Rasch item difficulty between the common-item set on the baseline scale and the common-item set on the live scale. The post-equating procedure involves the following steps:

1. Tests are assembled and evaluated using Rasch-based objective level and overall targets.
2. Data from the test administrations are sampled.
3. Rasch item calibrations are conducted using the sampled data.
4. A post-equating constant is calculated as the difference in mean Rasch item difficulty of items in the common item set on the baseline scale versus the live scale.
5. The post-equating constant is applied to the live scale item parameter estimates, and raw to scale score conversion tables are produced.

The full equating process is independently replicated by multiple psychometricians for verification. Any significant discrepancies between the replications are reviewed and resolved before results are used to score the tests of Texas students.

Field-Test Equating

To replenish the item bank as new tests are created each year, newly developed items must be field-tested and equated to the scale of the original form. TAKS, TELPAS reading, and EOC use both embedded and separate field-test designs to collect data on field-test items. TAKS English tests that contain only multiple-choice items and TELPAS reading both use an embedded field-test design; whereas TAKS tests containing short-answer or essay items and all TAKS–M tests use a separate, stand-alone field-test design. Additionally, the Spanish version of the TAKS uses a combination of embedded and separate field-test designs.

Once the field-test items are administered, it is necessary to place their difficulties onto the same scale as the original form of the test to enable pre-equating to be done during the test assembly process. Different variants of the common-items equating procedure are used for the TAKS, TAKS–M, TELPAS reading, and EOC tests because of the different field-test designs.

- In tests where field-test items are embedded into a base/live test form, live test items common to each form are used to equate the items to the original test form after the live administration of the test.
- In tests where no operational/live test form existed, a set of linking items common to each separate stand-alone form are used to equate the field-test items to each other after the test is administered.

Embedded Field Tests

After a newly constructed item has cleared the review process, it is embedded in an operational test along with the base-test items. The base-test items are common across all test forms and count toward an individual student's score, but the field-test items appear only in one form and do not count toward students' score. These forms are then spiraled, meaning they are packaged so that tests are randomly assigned to students. Test forms are spiraled so that a representative sample of test takers responds to the field-test items. Each test form is then calibrated separately, with both the base-test items and field-test items combined. Wright's common-items equating procedure, as described previously, is then used to transform the field-test items from each form to the same scale as the common items. Since the scale of the common items had previously been equated to the baseline form, so too are the equated field-test items.

Stand-Alone Field Tests

Newly constructed items that have cleared the review process are assembled into a specified number of forms per subject. These field-test forms then are spiraled across the state, and each student selected to participate in the field tests is administered a single form of the test.

Each stand-alone field-test form contains embedded field-test linking items. Within a subject area, these linking items are common across all field-test forms and serve as the basis for placing all field-test forms onto a common Rasch scale. The goal of stand-alone field-test equating is to take the newly field-tested items and link them to each other, rather than to a previously specified common scale. The stand-alone field-test items are placed on a common Rasch scale by using Wright's common-items equating procedure, as described above.

Comparability Analysis

When the same test is administered in paper and online delivery modes, studies should be conducted to determine whether the use of the same raw score to scale score table for both delivery modes is warranted. Texas uses a comparability method known as Matched Samples Comparability Analysis (MSCA; Way, Davis, & Fitzpatrick, 2006). In this design, a bootstrap sampling approach, which is described in the "Sampling" section of this chapter, is used to select online and paper student samples where each selected online student is matched to a paper student with the same demographic variables and proficiency level on previous test scores. Raw score to scale score conversions are then calculated using Rasch scaling as described above. The sampling is repeated many times. Conversion tables are retained and aggregated across replications, and the mean and standard deviation of the scale scores are taken at each raw score point to obtain the final conversion table scale scores and standard errors of linking respectively. The equivalency between online and paper scale scores is then evaluated using the standard errors as a guide. A separate conversion table is recommended if the two sets of scores are considered not comparable. Between-mode item level analyses and student group analyses for different demographics are also conducted. If mode effects are detected, it may be necessary to use a separate scoring table for each mode of delivery.

For detailed descriptions of past comparability analyses, see the "Comparability Studies" section of the technical report page on TEA's Student Assessment Division website.

Reliability

Internal Consistency Estimates

Reliability measures estimate the degree to which a test is consistent. In theory, reliable tests would generate similar results upon multiple administrations to the same student population. Reliability measures based on one test administration are known as internal consistency measures, because they measure the consistency with which students respond to the items within the test. As a general rule, reliability coefficients from 0.70 to 0.79 are considered adequate, 0.80 to 0.89 are considered good, and above 0.90 are considered excellent. Two types of internal consistency measures used to estimate the reliability of Texas assessments are described below:

- Kuder-Richardson 20 (KR20) is used for tests with only multiple choice items
- Stratified coefficient alpha is used for tests with a mixture of multiple choice and partial credit items

The KR20 is a mathematical expression of the classical test theory definition of test reliability, which expresses test reliability as the ratio of true score (i.e., no measurement error) variance to observed score variance (i.e., measurement error included). The KR20 formula, and the concept of reliability in general, can be expressed symbolically as:

$$Pxx' = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

where the reliability, Pxx', of test X is a function of the ratio between true score variance (σ_T^2) and observed score variance (σ_X^2), which is further defined as the combination of true score variance and error variance ($\sigma_T^2 + \sigma_E^2$). As error variance is reduced, reliability increases (that is, students' observed scores are more reflective of students' true scores or actual proficiencies). KR20 is a method of computing a type of reliability estimate called internal consistency (i.e., how consistently students performed across items within a test) , and can be represented mathematically as:

$$KR20 = \left[\frac{k}{k-1}\right]\left[\frac{\sigma_x^2 - \sum_{i=1}^{k} p_i\,(1-p_i)}{\sigma_x^2}\right],$$

where *KR20* is a lower-bound estimate of the true reliability, k is the number of items in test X, σ_x^2 is the observed score variance of test X, and p_i is the proportion of students who answered item i correctly. This formula is used when test items are scored dichotomously.

Coefficient alpha (also known as Cronbach's alpha) is an extension of KR20 to cases where items are scored polytomously (into more than two categories) and is computed as follows:

$$\alpha = \left[\frac{k}{k-1}\right]\left[1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_x^2}\right],$$

where α is a lower-bound estimate of the true reliability, k is the number of items in test X, σ_x^2 is the observed score variance of test X, and σ_i^2 is the observed score variance of item i.

The stratified coefficient alpha is a further extension of coefficient alpha used when a mixture of item types appears on the same test. In computing the stratified coefficient alpha as an estimate of reliability, each item type component (multiple-choice, short-answer, or essay) is treated as a subtest. A separate measure of internal-consistency reliability is computed for each component and combined as follows:

$$Strat\_\alpha = 1 - \frac{\sum_{j=1}^{c}\sigma_{x_j}^2\left(1-\alpha_j\right)}{\sigma_x^2}$$

where c is the number of item type components, α_j is the estimate of reliability for each item type component, σ_x^2 is the observed score variance for each item type component, and σ_x^2 is the observed score variance for the total score. For components consisting of multiple-choice and short-answer items, a standard coefficient alpha (see above) is used as the estimate of component reliability. The correlation between ratings of the first two raters is used as the estimate of component reliability for essay prompts.

Interrater Reliability

Assessments that are not traditional paper-and-pencil or multiple-choice tests may require reliability evidence that uses a different approach than the measures described above. Some tests, such as the TAKS–Alt, involve teachers observing and evaluating students completing appropriate TEKS-based instructional activities. As part of the process for ensuring the reliability of such tests, TEA must provide evidence that the teacher observation and resulting evaluation of student performance in the TEKS were appropriately conducted.

The interrater reliability study that Texas conducts is a process whereby two trained evaluators first observe the same student performance at the exact same time, and then independently enter the evaluation results at different online sites. Representative samples of observations in which paired evaluators submit student performance evaluations can then be analyzed, and the extent of agreement—or correlation—between the two sets of ratings can be estimated. The correlation between the two sets of ratings is considered to be a measure of the test's reliability. A detailed description of the interrater reliability study conducted to gather reliability evidence for the TAKS–Alt can be found in chapter 16 of the *2006–2007 Technical Digest.*

Measurement Error

Though test scores for Texas assessments are typically highly reliable, each test score contains a component of measurement error. This is the part of the test score that does not measure the characteristic of interest. The measurement error associated with test scores can be broadly categorized as systematic or random. Systematic errors are caused by a particular characteristic of the student or test that has nothing to do with the construct being measured. A language barrier that caused a student to answer a question incorrectly that he or she knew the answer to is an example of a systematic error. By contrast, random errors are chance occurrences. Guessing correctly is an example of random error. Texas computes the Classical Standard Error of Measurement, the Conditional Standard Error of Measurement, and Classification Accuracy for purposes of estimating the amount of random error in test scores.

Classical Standard Error of Measurement (SEM)

The SEM is calculated using both the standard deviation and the reliability of test scores; SEM represents the amount of variance in a score resulting from factors other than achievement. The standard error of measurement assumes that underlying traits such as academic achievement cannot be measured precisely without a perfectly precise measuring instrument. For example, factors such as chance error, differential testing conditions, and imperfect test reliability can cause a student's observed score (the score achieved on a test) to fluctuate above or below his or her true score (the true proficiency of the student). The SEM is calculated as

$$SEM = \sigma_X \sqrt{(1 - r)} ,$$

where r is the reliability estimate (for example, a KR20, coefficient alpha, or stratified alpha) and σ_X is the standard deviation of test X.

The SEM is helpful for quantifying the margin of uncertainty that occurs on every test. It is particularly useful for estimating a student's true score. Unless the test is perfectly reliable, a student's observed score and true score will differ. A standard error of measurement band placed around an observed score will result in a range of values that will most likely contain the student's true score. For example, suppose a student achieves a scale score of 2025 on a test with an SEM of 50. Placing a one-SEM band around this student's score would result in a scale score range of 1975 to 2075. Furthermore, if it is assumed that the errors are normally distributed, it is likely that across repeated testing occasions, this student's true score would fall in this band 68% of the time. Put differently, if this student took the test 100 times, he or she would be expected to achieve a scale score between 1975 and 2075 about 68 times.

It is important to note that the classical SEM index provides only an estimate of the average test score error for all students regardless of their individual proficiency levels. It is generally accepted (see, for example, Peterson, Kolen, & Hoover, 1989) that the SEM varies across the range of student proficiencies. For this reason, it is useful to report not only a test-level SEM estimate but also individual score-level estimates. Individual score-level SEMs are commonly referred to as conditional standard errors of measurement.

Conditional Standard Error of Measurement (CSEM)

The CSEM provides an estimate of reliability that is conditional on the proficiency estimate. In other words, the CSEM provides a reliability estimate, or error estimate, at each score point. Because there is typically more information about students with scores in the middle of the score distribution where scores are most frequent, the CSEM is usually smallest, and scores are most reliable, in the middle of the score distribution.

Item Response Theory (IRT) methods for estimating both individual score-level CSEM and test-level SEM are used because test- and item-level difficulties for TAKS, TAKS–M, TELPAS reading, EOC, and TAAS tests are calibrated using the Rasch measurement model, which is described in detail in the "Scaling" section of this chapter. The standard error of each overall test is calculated as the average conditional standard error across all students. By using CSEMs, which are specific to each scale score, a more precise error band can be placed around each student's observed score.

Classification Accuracy

Test scores are used to classify students into one or more performance categories. For the vast majority of students, we can be confident that these classifications are accurate reflections of their performance. However, since we know that all test scores contain some amount of error, we also know that some students will be misclassified. It is important to understand the expected degree of misclassification. To this end TEA and Pearson conduct an analysis of the accuracy in student classifications into performance categories based on results of tests for which performance standards have been established.

The procedures used for computing classification accuracy for Texas assessments are similar to those recommended by Rudner (2001, 2005).

Under the Rasch model, for a given true proficiency score, θ, the observed proficiency score, $\hat{\theta}$, is expected to be normally distributed with a mean of θ and a standard deviation of $SE(\theta)$. Using this information for a particular level, k, the expected proportion of all students that have a true proficiency score between c and d and an observed proficiency score between a and b is:

$$PropLevel_k = \sum_{\theta = c}^{d} \left(\phi\left(\frac{b-\theta}{SE(\theta)} \right) - \phi\left(\frac{a-\theta}{SE(\theta)} \right) \right) \varphi\left(\frac{\theta-\mu}{\sigma} \right),$$

where ϕ are the cumulative normal distribution functions at the observed score boundaries, and φ is the normal density associated with the true score (Rudner, 2005).

This formula is modified for the current case in the following ways:

- φ is replaced with the observed frequency distribution. This is necessary because the Rasch model preserves the shape of the distribution, which is not necessarily normally distributed.
- The lower bound for lowest performance category (Did Not Meet Standard for TAKS, TAKS–M, and EOC Algebra I, and Beginning for TELPAS reading) and the upper bound for highest performance category (Commended Performance for TAKS, TAKS–M, and EOC Algebra I, and Advanced High for TELPAS reading) are replaced with extreme, but unobserved, true proficiency/raw scores in order to capture the theoretical distribution in the tails.
- In computing the theoretical cumulative distribution, the lower bounds for the Met Standard performance level for TAKS, TAKS–M, and EOC Algebra I, and the Intermediate and Advanced performance levels for TELPAS reading, are used as the upper bounds for the adjacent lower levels, even though under the Rasch model there are no observed true proficiency scores between discrete and adjacent raw score points. This is necessary because a small proportion of the theoretical distribution exists between the observed raw scores, given that the theoretical distribution assumes a continuous function between discrete and adjacent raw score points.
- Actual boundaries are used for person levels, as these are the current observations.

To compute classification accuracy, the proportions are computed for all cells of an "*n* performance category by *n* performance category" classification table. The sum of the diagonal entries represents the classification accuracy for the test. An example of a classification accuracy table for Met Standard and Did Not Meet Standard is presented in Table 2.

Table 2. Classification Accuracy for Met Standard and Did Not Meet Standard

	Met Standard	Did Not Meet Standard
Met Standard	Proportion of Accurate "Met Standard" Classifications	Proportion of Inaccurate "Met Standard" Classifications
Did Not Meet Standard	Proportion of Inaccurate "Did Not Meet Standard" Classifications	Proportion of Accurate "Did Not Meet Standard" Classifications

See chapter 16 of the *2007–2008 Technical Digest* for information on previous classification accuracy studies conducted in Texas.

Validity

Validity in the Texas assessment program is concerned with the general question of to what extent test scores help educators make appropriate judgments about student performance. The concepts described in this section of the chapter are not types of validity, but types of validity evidence. Evidence to support validity can be based on and organized into the five categories described below: test content, response processes, internal structure, relations to other variables, and consequences of testing. Test validation is a matter of degree, not all or none, and is an ongoing process. Furthermore, evidence supporting the validity of a test is evidence supporting the way scores are used and not the scores themselves.

Evidence Based on Test Content

Validity evidence based on test content supports the assumption that the content of the test adequately measures the intended construct. For example, TAKS test scores are designed for making inferences about students' knowledge and understanding of the TEKS. Therefore, evidence supporting the content validity of the TAKS test maps the test content to the TEKS. Validity evidence supporting Texas' test content comes from the established test development process and documentation of subject matter expert judgments about the relationship between the test items and the test construct.

Test Development Process

The test development process starts with a review of the TEKS by Texas educators. The educators then work with TEA to prepare test objectives and determine how the objectives would best be assessed. A test blueprint is developed with educator input, mapping the items to the domains they are intended to represent. Items are then developed based on the test blueprint. Below is a list of steps in the development process that are followed each year to support the validity of test content in Texas beginning with item development:

- writing items based on test objectives and item guidelines
- reviewing items on more than one occasion for appropriateness of item content and identification of item bias
- field-testing of items
- reviewing field-test data with educators
- building tests to pre-defined criteria
- reviewing high-school tests for accuracy of the advanced content by university-level experts

A more comprehensive description of the test development process can be found in chapter 2.

Evidence Based on Response Processes

Response processes refer to the cognitive behaviors required to respond to a test item. Texas collects evidence that the manner in which students are required to respond to test items supports the accurate measurement of the construct of interest. For example, the TAKS writing test includes a writing component in addition to multiple-choice questions because requiring students to respond both to multiple-choice questions and in direct writing provides the most appropriate manner for students to illustrate their ability to write. Student response processes on Texas' assessments differ by item type and administration mode. Texas gathers theoretical and empirical evidence to support the assumption that variability in students' responses is due to differences in the construct of interest and not to an irrelevant construct. Variability in student responses should not depend, for example, on student group. Therefore, Texas gathers supporting evidence that the nature of the test items do not advantage or disadvantage one or more student groups.

Texas gathers evidence to support validity based on response processes from several sources. Test items are pilot tested to gather information about different item types and formats. After item types and formats are determined to be appropriate for tests, evidence is gathered about student responses through field testing, including statistical information such as item difficulty, item-test correlations, and differential item functioning. The evidence is then subjected to educator and expert review.

When students are given the option to take tests either on paper or online, evidence is necessary to show that paper and online response processes lead to comparable score interpretations. Texas conducts the comparability studies described in the "Equating" section of this chapter to evaluate the comparability of online and paper test score interpretations at the test level and at the item level. Score adjustments are made when evidence suggests that students' responses on paper and online are not comparable.

Evidence Based on Internal Structure

When tests are designed to measure a single construct, the internal components of the test should exhibit a high level of homogeneity, which can be evaluated in terms of the internal consistency estimates of reliability described above. Internal consistency estimates are evaluated for Texas tests for reported student groups including all students, female students, male students, African American students, Asian students, Hispanic students, Native American students, and white students. Estimates are made for the full test as well as for each objective within a content area.

Validity studies have also been conducted to evaluate the structural composition of assessments, such as the comparability between the two language versions of the same test. As an example, a study conducted by Pearson (Davies, O'Malley & Wu, 2007) on the structural equivalence of transadapted reading and mathematics tests provided evidence that the Spanish and English version TAKS tests were measuring the same construct, supporting the internal structure validity of the tests.

Evidence Based on Relationships to Other Variables

Another source of validity evidence is the relationship between test performance and performance on some other measure, sometimes called criterion-related validity. The relationship can be concurrent, meaning performance on two measures taken at the same time are correlated, or predictive, meaning the current performance on one measure predicts performance on a future measure. For example, as part of the TAKS Higher Education Readiness Component, a concurrent validity study was conducted in 2004–2005 to evaluate the correlation between performance on exit level TAKS and performance on national testing programs.

Another concurrent validity study conducted annually is the grade correlation study, which compares the passing rates of Texas students on the TAKS tests with their passing rates in related courses. The passing rates are compared overall and across student groups. High percentages of students who pass both the TAKS and the related course provide convergent validity evidence, meaning students are expected to perform similarly on both measures. Low percentages of students who pass the TAKS but fail the related course provide divergent validity evidence, meaning students are not expected to perform well on one measure and poorly on the other measure. Grade correlation studies for grade 10 social studies and science are available on TEA's Student Assessment Division website.

Evidence Based on Consequences of Testing

Consequential validity refers to the idea that the validity of an assessment program should account for both intended and unintended consequences resulting from test score based inferences. For example, TAKS tests are intended to have an effect on curriculum, instructional content, and delivery strategies; however, an unintended consequence could be the narrowing of instruction, or "teaching to the test." Consequential validity studies in Texas use surveys to collect input from various assessment program stakeholders to measure the intended and unintended consequences of the assessments. Pearson worked with TEA, for example, to develop and implement a plan to formally document the evidence of the consequential validity of the TAKS program. Surveys asking about the intended and unintended consequences resulting from the Texas assessment program were administered to the District Advisory Committee (DAC) and educator groups. Results from these surveys are disseminated and considered to help ensure the continuous improvement of the Texas assessment program. The survey results from the TAKS–Alt consequential validity study can be found in the *TAKS–Alt Technical Report 2006–2007*.

Measures of Annual Improvement

During the 2008–2009 school year, the TPM was implemented as a progress measure specifically for the TAKS assessment program. Because the TPM was specific to the TAKS, the technical details and procedures concerning its use can be found in chapter 2.

Sampling

Key Concepts of Sampling

Target Population

A target population is the complete collection of objects (for example, students) we want to learn something about (Lohr, 1999). This is the set of students to whom we want to be able to generalize the results. For example, for a study with the goal of understanding how grade 3 English language learners (ELLs) perform on a set of test questions, the target population could be all grade 3 ELLs in Texas. Defining the target population is an important task, and careful consideration is given to defining the target population prior to sampling.

Sampling, Sample and Observation Unit

Sampling is the process of selecting a subset of the target population to participate in a study. A well-drawn sample will enable reliable and valid inferences to be made about the target population. Thus, the primary goal of sampling is to create a small group from the population that is as similar as possible to the larger population.

A sampling unit is the unit to be sampled from the target population. A sampling unit could be a student, a campus, a district, or even a region. For instance, if 20 campuses are randomly chosen from a list of all campuses in the state, then campus is the sampling unit.

An observation unit is the unit on which data are actually collected. An observation unit may or may not be the same as the sampling unit. For example, a study designed to estimate the number of computers per campus in the state might request each of 20 sampled campuses to report the number of computers it has. In this case, the campus is both the sampling unit and the observation unit. Alternatively, if a study required data to be collected on all the students in the selected campuses, then the sampling unit would be campus and the observation unit would be student.

Reasons for Sampling

Texas employs sampling instead of studying entire target populations for several reasons, including:

- **Size.** It is more efficient to examine a representative sample when the size of the target population is quite large.
- **Accessibility.** There are situations where collecting data from everyone that forms the target population is not feasible.
- **Cost.** It is less costly to obtain data for a selected subset of a population than it is for the entire population.
- **Time.** Using the tool of sampling to study the target population is less time-consuming. This consideration may be vital if the speed of the analysis is important.
- **Burden.** Sampling minimizes the requirements for campus and district participation in field testing and research studies, reducing testing burden.

Sampling Designs

Probability Sampling

In a probability sample, all sampling units have a known probability of being selected. Probability sampling requires that the number of sampling units in the target population is known and can be listed. For example, if student is the sampling unit, probability sampling would require an accurate list of all the students in the target population. Random selection, meaning each sampling unit has the same probability of being selected from the list of sampling units, is a key component of probability sampling. The major probability sampling designs include:

- Simple random sampling—all sampling units in the target population have the same probability of being selected.
- Stratified sampling—the list of sampling units is first grouped (i.e., stratified) according to variables of interest; then, a random sample is selected from each group.
- Cluster sampling—the list of all sampling units is first grouped into clusters according to variables of interest. Then, unlike stratified sampling, a predetermined number of clusters are randomly selected. All sampling units within the selected clusters are observed.

The Texas assessment program uses all three of these sampling designs to create samples for the purpose of, for example, field testing or conducting research studies.

Nonprobability (Convenience) Sampling

A sample that is created without the use of random selection is called a nonprobability (or convenience) sample. Convenience samples are selected when collecting a complete list of sampling units is either impractical or impossible. When using convenience sampling, the list of sampling units is incomplete, and sampling units have no known probability of being selected. Convenience sampling introduces sources of potential bias into the resulting data, which makes it difficult to generalize results to the target populations.

Sampling with or without Replacement

Regardless of the type of probability sampling design used, one decision that needs to be made is whether to sample with or without replacement. To help illustrate the distinction between the two sampling methods, consider simple random sampling with or without replacement. Suppose a simple random sampling with replacement of size n is obtained from a population of N. In this case, each time a sampling unit is randomly chosen, it is placed back into the target population and can be chosen again. In other words, when sampling with replacement, it is possible for any given sampling unit to be selected multiple times and have its data duplicated in the resulting samples of size n.

On the other hand, in a simple random sampling without replacement of size *n* from a population of *N*, once a sampling unit is chosen, it is ineligible to be selected again. Thus, when sampling without replacement, each sample consists of *n* distinct, non-duplicate units from the population of size *N*.

Typically, sampling without replacement is preferred over sampling with replacement, because duplicate data adds no new information to the sample (Lohr, 1999). The method of sampling with replacement, however, is very important in resampling and replication methods, such as bootstrapping.

Resampling and Replication Methods: Bootstrap

Resampling and replication methods, such as bootstrapping, treat the sample like a population. They repeatedly take pseudo-samples from samples to estimate variances and standard errors. Thus, sampling with replacement is assumed with these methods. The bootstrap method was developed by Efron (1979), and described in Efron and Tibshirani (1993). Texas uses bootstrapping methods when conducting comparability studies, which compare online and paper versions of a test form.