Chapter 3 Standard Technical Processes

OVERVIEW

The *Standards for Educational and Psychological Testing* (1999) by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1999) provide a set of guidelines for evaluating the quality of testing practices. By using these standards to guide test development, TEA is confident that Texas assessments are technically adequate and appropriate for the purposes for which the tests are used.

The purpose of this chapter is to provide a general description of the technical processes TEA follows to promote fairness, accuracy, validity, and reliability in the Texas assessment program. The specific processes within the assessment program are provided in subsequent chapters. This chapter is divided into two sections: an Overview section and a Technical Details and Procedures section. The Overview section provides an overview of eight technical concepts. The Technical Details and Procedures section elaborates on these eight concepts.

The eight technical concepts described in this chapter are:

Performance Standards. Performance standards relate levels of test performance directly to what students are expected to learn as described in the statewide curriculum.

Item Analyses. Statistical analyses are conducted on the student performance data collected for field-test items. These analyses are used to gauge the level of difficulty of the item, the degree to which the item appropriately distinguishes between students of different proficiency levels, and as a tool to evaluate the item for potential bias.

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

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

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

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

Measures of Student Progress. Measures of student progress describe changes in student performance across time.

Sampling. Sampling is a procedure that is used to select a small number of observations that are representative of the whole. In this case, sampling involves the selection of a set of Texas students that is able to represent the entire body of Texas students. The results from well-drawn samples allow TEA to estimate characteristics of the larger Texas student population.

Performance Standards

A critical aspect of any statewide testing program is the establishment of performance levels that provide a frame of reference for interpreting test scores. After an assessment is administered, students, parents, educators, administrators, and policymakers want to know, in clear language, how students performed on that assessment.

Performance standards help to relate test performance directly to the student expectations expressed in the state curriculum in terms of what knowledge and skills students are expected to demonstrate upon completion of each grade or course. Standards, therefore, describe the level of competence students are expected to exhibit on an assessment.

Standard setting is the process of establishing the cut scores on an assessment that define the performance levels. For example, the STAAR standard-setting process established two cut scores on each assessment, creating three performance levels: Level I: Unsatisfactory Academic Performance; Level II: Satisfactory Academic Performance; and Level III: Advanced Academic Performance.

The Technical Details and Procedures section of this chapter provides information about the standard-setting framework and the specific standard-setting processes that are used to establish the performance standards for the various tests in the Texas assessment program.

Item Analyses

Various statistical analyses are conducted on the student performance data collected for field-test items. Item analyses are conducted annually for the purpose of reviewing the quality of newly field-tested items and help in the evaluation of which items might be included as operational test items on a future administration. The Technical Details and Procedures section of this chapter provides details about the various item statistics that are generated as part of the item analyses.

Scaling

Scaling is the process of associating an array of numbers to a characteristic of interest. Scales of all kinds are used every day to provide information about temperature, time, speed, etc. For example, temperature is frequently described using the Fahrenheit scale: "The high today will be 102 degrees Fahrenheit." However, the same temperature can also be described using a different scale such as the Celsius scale: "The high today will be 39 degrees Celsius." The numbers 102 and 39 both refer to the same temperature, but they describe it using different scales. Test scores can also be described using more than one scale.

The number of items that a student answers correctly on a given test is known as the raw score, and this raw score is interpreted in terms of the specific set of test questions answered. Raw scores from different test forms should not be compared with each other unless the sets of questions that count toward the raw score on both forms are identical. A hypothetical example can help illustrate the reasons for this. If 75% of students earn a raw score of 34 out of 50 on one test form, and 80% of students earn a raw score of 34 out of 50 on a separate test form 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 form was easier than the first test form. Whenever subsequent administrations of a test use new forms, the questions on the new forms are likely to be different from those on the old forms. And despite the fact that different test forms might target the same knowledge and skills, it is likely that some forms will be slightly easier or slightly more difficult than others. As a result, in most

cases differences in student performance cannot be directly compared across testing administrations using raw scores. Instead, the test scores must be placed on a common scale to allow for comparisons.

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

Horizontal scale scores are used to describe student performance within a given grade and content area. A horizontal scale score is used to evaluate student performance relative to the performance standards for a specific grade and content area. Horizontal scales are created separately for each grade and content area, making no reference to potential links or similarities of content across grades. By contrast, vertical scale scores can be used to describe student performance across grade levels within a content area. A vertical scale places scores of tests that measure student performance in similar content areas at different grades onto a common scale, thereby making those scores easily comparable with one another and allowing for simple interpretations of student progress.

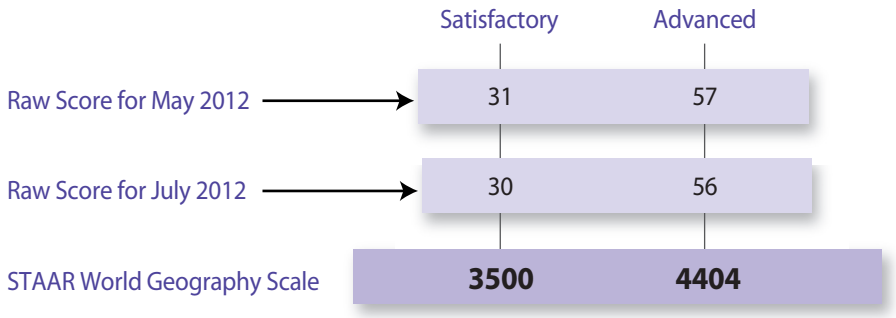
For the STAAR assessments three vertical scales were developed for the following grades and content areas: STAAR grades 3–8 mathematics (one scale for English and Spanish assessments), STAAR grades 3–8 English reading, and STAAR grades 3–5 Spanish reading. TELPAS reading is also reported on a vertical scale. All other STAAR assessments—STAAR grades 4 and 7 writing, grades 5 and 8 science, grade 8 social studies, STAAR EOC assessments, and STAAR Modified assessments—are reported on horizontal scales. Grades 10 and exit level TAKS and grades 10 and 11 TAKS–M are also reported on a horizontal scale.

Equating

Used in conjunction with the scaling process, equating is the statistical process that takes into account the slight differences in difficulty across test forms and administrations and allows scores to be placed onto a common scale. By using statistical methods, TEA "equates" different tests, enabling the comparison of scale scores across test forms and testing administrations.

The example below can help explain the reasoning and purpose behind equating. Figure 3.1 illustrates the relationship among raw scores, cut scores, and scale scores from two different STAAR world geography test forms that vary slightly in difficulty. The scale scores required for Level II: Satisfactory Academic Performance and Level III: Advanced Academic Performance remain the same across both test forms: 3500 is the cut score for Level II, and 4404 is the cut score for Level III. However, the raw scores required to achieve Level II and Level III on the May 2012 form were 31 and 57, respectively, while the raw scores required to achieve Level II and Level III on the July 2012 form were 30 and 56, respectively. At first glance, it might appear that less was expected of students for them to achieve Level II and Level III on the STAAR world geography assessment in July 2012 than in May 2012, but this would be a misinterpretation. Rather, because the items on the July 2012 test form were slightly more difficult than the items on the May 2012 test form when taken as a whole, a student who, for example, scored a 31 on the May test form would have been expected to achieve a score of 30 on the July test form to demonstrate the same level of performance.

Figure 3.1. Relationship Between Raw Scores and Scale Scores at the Performance Standards



Equating is done to promote equitability. By accounting for the slight differences across test forms and administrations, equating enables fair comparisons of results when test forms are not exactly equal in difficulty.

Reliability

The concept of reliability is based on the idea that repeated administrations of the same test should generate consistent results. Reliability is a critical technical characteristic of any measurement instrument, because unreliable instruments cannot be interpreted in a valid way. 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 score 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 schools and students. Therefore, reliability estimation methods that require only one test administration have been developed and are commonly used for various assessments including STAAR.

Validity

The results of STAAR, STAAR Modified, STAAR Alternate, TAKS, and TAKS–M are used to make inferences about students' knowledge and understanding of the TEKS curriculum. Similarly, TELPAS test results are used to make inferences regarding English language acquisition in alignment with the ELPS.

When test scores are used to make inferences about student achievement, it is important that the assessment support those inferences. In other words, the assessment should measure what it was intended to measure in order for inferences about test results to be valid. For this reason test makers are responsible for collecting evidence that supports the intended interpretations and uses of the scores (Kane, 2006). Evidence that supports the validity of interpretations and uses of test scores can be classified into one of the following categories:

- Evidence based on test content
- Evidence based on response processes
- Evidence based on internal structure
- Evidence based on relations to other variables
- Evidence based on consequences of testing

Measures of Student Progress

Student performance is commonly described using performance levels. For example, each STAAR 3–8 and EOC assessment has three performance levels: Level III: Advanced Academic Performance; Level II: Satisfactory Student Performance; and Level I: Unsatisfactory Academic Performance. While this information is useful in describing students' current knowledge and skills, the performance levels alone do not provide information about how students' performance this year compares to performance achieved in the prior year. Student progress measures provide additional feedback by considering not only current performance but also past performance. These measures provide information about the progress students have achieved over time. Performance level descriptors and progress measures together provide an enhanced description of student achievement.

Sampling

Sampling plays a critical role in the research and annual test development activities that are necessary to support the Texas assessment program. The assessment program affects all students (i.e., the "population" of students) in Texas. A sample is simply a group of students smaller than the entire 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 assessments while minimizing campus and district burden. In particular, sampling is used in the Texas assessment program for research studies, audits, and field testing.

In general, research studies involve assessing a subset of students from the state population using various testing conditions in order to collect evidence that supports the reliability and validity of the assessment program. Audits allow for the collection of information from school districts that can be used to evaluate training, administration, and scoring of the STAAR assessments. Results from field testing are used to evaluate statistical properties of newly developed test items as well as items that have not yet been used on an operational 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 are carefully selected to mirror important characteristics of the state population such as ethnicity and campus size. For example, the results of field testing can be generalized and used to make recommendations for item selection in future Texas assessments.

TECHNICAL DETAILS AND PROCEDURES

Performance Standards

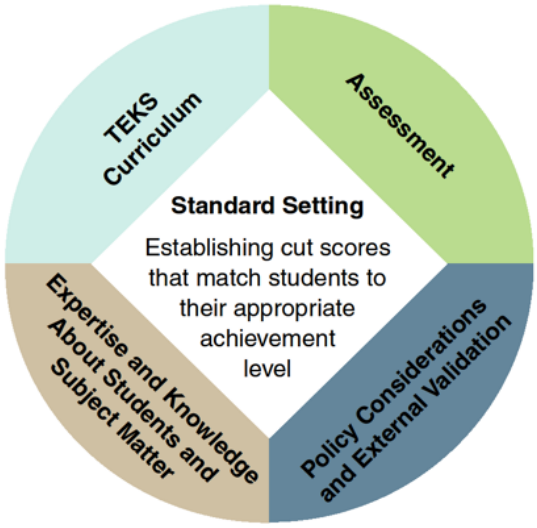
Performance standards relate levels of test performance directly to what students are expected to learn as described in the statewide curriculum. This is done by establishing cut scores that distinguish performance levels or categories. The STAAR programs including the general STAAR, STAAR Spanish, and STAAR L have two cut scores (Level II and III) that identify three performance categories, Level I: Unsatisfactory; Level II: Satisfactory; Level III: Advanced. For the STAAR Alternate, the three performance categories are Level I: Developing, Level II: Satisfactory, and Level III: Accomplished. Standard setting is the process of establishing cut scores on an assessment that define the performance levels. This section describes the standard-setting framework and process for the STAAR (including STAAR 3–8, STAAR EOC, STAAR Modified, and STAAR Alternate), TELPAS, and TAKS (including TAKS–M) testing programs.

Standard Setting for STAAR

As Texas implemented the STAAR program, TEA used an evidence-based standard-setting approach (O'Malley, Keng, & Miles, 2012) to determine the cut scores for the three performance categories (Level III: Advanced (Accomplished) Academic Performance; Level II: Satisfactory Academic Performance; Level I: Unsatisfactory Academic Performance). Standard setting for STAAR involved a process of combining considerations regarding policy, the TEKS content standards, educator knowledge about what students should know and be able to do, and information about how student performance on statewide assessments aligns with performance on other assessments. Standard-setting advisory panels, made up of diverse groups of stakeholders, considered the interaction of all these elements for each STAAR assessment.

Figure 3.2 illustrates the critical elements of the evidence-based standard-setting approach that was used by Texas to establish the STAAR performance standards.

Figure 3.2. Critical Elements of the Evidence-Based Standard-Setting Approach



Each element of the evidence-based standard-setting approach as it relates to the STAAR assessments is described below.

- **TEKS Curriculum Standards:** The TEKS curriculum standards contain the content standards designed to prepare students ultimately to succeed in college and careers and to compete globally. They provide the underlying basis for several key components of the standard-setting process, including the performance labels, policy definitions, and specific performance level descriptors.
- **Assessment:** Each STAAR assessment has been developed to test the knowledge and skills described in the TEKS curriculum standards. Each STAAR assessment is based on the student expectations and reporting categories specified in the STAAR assessed curriculum document and the STAAR test blueprint.
- **Policy Considerations and External Validation:** Research studies, which empirically correlate performance on the STAAR assessments with scores on other related measures or external assessments, were conducted and used to inform the standard-setting process. Stakeholders and experts with experience in educational policy and knowledge of the Texas assessment program considered the results of the research studies when making recommendations about reasonable ranges for setting performance standards.
- **Expertise and Knowledge about Students and Subject Matter:** Texas educators, including classroom teachers and curriculum specialists from elementary, secondary, and higher education, brought content knowledge and classroom experience to the standard-setting process. They played an integral role in developing the performance labels, policy definitions, and specific performance level descriptors and in recommending the performance standards.

- **Standard Setting:** Within the framework of evidence-based standard-setting, an established standard-setting method, such as an item-mapping method with external data method (Ferrara, Lewis, Mercado, D'Brot, Barth, & Egan, 2011; Phillips, 2012), was used to recommend the performance standards.

Using this standard-setting framework, TEA defined and implemented a nine-step process to establish the performance standards for the STAAR, STAAR Modified, and STAAR Alternate assessments. The nine steps are:

1. Conduct validity and linking studies
2. Develop performance labels and policy definitions
3. Develop grade-level/course- specific performance level descriptors
4. Convene a policy committee and/or develop reasonable ranges for performance standards
5. Convene standard-setting committees
6. Review performance standards for reasonableness
7. Approve performance standards
8. Implement performance standards
9. Review performance standards

Table 3.1 provides descriptions of each of the steps in the STAAR standard-setting process. More detail about each of the steps is provided in the *State of Texas Assessments of Academic Readiness (STAAR) Standard Setting Technical Report*, available on the STAAR Resources page of TEA's Student Assessment Division website. Standard-setting reports are also available for the STAAR Modified and STAAR Alternate assessment programs.

Table 3.1: The Nine-Step STAAR Standard-Setting Process

Standard-Setting Step	Description
1. Conduct validity and linking studies	External validity evidence was collected to inform standard setting and support interpretations of the performance standards. Scores on each assessment were linked to performance on other assessments in the same content area.
2. Develop performance labels and policy definitions	Committees recommended performance categories, performance category labels, and general policy definitions for each performance category.
3. Develop grade/course specific performance level descriptors (PLDs)	Committees consisting primarily of educators developed performance level descriptors (PLDs) as an aligned system, describing a reasonable progression of skills within each content area (English, mathematics, science, and social studies).
4. Convene a policy committee and/or develop reasonable ranges for performance standards	For the STAAR EOC assessments, a committee considered policy implications of performance standards and empirical study results and made recommendations to identify reasonable ranges for performance standards ("neighborhoods") for the cut scores. For the STAAR 3–8, STAAR Modified, and STAAR Alternate assessments, the STAAR EOC performance standards and additional empirical study results were used to identify the neighborhoods for Levels II and III cut scores.
5. Convene standard-setting committees	Committees consisting of K–12 educators and higher education faculty used the performance labels, policy definitions, PLDs, and neighborhoods to recommend cut scores for each STAAR assessment.
6. Review performance standards for reasonableness	TEA reviewed the cut-score recommendations across content areas.
7. Approve performance standards	The commissioner of education approved performance standards.
8. Implement performance standards	Once established, performance standards were reported to students for the spring 2012 administration with phase-in standards applied.
9. Review performance standards	Performance standards will be reviewed at least once every three years.

Standard Setting for TELPAS

For the TELPAS grades 2–12 reading tests, a two-phase approach was used to set proficiency level standards in 2008. During the first phase, an internal work group reviewed item-level data, test-level data, and impact data to recommend a set of cut-score ranges for each grade/grade cluster assessment. During the second phase, an external review group of state educators recommended specific cut scores after reviewing the cut-score ranges from the first phase, the test forms on which the phase-one recommendations were based, and impact data.

Standard Setting for TAKS and TAKS–M

To set performance standards on TAKS and TAKS–M, the modified item-mapping method, often referred to as the "bookmark procedure" (Lewis, Green, Mitzel, Baum, & Patz, 1998) was used. 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) students would be expected to answer correctly. A thorough description of the components of the process used to set performance standards on the TAKS is available in chapter 11 of the 2003–2004 Technical Digest.

Item Analyses

Various statistical analyses, based on classical test theory and item response theory (e.g., the Rasch model measurement), are used to analyze the data collected for field-test items. Item analyses are conducted annually for the purpose of reviewing the quality of newly field-tested items and helping in the evaluation of which items may be included as operational test items on a future administration.

Statistics generated for each item as part of the item analyses include: p-value, point-biserial, Rasch item difficulty, Rasch fit, response distribution, and group difference analyses. Detailed descriptions of each statistic along with the statistical guidelines and/or criteria used to evaluate each item are given next.

P-Value

The p-value indicates the percentage of the total group of students answering a multiple choice or gridded response item correctly. An item's p-value shows how difficult the item was for the students who took the item. Items with high p-values, such as 90 (meaning 90% of the students correctly answered this item), are relatively easy items. Those with p-values below 30 (meaning only 30% of the students correctly answered this item) are relatively difficult items.

Point-Biserial

The point-biserial describes the relationship between a student's performance on a multiple choice or gridded response item (correct or incorrect) and performance on the assessment as a whole. A high point-biserial correlation indicates that students who answered the item correctly achieved higher scores on the assessment than those who missed the item. In general, values less than 0.20 indicate a potentially weaker than desired relationship. It should be noted, however, that on items with very high or very low p-values, the point-biserial may be artificially depressed.

Rasch Item Difficulty

Another indication of the difficulty of an item is the Rasch item difficulty estimate. The Rasch Model allows for the comparison of item difficulty across test forms and across different samples of students taking an item across test

administrations. On average, items with negative Rasch item difficulty values (e.g., –1.5) are considered relatively easy, while items with positive values (e.g., +1.2) are considered relatively difficult.

Rasch Fit

The Rasch fit statistic indicates the extent to which student performance on a multiple choice or gridded response item conforms to the technical requirements of the Rasch measurement model. For an item to fit the model well, there should be relatively few cases of uncharacteristic responses (e.g., low-scoring students answering difficult items correctly or high-scoring students missing easy items). In general, a Rasch fit value greater than 1.3 may indicate that the item does not fit the Rasch model.

Response/Score Point Distribution

Response/score point distribution gives the percentage of students responding to each of the answer choices (e.g., A, B, C or D) for a multiple choice item or the percentage of students who received each of the score points (e.g., 0, 1, 2, 3 etc.) for an extended response item (short answer or written composition). Response/score point distributions are provided for the total group (i.e., all students) and by various demographic groups (i.e., gender or ethnicity for STAAR, including STAAR Modified, and TAKS) or by proficiency level groups (i.e., Beginning, Intermediate, Advanced, and Advanced High for TELPAS).

Group Difference Analysis

Statistics from a group difference analysis provide information about how different disaggregated student groups (e.g., males, females, African American, Hispanic, or White students) performed on an item and can help in examining differential item functioning (DIF) across different student groups. Two types of statistical indicators of potential DIF are used in Texas assessment program: the Mantel-Haenszel Alpha and the "ABC" DIF classification (also known as the ETS DIF classification; Petersen, 1987; Zieky, 1993).

Mantel-Haenszel Alpha

The Mantel-Haenszel Alpha is a ratio indicating conceptually the likelihood of the members of each compared student group answering the item correctly, provided that the members of the two groups are performing equally on the assessment. If the value of the reported Alpha is 1, the groups are equally likely to answer the item correctly. If the Alpha value is statistically significantly greater than 1, the chance of success on the item may be better for the designated reference group (e.g., male) than for the focal group (e.g., female). A value statistically significantly less than 1 indicates the item is easier for the focal group. The Mantel-Haenszel Alpha is currently used as the DIF indicator for items on the STAAR, including STAAR Modified, and TAKS assessments.

ABC DIF Classification

The ABC DIF classification is an indicator based on the Mantel-Haenszel Alpha that takes into account both statistical and practical significance when examining an item for DIF. Each item is classified into one of three categories based on each group comparison: "A" means negligible or no presence of DIF, "B" means moderate DIF and "C" means large DIF (for more information refer to Zieky, 1993). Plus and minus signs (+/–) are used to indicate the direction of DIF. A plus sign indicated the item is easier for the focal group (e.g., female); a minus sign indicated the item is easier for the reference group (e.g., male). The ABC DIF classification is currently used as the DIF indicator for items on the TELPAS reading assessments.

Use of DIF Analysis Results

It should be noted that DIF analyses merely serve to identify test items that have unusual characteristics relative to student group performance. They do not specifically identify items that are "biased;" such decisions are made by item reviewers who are knowledgeable about the state's content standards, instructional methodology, and student testing behavior.

Scaling

There are three scales that underlie the STAAR and TELPAS reading assessments: the raw score scale, the Rasch scale, and the reported scale score.

- The raw score scale is defined as the number of items answered correctly regardless of difficulty and includes weighting of short answer responses or written compositions, if applicable.
- The Rasch scale is defined as a transformation of the raw score into a scale that takes into account the difficulty level of the items and the proficiency of the students.
- The reported scale score is defined as a linear transformation of the Rasch scale, through scaling constants, onto a user-friendly scale that takes into account the difficulty of the items and maintains performance standards across test forms and test administrations.

The following sections detail the scaling process in terms of estimating the Rasch scale and transforming the Rasch scale into the reported scale scores.

The Scaling Process

The scaling process places test score data from different tests onto a common scale. There are three primary 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 comparison, 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 and therefore are the most complex. Response-centered approaches simultaneously locate students and items on a scale based on how students respond to the items and how difficult the items are.

TEA scales its assessments using a response-centered approach that involves specialized statistical methods that can estimate both student proficiency and the difficulty of a particular set of test items. Specifically, Texas assessments use a statistical model known as the Rasch Partial-Credit Model (RPCM) to place test items and measures of student proficiency on the same Rasch scale across test forms and test administrations. The Rasch scale is then transformed to the more user-friendly scale score to facilitate interpretation of the test scores.

Rasch Partial-Credit Model

Test items (whether multiple-choice, gridded response, or written composition) 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 (1966), 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., short answer items worth zero to three 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 that 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:

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

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, and the step difficulties, δ_{ij}, of the m steps in prompt i (refer to Masters, 1982, for an example). Note that for multiple-choice and gridded-response questions, there are only two score categories: 0 for an incorrect response and 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.

Some of the advantages of RPCM scaling are:

- All items, regardless of type, are placed on the same common Rasch scale.
- Students' achievement results are placed onto the same scale. Hence, direct comparisons can be made with respect to the difficulty levels 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 Rasch scale as items on the operational tests. This enables student performance on the field-test items to be linked to all items in the item 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 then converted using a linear transformation from the Rasch scale to a more user-friendly reporting scale.

Horizontal Scaling

The STAAR scale scores represent linear transformations of Rasch-based performance estimates (θ). For horizontal scale scores, this transformation is made by first multiplying any given θ by a slope (A) and then adding an intercept (B). This simple operation is represented by the equation below:

$$SS_{\theta} = A \times \theta + B \qquad (1)$$

The slope (A) and intercept (B) in Equation (1) are called scaling constants, and they are derived using a method described by Kolen and Brennan (2004). For STAAR and STAAR Modified, two features of the desired scale score system were established in advance: a scale score value at the passing standard and the standard deviation of the scale. The A scaling constant is calculated as follows:

$$A = \frac{\sigma_{ss}}{\sigma_{\square}} \qquad (2)$$

In Equation (2), σ_{ss} represents the desired standard deviation of the scale, and $\sigma_{\square}$ represents the standard deviation of Rasch-based θ values among a sample group. For example, to construct a horizontal scale for a given STAAR EOC assessment, the sample group for the STAAR EOC assessment consisted of all

students who took that assessment in 2011. For the STAAR 3–8 horizontal scales, the sample group for a given STAAR 3–8 assessment consisted of all students who took that assessment in spring 2012.

The *B* scaling constant is calculated as follows:

$$B = SS_{Level\_II} - \frac{\sigma_{ss}}{\sigma_{\theta}} \times \theta_{Level\_II} \quad (3)$$

Because each assessment's horizontal scale is derived using its own sample group, σ_{θ} varies across assessments. Likewise, each assessment has a unique Level II performance standard in Rasch units, so θ_{Level_II} varies across assessments. SS_{Level_II} and σ_{ss} are set to be consistent within academic content areas but not across all assessments. Once these constants are established, the same transformations are applied each year to the Rasch proficiency level estimates for that year's set of test questions.

A similar process was conducted for TAKS and TAKS–M and is discussed in chapter 8. A unique scale transformation was then developed for each assessment and applied thereafter.

Vertical Scaling

A vertical scale is a scale score system that allows for direct comparison of student test scores across grade levels within a content area. Vertical scaling refers to the process of placing scores of tests that measure similar content areas but at different grade levels onto a common scale. In order to implement a vertical scale, research studies were needed to determine differences in difficulty across grade levels or grade clusters. Such studies were conducted for the STAAR grades 3–8 reading and mathematics and the STAAR Spanish grades 3–5 reading in spring 2012, and for TELPAS in spring 2008. For these studies, embedded field-test positions from several regular field-test forms (refer to the Field-Test Equating section of this chapter) were used to include vertical linking items. The studies assumed a common-item nonequivalent groups design (refer to the Equating section of this chapter) in which items from different grade levels appear together on consecutive grade-level tests, allowing for direct comparison of item difficulties across grade levels. By embedding vertical-linking items across grade levels, the average differences in item difficulties of vertical-linking items can be calculated for each adjacent grade pair, and these linking constants between adjacent grades are then used to create a vertical scale.

For detailed information about these studies, refer to the "Vertical Scaling Studies" section of the Technical Digests and Reports page on TEA's Student Assessment Division website.

Similar to the horizontally-scaled assessments, vertically-scaled scale scores also represent linear transformations of Rasch-based performance estimates (θ). Vertically scaled scores, however, additionally include an extra scaling constant (V_g) that varies across each grade (g). This is given by the equation below:

$$SS_\theta = A \times (\theta - V_g) + B \quad (4)$$

The scaling constants A and B in Equation (4) are derived in the same way as for horizontal scale score systems, except that the scale score for one of the performance standards (e.g., Level II) is fixed only for one of the assessments in the vertical scale (e.g., the STAAR grade 8 mathematics for the STAAR mathematics vertical scale), and the standard deviation is taken across all of the assessments (e.g., all STAAR grades 3–8 mathematics assessments). The A scaling constant is calculated as follows:

$$A = \frac{\sigma_{ss}}{\sigma_{\square}} \quad (5)$$

In Equation (5), σ_{ss} represents the desired standard deviation of the scale across all assessments, while $\sigma_{\square}$ represents the standard deviation of Rasch-based θ values among a sample group. For the STAAR 3–8 vertical scales, for example, the sample group consisted of all students who took a test form with embedded vertical scale items in spring 2012 (vertical scale items are not used to calculate student scores).

The B scaling constant is calculated as follows:

$$B = SS_{Level\_II} - \frac{\sigma_{ss}}{\sigma_\theta} \times \theta_{Level\_II} \quad (6)$$

In Equation (6), SS_{Level_II} represents the desired scale score at the Level II cut for the final assessment in the vertical scale, and θ_{Level_II} represents the approved Level II performance standard (in Rasch units) for the final assessment in the vertical scale. In Equation (6), σ_{ss} represents the desired standard deviation of the scale, while $\sigma_{\square}$ represents the standard deviation of Rasch-based θ values in the sample group.

Equating

As mentioned in the Scaling section of this chapter, Texas uses the RPCM to scale its assessments. Kolen and Brennan (2004) describe three data collection designs that facilitate the equating of IRT scales:

1. **Single group design:** Two or more tests are administered to the same group of examinees, with the test administration order counterbalanced to compensate for time and/or fatigue effects.

2. **Randomly equivalent groups design:** Two or more tests are administered to randomly equivalent groups of examinees.

3. **Common-item nonequivalent groups design:** A common set of test items are 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 campuses. Under the common-item 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 equating items, can be embedded within the test or can stand alone as a separate test. The specific data collection designs and equating methods used in Texas are described below. Refer to Kolen and Brennan (2004) or Petersen, Kolen, and Hoover (1989) for a more detailed explanation of equating designs and methods.

Types of Equating

There are essentially three stages in the item and test development process with regard to equating:

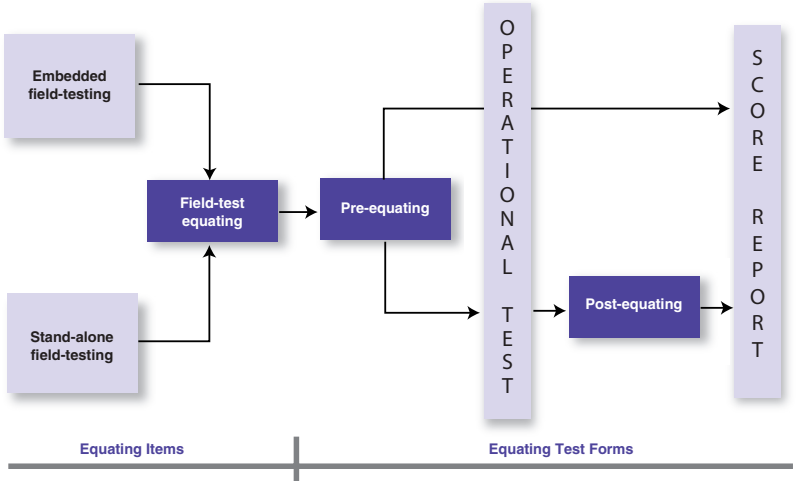
1. Pre-equating test forms that are under construction
2. Post-equating operational test forms after administration
3. Equating field-test items after administration

These three stages allow the established performance standards for the assessments to be maintained on all subsequent test forms. For example, the STAAR EOC performance standards were established by the commissioner of education in April 2012. The STAAR EOC assessments were administered for the first time in spring 2012. Thus, the scale score system for each STAAR EOC assessment was first implemented with the spring 2012 administration. All subsequent test forms for a given STAAR EOC assessment have been or will be equated to this scale score system. STAAR, STAAR Modified, TELPAS reading, TAKS, and TAKS–M are all assessments that require annual equating.

Figure 3.3 illustrates the three stages of the equating process. While field-test equating focuses on equating individual items to the Rasch scale of the item bank, pre-equating and post-equating both focus on equating test forms to maintain score comparability and consistent performance standards.

Pre-equating and post-equating methods take into account differences in the difficulty of test forms.

Figure 3.3. Three Stages of the Equating Process



Pre-equating

The pre-equating process occurs when a newly developed test is linked to the Rasch scale, prior to administration, by a set of items that previously appeared 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. Because the difficulty of the items is established in advance, the difficulty level of newly developed tests can be determined, and the anticipated connection among the raw scores, scale scores, and performance level standards can be identified. Once the anticipated connection among raw scores, scale scores, and performance levels has been established, a raw score to scale score (RSSS) conversion table can be produced that maps each raw score to a scale score and the performance level cut scores.

The pre-equating process involves these steps:

1. Select items that have been equated to the Rasch scale and are available in the item bank.
2. Construct a new test that meets the content specifications.
3. Evaluate the test being constructed using Rasch-based difficulty targets.
4. Develop a RSSS conversion table for the operational test form using the Rasch-based item difficulties.

Pre-equating is conducted for all assessments for which scale scores are reported as part of the test construction process. In many cases, post-equating (described below) is also conducted. However, for some tests post-equating is not conducted and the pre-equated RSSS conversion table is used to assign scale scores. A "pre-equating only" model might be preferred when a small or non-representative sample of students is taking the operational test or when

faster reporting of scores is a priority. For example, for the STAAR EOC mathematics, science, and social studies assessments, pre-equating is used in order to report scale scores faster.

Post-equating

Post-equating uses data from the operational test administration to re-estimate item difficulties and place them onto the Rasch scale for an assessment. The method of estimating the Rasch item difficulties is referred to as calibration. These updated item difficulty estimates are then used to create the RSSS conversion table that is used to report scale scores. Post-equating might be preferred when changes in item presentation (i.e., position, formatting, etc.) or instructional practice have occurred since the time an item was field-tested that might impact the estimated difficulty of the item. Wright (1977) outlines the procedure performed on the common-item set to calculate an equating constant to transform the Rasch difficulty obtained from the current calibration to the Rasch difficulty established by the original test form. This equating constant is defined as:

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

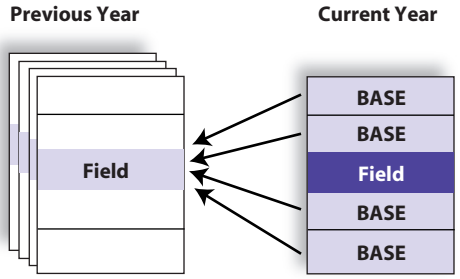
where $t_{a,b}$ is the equating constant, $d_{i,a}$ is the Rasch difficulty of item i on current test a, $d_{i,b}$ is the Rasch difficulty of item i on previous test b, and k is the number of common items. Once the equating constant is obtained, it is applied to all item difficulties, transforming them so 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). Both item difficulty and person proficiency are on the same scale under the Rasch model. Therefore, the resulting scale scores are also comparable from year to year.

The post-equating of assessments in the Texas assessment program uses conventional common-item nonequivalent groups equating procedures, whereby the Rasch item difficulties for a set of equating items are compared with their previously estimated values on the scale of the item bank to derive a post-equating constant. The post-equating constant is calculated as the difference in mean Rasch item difficulty between the set of equating items on the base Rasch scale and on the operational Rasch scale.

The way in which equating items are identified differs among the TAKS, TELPAS, and STAAR programs. For TAKS and TELPAS, the equating item set consists of all the base-test items. The base-test items' Rasch difficulty values from field testing are compared to their values from operational testing to calculate the equating constant. Figure 3.4 illustrates the source of the equating items for the TAKS post-equating design. The

arrows in the figure indicate the transformation of the base-test Rasch item difficulties for the current year onto the Rasch scale for an assessment through the same items' field-test Rasch item difficulties from the previous year. In general, the field-test items are located around the middle of the test form. Therefore, the base-test position of an item is different from its field-test position. Item position effects are assumed to cancel out because half of the items will move to an earlier serial position in the test and half will move to a later serial position in the test.

Figure 3.4. TAKS Common-Item Post-Equating Design

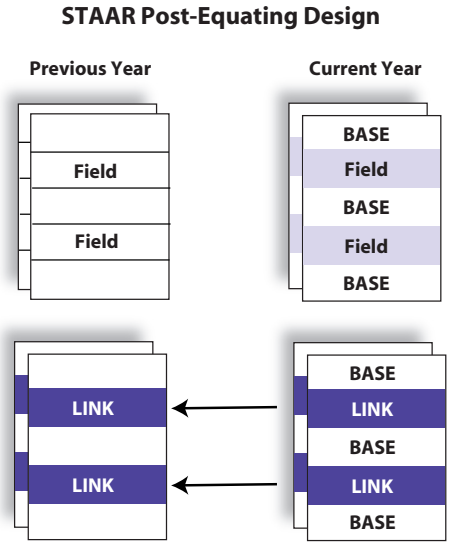


TAKS post-equating is conducted on a sample of students. The requirements for the sample include a minimum sample size of 100,000 students, regional representation similar to the student population, ethnic distribution similar to the student population, and a stable raw score distribution. TELPAS post-equating is conducted using all or nearly all of the student data so no sampling is needed. The initial equating item set for most of the TAKS and TELPAS assessments consists of all the base-test items. However, the stability of the Rasch item difficulty estimates is monitored from field test to base test and, if an item's Rasch item difficulty appears less stable than expected, the item will be excluded from the equating item set during the stability check analysis. Prior to applying the final equating constant, the equating item set is evaluated for content representation and the number of items in the equating set is compared to the base test. The content representation of the equating item set is compared to the base test to verify the test objectives are appropriately represented in the equating item set.

Unlike TAKS, the STAAR base-test items do not comprise the equating item set. Instead, the equating item set is placed in field-test positions on a small number of test forms instead of field-test items. For STAAR, the equating item set consists of a previously designated group of "equating items" that have been evaluated for statistical properties and content alignment. Figure 3.5 illustrates the source of the common-item sets (LINK) for the STAAR post-equating design. The equating items appear in the same item positions on both the current year and previous year test forms. This design minimizes item position effects for the equating items. The number of equating forms required

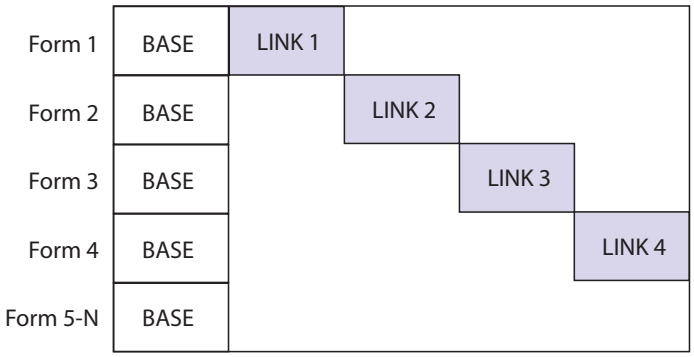
for the equating item sets is dependent on the number of items per form, the number of items needed for the equating items to be content representative of the base test, and the size of the testing population for each STAAR assessment.

Figure 3.5. STAAR Common-Item Post-Equating Design



In order to obtain Rasch estimates for the equating items on the same scale, a concurrent calibration across the equating forms is conducted using an incomplete data matrix (Kolen & Brennan, 2004; Wingersky and Lord, 1984). Figure 3.6 illustrates the incomplete data matrix for the equating forms with the equating item sets (LINK). Forms 1 through 4 represent forms with equating items and forms 5 through N represent forms with field-test items.

Figure 3.6. STAAR Incomplete Data Matrix for Common-Item Sets



STAAR post-equating is conducted on a sample of students. The requirements for the sample include a minimum sample size of 100,000 students, regional representation similar to the student population, ethnic distribution similar to the student population, and a stable raw score distribution. Only the test forms with the equating item sets are used in determining the equating constant that will place the base-test Rasch item

difficulties on the Rasch scale common across administrations for an assessment. However, student data from all test forms are used in estimating the Rasch item difficulties for the base test items. The initial equating item set for most of the STAAR assessments consists of all equating items. However, the stability of the Rasch item difficulty estimates for the equating items is monitored from year to year and if an item's Rasch item difficulty appears less stable than expected, the item will be excluded from the equating item set during the stability check analysis. Prior to applying the final equating constant, the equating item set is evaluated for content representation and the number of items in the equating set is compared to the base test. The content representation of the common-item set is compared to the base test to verify the test objectives (reporting categories) are appropriately represented.

The post-equating procedure involves these steps:

1. Tests are assembled and evaluated using Rasch-based difficulty targets.
2. Data from the test administrations are sampled.
3. Rasch item difficulty 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 equating item set on the scale of the item bank versus the operational scale.
5. The post-equating constant is applied to the Rasch difficulty estimates for the operational test items, and RSSS conversion tables are produced.

The full equating process is independently replicated by multiple psychometricians (from TEA and external vendors) for verification.

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 Rasch scale of the assessment. The STAAR and TELPAS reading assessments, for example, use embedded field-test designs to collect data on field-test items. STAAR Modified, by comparison, used an approach during its initial administration in spring 2012 in which field-test items were immediately analyzed for use as part of a student's operational test score. However, STAAR Modified will use an embedded field-test design for future administrations.

In embedded field test designs, after a newly constructed item has cleared the review process, it is embedded in a test form along with the operational items. The operational items are common across all test forms and count toward an individual student's score, but each field-test item appears on only a small number of test forms (typically one form) and does not count toward students' scores. These forms are then spiraled, meaning that they are packaged in such a way that the test forms are assigned to students randomly. Test forms are

spiraled so that a representative sample of examinees responds to the field-test items. A calibration of the Rasch item difficulties for both the base-test items and the field-test items is conducted. Wright's (1977) common-items equating procedure is then used to transform the Rasch difficulty of the field-test items to the same Rasch scale as the common items, as described below:

1. Obtain Rasch item difficulty estimates for the combination of operational and field-test items.
2. Using the operational items as the common items, calculate an equating constant as the difference in mean Rasch item difficulty between the Rasch item estimates of the common items on the base Rasch scale and the Rasch item difficulty estimates of the common items as estimated with the field-test items.
3. The field test item difficulties are placed on the scale of the item bank by adding the equating constant to each of the field-test Rasch item difficulties.

Because the Rasch scale of the common items had previously been equated to the scale of the item bank, so too are the equated field-test items.

During the initial operational administration of the STAAR Modified assessments in spring 2012, each test form included items that had not been previously field-tested. Statistical characteristics of all the items were examined immediately after the test administration. Only items with sound psychometric properties were used to generate the RSSS conversion table for each STAAR Modified test form. In subsequent operational administrations of the STAAR Modified assessments, the embedded field-test design will be used to collect data on field-test items. The equating procedure described above for embedded field tests will be used to put field-test items on the Rasch scale for an assessment. Refer to chapter 5, "STAAR Modified," for more information on the spring 2012 administration of STAAR Modified.

Matched Sample Comparability Analysis

When the same test is administered in paper and online delivery modes, studies can be conducted to determine whether the use of the same RSSS conversion 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, described in the Sampling section of this chapter, is used to select online and paper student samples wherein each selected online student is matched to a paper student with the same demographic variables and proficiency level on previous test scores. Item statistics, such as item p-values and Rasch item difficulties, are compared between the matched samples. Raw score to scale score conversions are calculated using Rasch scaling as described above. The sampling is then repeated many times. RSSS conversion tables are retained and aggregated across replications, and the mean and the standard deviation of the scale scores are taken at each raw score point to obtain the final RSSS conversion table and the standard errors of linking, respectively. The equivalency of online and paper scale

scores are then evaluated using the standard errors and raw scores as guides. If the two sets of scores are considered not comparable, it might be necessary to use a separate RSSS table for each mode of delivery.

For detailed descriptions of past comparability analyses, refer to the Comparability Studies section of the Technical Report series on TEA's Student Assessment Division website.

Reliability

The concept of reliability is based on the idea that repeated administrations of the same test should generate consistent results. Reliability measures estimate the degree to which a test produces consistent results. In theory, reliable assessment instruments would generate similar results upon multiple administrations to the same student population. Reliability is a critical technical characteristic of any measurement instrument, because unreliable instruments cannot be interpreted in a valid way.

Internal Consistency Estimates

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, those from 0.80 to 0.89 are considered good, and those at 0.90 or above are considered excellent. However, what is considered appropriate might vary in accordance with how assessment results are used. 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 containing a mixture of multiple-choice and constructed-response items.

The KR20 is a mathematical expression of the classical test theory definition of test score reliability that expresses test score 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:

$$P'_{xx} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

where the reliability, P'_{XX}, of test X is a function of the ratio between true score variance (σ^2_T) and observed score variance (σ^2_X), which is further defined as the combination of true score variance and error variance ($\sigma^2_T + \sigma^2_E$). As error variance is reduced, reliability increases (that is, students' observed scores are more reflective of students' true scores or actual proficiencies). KR20 can be represented mathematically as:

$$KR_{20} = \left[\frac{k}{k-1}\right]\left[\frac{\sigma^2_X - \sum_{i=1}^{k} p_i\,(1-p_i)}{\sigma^2_X}\right]$$

where KR_{20} is a lower-bound estimate of the true reliability, k is the number of items in test X, σ^2_X 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 KR_{20} to cases where items are scored polytomously (into more than two categories) and is computed as follows:

$$\alpha = \left[\frac{k}{k-1}\right]\left[1 - \frac{\sum_{i=1}^{k}\sigma^2_i}{\sigma^2_X}\right]$$

where α is a lower-bound estimate of the true reliability, k is the number of items in test X, σ^2_X is the observed score variance of test X, and σ^2_i 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:

$$\textit{Stratified}\ \ \alpha = 1 - \frac{\sum_{j=1}^{c}\sigma^2_{x_j}\left(1-\alpha_j\right)}{\sigma^2_X}$$

where c is the number of item type components, α_j is the estimate of reliability for each item type component, $\sigma^2_{x_j}$ is the observed score variance for each item type component, and σ^2_X is the observed score variance for the total score. For components consisting of multiple-choice and short answer items, a standard coefficient alpha (refer to 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 might require reliability evidence that uses a different approach than the measures described above. Some tests, such as STAAR Alternate, involve teachers observing and evaluating students who are completing appropriate TEKS-based assessment tasks. As part of the process for evaluating the reliability of such tests, TEA must provide evidence that the teacher observation and resulting evaluation of student performance 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 provide ratings of that student performance. These ratings 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 reliability of the test scores.

Measurement Error

Though test scores for Texas assessments are typically highly reliable, each test score does contain 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. An example of a systematic error would be a language barrier that caused a student to answer a question incorrectly that he or she actually knew the answer to. By contrast, random errors are chance occurrences. An example of a random error would be a student guessing the correct answer to a test question. Texas computes the classical standard error of measurement (SEM), the conditional standard error of measurement (CSEM), and classification accuracy for the purpose of estimating the amount of random error in test scores.

Classical Standard Error of Measurement (SEM)

Classical SEM represents the amount of variance in a score resulting from factors other than what the assessment is designed to measure. 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 score 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 using both the standard deviation and the reliability of test scores, as follows:

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

where r is the reliability estimate (for example, a KR_{20}, 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 most likely contains the student's true score. For example, suppose a student achieves a raw score of 50 on a test with an SEM of 3. Placing a one-SEM band around this student's score would result in a raw score range of 47 to 53. 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 another way, if this student took the test 100 times, he or she would be expected to achieve a raw score between 47 and 53 about 68 times out of the 100.

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 (refer to, 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)

CSEM also represents the amount of variance in a score resulting from factors other than what the assessment is designed to measure, but it provides an estimate conditional on proficiency. In other words, the CSEM provides a measurement error estimate at each score point on an assessment. 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 thus the scores are most reliable, in the middle of the score distribution.

IRT methods for estimating score-level CSEM are used because test- and item-level difficulties for the STAAR, STAAR Modified, TELPAS reading, TAKS, and TAKS–M tests are calibrated using the Rasch measurement model, described in detail in the Scaling section of this chapter. By using CSEMs that 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 performance levels. For the vast majority of students, these classifications are accurate reflections of their performance. However, because it is known that all test scores contain some amount of error, some students might be misclassified. It is important to understand the expected degree of misclassification. To this end TEA and Pearson conduct an analysis of the accuracy of student classifications into performance levels based on results of tests for which performance standards have been previously 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 the lowest performance level (Level I for STAAR and STAAR Modified, Beginning for TELPAS reading , and Did Not Meet Standard for TAKS and TAKS–M) and the upper bound for highest performance level (Level III for STAAR and STAAR Modified, Advanced High for TELPAS reading, and Commended Performance for TAKS and TAKS–M) 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 Level II performance level for STAAR, STAAR Modified, the Intermediate and Advanced performance levels for TELPAS reading, and Met Standard for TAKS and TAKS–M 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 level by *n* performance level" classification table. The sum of the diagonal entries represents the classification accuracy for the test. An example of a classification accuracy table for the STAAR Level II: Satisfactory Academic Performance standard is presented in Table 3.2.

Table 3.2. Classification Accuracy for STAAR Level II

		STAAR Classification	
		At or above Level II	**Below Level II**
True Classification	At or above Level II	Proportion of accurate "At or above Level II" classifications	Proportion of inaccurate "At or above Level II" classifications
	Below Level II	Proportion of inaccurate "Below Level II" classifications	Proportion of accurate "Below Level II" classifications

Validity

In the Texas assessment program, validity refers to the extent to which test scores help educators make appropriate inferences about student performance. The concepts described here are not types of validity, but types of validity evidence. Evidence to support validity can be based on and organized into these five categories (described in detail below): test content, response processes, internal structure, relations to other variables, and consequences of testing (AERA/APA/NCME, 1999; Schafter, Wang, & Wang, 2009). Assessment validation is a matter of degree and is an ongoing process. Furthermore, evidence that supports the validity of a test is evidence regarding 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, the STAAR test scores are designed to help make inferences about students' knowledge and understanding of the TEKS. Therefore, evidence supporting the content validity of the STAAR assessments maps the test content to the TEKS. Validity evidence supporting Texas' test content comes from the established test development process and the judgments of subject matter experts' about the relationship between the items and the test construct.

The test development process starts with a review of the TEKS by Texas educators. The educators then work with TEA to define the readiness and supporting standards in the TEKS and help determine how each standard would best be assessed. A test blueprint is developed with educator input, which maps the items to the reporting categories they are intended to represent. Items are then developed based on the test blueprint. Below is a list of steps in the test development process that are followed each year to support the validity of test content in Texas:

- write items based on test reporting categories and item guidelines
- review items on more than one occasion for appropriateness of item content and identification of item bias
- field-test items
- review field-test data
- build tests to pre-defined criteria
- have university-level experts review high school assessments for accuracy of the advanced content

A more comprehensive description of the test development process is available in chapter 2, "Building a High-Quality Assessment System."

Evidence Based on Response Processes

Response processes refer to the cognitive behaviors required to respond to a test item. Texas collects evidence that supports 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 STAAR writing test includes a writing component in addition to multiple-choice questions because requiring students to answer multiple-choice questions as well as to respond to writing prompts provides the most appropriate manner for students to demonstrate their ability to write. Student response processes on Texas' assessments differ by both item type and administration mode.

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, 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 submitted 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 comparability studies using the methodology described in the Equating section of this chapter to

evaluate the comparability of online and paper test score interpretations. Score adjustments might be made when evidence suggests that student 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 that can be evaluated in terms of the internal consistency estimates of reliability described above in the Reliability section of this chapter. Internal consistency estimates are evaluated for Texas assessments for reported student groups, including all students, female, male, African American, Hispanic, and white students. Estimates are made for the full assessment as well as for each reporting category within a content area.

Validity studies have also been conducted to evaluate the structural composition of assessments, such as the comparability between two language versions of the same test. For example, a study conducted on the structural equivalence of transadapted tests (Davies, O'Malley & Wu, 2007) provided evidence that the Spanish and English versions of Texas assessments were measuring the same construct, which supports 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 another measure, sometimes called criterion-related validity. The relationship can be concurrent, meaning that performance on two measures taken at the same time are correlated, or the relationship can be predictive, meaning that the current performance on one measure predicts performance on a future measure. The relationship can also be convergent, meaning performance on two measures that are meant to assess the same or similar construct should be strongly correlated, or the relationship can be divergent, meaning performance on two measures that are meant to assess unrelated constructed should have a weak correlation or even no correlation.

A large number of research studies were conducted to evaluate the relationship between performance on the STAAR assessments and performance on other related tests or criteria. The studies included:

- The STAAR-to-TAKS comparison studies, which link performance on the STAAR assessments to performance on TAKS assessments (for example, the STAAR grade 7 mathematics to the TAKS grade 7 mathematics)
- The STAAR linking studies, which link performance on the STAAR assessments across grade levels or courses in the same content areas (for example, grade 4 reading to grade 5 reading, and Algebra I to Algebra II)
- The STAAR inter-correlation estimates, which evaluate the strength of relationship (or lack thereof) among scores on the STAAR assessments across different content areas (for example, grade 4 mathematics to grade 4 reading, and biology to world geography)

- Grade correlation studies, which link performance on the STAAR EOC assessments to course grades
- External validity studies, which link performance on the STAAR assessments to external measures (specifically: SAT, ACT, THEA, ACCUPLACER, Explore, and Readistep)
- College students taking the STAAR studies, which link performance on the STAAR EOC assessments to college course grades

For detailed descriptions and results of some of these studies, refer to the STAAR Resources page of 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, the STAAR assessments 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. TEA, for example, has developed and implemented a plan to formally document the evidence of the consequential validity of the STAAR Modified program. Surveys asking about the intended and unintended consequences resulting from the STAAR Modified assessments were administered to the standard-setting committee. Once analyzed, results from the STAAR Modified consequential validity surveys will be reported and used to help promote the continuous improvement of the STAAR Modified program. Collection of consequential validity will continue in upcoming years for the STAAR assessments.

Measures of Student Progress

Measures of student progress compare and describe current student performance and previous student performance. Student progress information provides essential context to understanding students' current performance. For example, consider a student that achieves Level II: Satisfactory Academic Performance on a STAAR assessment. The interpretation of Level II performance would depend on the performance the student achieved in the previous year. If the student achieved Level I: Unsatisfactory Academic Performance in the previous year, then the student made notable progress this year by advancing a performance level. However, if the student had achieved Level III: Advanced Academic Performance in the previous year, then the interpretation of Level II this year would be quite different.

Student progress information can also provide insight into future performance goals. For example, a worthy goal would be for all students to achieve at or above Level II on the STAAR assessments. Student progress measures provide information about where a student performed previously as well as currently. This information can then be used to set reasonable, individual student goals for the future. For example, for those students who have not yet reached Level II, progress measures can be used to evaluate whether a student is "on track" to meet Level II in a future year.

For students who took TAKS in 2011–2012, progress information was provided using the Texas Projection Measure (TPM). For more information on the TPM, refer to Procedures for Developing the Texas Projection Measure Equations on TEA's Student Assessment Division website.

Types of Student Progress Measures

Given the importance of progress information, student progress measures will be calculated and reported for STAAR. Several types of progress measures can be used and each was considered for use with STAAR.

- Regression models—Regression models utilize past and present student performance to statistically predict future performance. These models are commonly used to predict whether a student will achieve a higher performance level, such as Level II, in the future. This type of model was used to report progress for students taking TAKS and TAKS–M grade 10 assessments in 2012. For more information on TAKS progress measures see the Student Progress Measures resources page on TEA's website.
- Growth percentile models—Like regression models, growth percentile models statistically predict future student performance and achievement. These models also provide information regarding one student's growth as compared with the growth of the student's peers.
- Growth to proficiency models—Growth to proficiency models do not predict future performance with statistical procedures. Rather, these models consider students' current achievement and a future goal (for example, achievement of Level II) and measure the annual progress needed in order to achieve the goal.
- Value/transition tables—Similar to growth to proficiency models, value/transition tables establish annual progress goals to reach desired performance in a future year. This is done by subdividing performance levels. For example, Level I could be further divided into smaller categories, and then progress can be tracked through these smaller categories.

These measures differ in the types of information they use, the complexity of their calculations, the feedback they provide, and the ease with which they can be explained. These factors are all important to consider when selecting a model for measuring student progress.

Each of the student progress measures require at least two years of student data. Because STAAR was administered for the first time in 2011–2012, student progress information is not yet available. However, student progress will be measured and available after the 2012–2013 STAAR administrations.

Sampling

Sampling is a procedure that is used to select a small number of observations that are representative of the whole population. In this case, sampling involves the selection of a set of Texas students that is able to represent the entire body of Texas students. The results from well-drawn samples allow TEA to estimate characteristics of the larger Texas student population.

Key Concepts of Sampling

Target Population

A target population is the complete collection of objects of interest (for example, students) (Lohr, 1999). This is the set of students to which the results generalize. For example, for a study with the goal of understanding how grade 3 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 before sampling takes place.

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 allow 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 entire 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 example, if 20 campuses are randomly chosen from a list of all campuses in the state, then the campus is the sampling unit.

An observation unit is the unit on which data are actually collected. An observation unit might or might not be the same as the sampling unit. For example, a study designed to estimate the number of computers per campus in the entire state might involve requesting each of 20 randomly selected campuses to report the number of computers it has. In this case, the campus is both the sampling unit and the observation unit. By comparison, if a study is designed to estimate the number of students per computer in the entire state, and each of the same 20 sampled campuses is requested to report data on how many students per computer there are at that campus, then even though

the sampling unit is still the campus (because 20 campuses were picked), the observation unit is the student (because the data being collected are in regard to students).

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 very large.
- Accessibility—There are situations where collecting data on every member of the target population is not feasible.
- Cost—It is less costly to obtain data for a carefully selected subset of a population than to collect the same data for the entire population.
- Time—Using the tool of sampling to study the target population is less time-consuming. Sampling might be needed when the speed of the analysis is important.
- Burden—Sampling minimizes the participation requirements for campus and district therefore reducing testing burden.

Sampling Designs

The Texas assessment program uses the following types of sampling designs to collect data for the purpose of field testing, audits, and empirical research studies (such as linking studies, linguistic accommodations studies and comparability studies).

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. For example, if the student is the sampling unit, probability sampling would require an accurate list of all the students in the target population. Random selection, meaning that each sampling unit has the same probability of being selected from the list of sampling units, is a key element of probability sampling. The three major types of probability sampling designs are:

- 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.

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 clarify the distinction between the two sampling methods, consider first simple random sampling *with* replacement and then simple random sampling without replacement. First, suppose that a simple random sample with replacement of size *n* is obtained from a population of *N*. In this case, when a sampling unit is randomly selected, that unit remains eligible to be selected again. In other words, after the sample is "picked," it is also "put back" and can be selected again. When sampling with replacement, a sampling unit might be selected multiple times and have its data duplicated in the resulting samples of size *n*.

By comparison, suppose that a simple random sample *without* replacement of size *n* is obtained from a population of *N*. In this case, once a sampling unit is chosen, it is ineligible to be selected again. In other words, after the sample is "picked," it is not "put back." 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.

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 it is impractical or impossible to collect a complete list of sampling units. 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.

Resampling And Replication Methods: Bootstrap

Resampling and replication methods, such as bootstrapping, treat the sample like a population. These methods repeatedly take pseudo-samples from samples to estimate the parameters of distributions. 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 that compare online and paper versions of a test form.