Chapter 3 Standard Technical Processes

Overview

The Standards for Educational and Psychological Testing by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (2014) provide a set of guidelines for evaluating the quality of testing practices. By using these standards to guide test development, the Texas Education Agency (TEA) is confident that Texas assessments are technically defensible and appropriate for the purposes for which they are used.

The objective 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. In-depth discussions of the specific processes are covered 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 directly relate levels of test performance 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, examine the degree to which the item appropriately distinguishes between students of different proficiency levels, and assess the item for potential bias.

Scaling — Scaling is a process that transforms test scores from one set of numbers to another so that they are easier to interpret.

Equating — Equating is used in conjunction with scaling to place scores from different test forms on a common scale, thereby making test scores comparable across test administrations.

Reliability — Reliability indicates the precision of test scores, which also reflects the consistency of test results across testing conditions.

Validity — Validity refers to the extent to which test scores can be interpreted as indicators of what the test 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 representative of a population. For the State of Texas Assessments of Academic Readiness (STAAR®) program, sampling involves the selection of a set of Texas students' representative of the entire body of Texas students. The results from well-drawn samples allow TEA to estimate characteristics of the 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 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. Performance 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 performance levels. In 2012, the STAAR standard-setting process established cut scores on each assessment, creating the following performance levels: Level I: Unsatisfactory Academic Performance, Level II: Satisfactory Academic Performance (this included a phase-in standard and a final standard), and Level III: Advanced Academic Performance. These performance level labels for STAAR, including STAAR Spanish, were revised in the 2016–2017 school year to Did Not Meet Grade Level, Approaches Grade Level, Meets Grade Level, and Masters Grade Level. Standards were set for STAAR Alternate 2 in spring 2015 to establish the following performance levels: Level I: Developing Academic Performance, Level II: Satisfactory Academic Performance, and Level III: Accomplished Academic Performance. The most recent standard-setting for the Texas English Language Proficiency Assessment System (TELPAS) was conducted in 2018 to create proficiency level cuts (Beginning, Intermediate, Advanced, and Advanced High) for the reading, listening, and speaking domains. In 2019, standard setting was conducted for TELPAS Alternate to establish the following five proficiency levels: Awareness, Imitation, Early Independence, Developing Independence, and Basic Fluency for the reading, listening, and speaking domains.

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

Item Analyses

Several statistical analyses are conducted using the student response data collected for each item. Item analyses are conducted annually for the purpose of reviewing the quality of newly field-tested items to help determine which items might be included as operational items in future test administrations. The Technical Details and Procedures section of this chapter provides information about the various item statistics that are generated as part of the item analyses.

Scaling

Scaling is the process of associating numbers with a characteristic of interest such as temperature, time, speed, etc. Multiple scales can be used to provide information about measurable quantities for a single characteristic of interest. 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. Similarly, test scores can also be reported using more than one scale, as explained in the following paragraphs.

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. In general, raw scores from different test forms are not comparable, as the

following hypothetical example helps illustrate. Suppose there are two forms of an assessment that are not equally difficult. In this example, Form A is harder than Form B. Suppose also that a student (Student A) takes Form A and earns a raw score of 34 out of 50, while another student (Student B) takes Form B and also earns a raw score of 34 out of 50. Here, Student A's performance reflects greater achievement than Student B's performance even though both students receive the same raw score. When a new form of an assessment is administered, the questions on the new form are generally different from those on older forms. Despite the fact that different test forms target the same knowledge and skills, some forms will be slightly easier or slightly more difficult than others. As a result, in most cases, student performance cannot directly be compared across testing administrations using raw scores. To facilitate comparisons, raw scores from different test forms are transformed into scale scores on a common scale.

When scores from 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 allow for direct comparisons of student performance across separate test forms and different test administrations. A 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 test forms. Scaling is the process of creating these scale scores.

Horizontal scale scores are used to describe student performance within a given grade level and content area. Horizontal scales are created separately for each grade level and content area, making no reference to potential similarities in content across grade levels. 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 assessments that measure student performance in the same content area at different grade levels onto a common scale, thereby facilitating inferences about changes in students' scores across grades.

For the STAAR program, vertical scales have been developed for the following grade levels and content areas: STAAR grades 3–8 mathematics (a single scale for English and Spanish assessments), STAAR grades 3–8 English reading, and STAAR grades 3–5 Spanish reading.

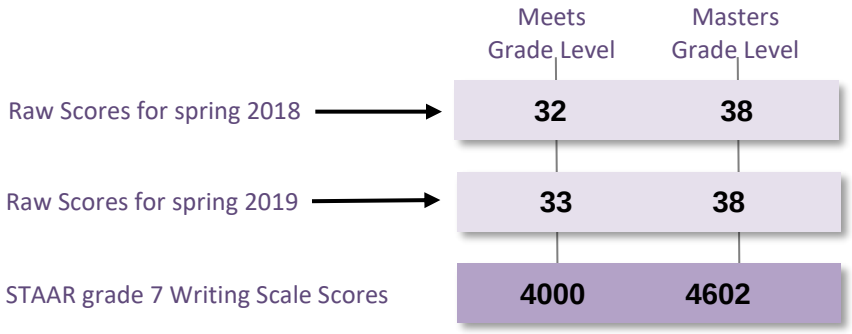
STAAR grades 4 and 7 writing, grades 5 and 8 science, grade 8 social studies, end-of-course (EOC) assessments, STAAR Alternate 2, TELPAS, and TELPAS Alternate assessments are reported on horizontal scales.

Equating

Used in conjunction with the scaling process, equating is the statistical process that takes into account the differences in difficulty across test forms and administrations and allows scores to be placed onto a common scale. Through the equating process, TEA enables the comparison of scale scores across test forms and test administrations.

Due to the impact of COVID-19 and the subsequent cancellation of STAAR testing, no operational STAAR equating was conducted in the spring and summer of 2020. The following 2018–2019 example illustrates the purpose of equating. Figure 3.1 provides an example of the relationship between raw scores and scale scores relative to the performance standards (or cut scores) on two STAAR grade 7 writing test forms that vary slightly in difficulty. The scale scores required for Meets Grade Level and Masters Grade Level remain the same across both test forms: 4000 is the cut score for Meets Grade Level, and 4602 is the cut score for Masters Grade Level. The raw scores required to achieve these performance levels on the spring 2018 test were 32 and 38, respectively. The raw scores required to achieve these performance levels on the spring 2019 test were 33 and 38, respectively. At first glance, it might appear that more was expected of students for them to achieve these performance standards in 2019 than in 2018, but this would be a misinterpretation. Rather, the set of test questions on the 2019 test were slightly less challenging than the set of test questions on the 2018 test. So, a student who scored a 32 on the more difficult 2018 test would have been expected to achieve a score of 33 on the easier 2019 test.

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



Equating is done to ensure equitability. By accounting for the 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 assessment should generate consistent results. Reliability is a critical technical characteristic of any measurement instrument because unreliable scores cannot be interpreted in a valid way. 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, obtaining these types of reliability estimates is burdensome on schools and students. Therefore, reliability estimation methods that require only one test administration have been developed and are commonly used for large-scale assessments, including STAAR, STAAR Alternate 2, TELPAS, and TELPAS Alternate.

Validity

The results of STAAR, including STAAR Spanish and STAAR Alternate 2, are used to make inferences about how well students know and understand the Texas Essential Knowledge and Skills (TEKS) curriculum. Similarly, TELPAS and TELPAS Alternate test results are used to make inferences regarding English language acquisition aligned with the English Language Proficiency Standards (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 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. Beginning in the 2016–2017 school year, each STAAR assessment, including STAAR Spanish, has the following four performance levels: Did Not Meet Grade Level, Approaches Grade Level, Meets Grade Level, and Masters Grade Level. This information is useful in describing students' current knowledge and skills. However, the overall description of student achievement can be enhanced by providing student progress measures that convey information about how performance in the current year compares to performance in the prior year. Individual student progress is compared to progress targets so that progress for STAAR can be classified as *Limited*, *Expected*, or *Accelerated*.

The STAAR Alternate 2 progress measure is based on a comparison of a student's test score last year with his or her score this year. For STAAR Alternate 2, students' progress is classified as either *Did Not Meet*, *Met*, or *Exceeded* the progress target. Due to the cancellation of STAAR and STAAR Alternate 2 testing in the spring and summer of 2020, there is no progress measure in 2020.

Sampling

Sampling plays a critical role in the annual test-development and research 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 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 the burden on campuses and districts. In particular, sampling is used in the Texas assessment program for field testing, audits, and research studies (e.g., linking studies, linguistic accommodations studies, cognitive labs, and comparability studies).

Results from field testing are used to evaluate statistical properties of newly developed test items that have not yet been used on an operational test form. Audits allow for the collection of information from school districts that can be used to evaluate training, administration, and scoring of the assessments. In general, research studies involve assessing a sample of students under various testing conditions in order to collect evidence to support the technical quality of and make improvements to the Texas assessment program.

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 gender, ethnicity, and campus size.

Technical Details and Procedures

Performance Standards

Performance standards directly relate levels of test performance 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 assessments (including STAAR Spanish) have three cut scores that identify four performance levels:

- Did Not Meet Grade Level
- Approaches Grade Level
- Meets Grade Level
- Masters Grade Level

The STAAR Alternate 2 assessments have two cut scores that identify three performance levels:

- Level I: Developing Academic Performance
- Level II: Satisfactory Academic Performance

- Level III: Accomplished Academic Performance

The TELPAS assessments have three cut scores that identify four English proficiency levels:

- Beginning
- Intermediate
- Advanced
- Advanced High

The TELPAS Alternate assessments have four cut scores that identify five English proficiency levels:

- Awareness
- Imitation
- Early Independence
- Developing Independence
- Basic Fluency

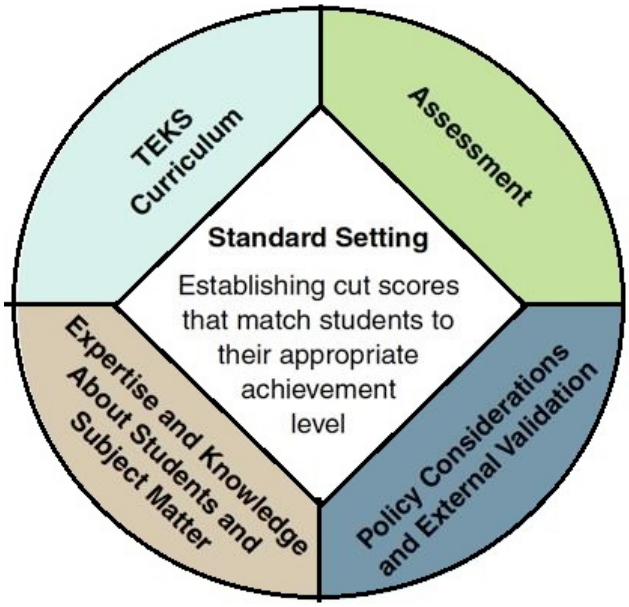
Standard-setting is the process of establishing cut scores that define the performance levels on an assessment. This section describes the standard-setting framework and process for the STAAR, STAAR Alternate 2, TELPAS, and TELPAS Alternate programs.

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 levels (Level I: Unsatisfactory Academic Performance, Level II: Satisfactory Academic Performance [this included a phase-in standard and a final recommended standard], Level III: Advanced Academic Performance). In the 2016–2017 school year, the performance level labels were changed to Did Not Meet Grade Level, Approaches Grade Level, Meets Grade Level, and Masters Grade Level for better communication of testing results to educators, parents, and students. This was a part of the effort to report students' test results with a family-friendly STAAR Report Card (for more information, visit The STAAR Report Card on the TEA website).

Standard-setting for STAAR involved a process of combining policy considerations, 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 STAAR is described below.

- **TEKS Curriculum Standards —** The TEKS curriculum standards are designed to reflect the knowledge and skills students need to succeed in their postsecondary endeavors and to compete globally. The standards 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, including STAAR Spanish, has been developed to measure 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 corresponding STAAR assessed curriculum document and the STAAR test blueprint.
- **Policy Considerations and External Validation —** Research studies that empirically correlated 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 item mapping with external data (Ferrara, Lewis, Mercado, D'Brot, Barth, & Egan, 2011; Phillips, 2012), was used to make recommendations for the performance standards.

Using this standard-setting framework, TEA defined and implemented a nine-step process to establish the performance standards for the STAAR assessments. Table 3.1 provides descriptions of each of the steps in the STAAR standard-setting process.

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

Standard-Setting Step	Description	Timeline
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.	Studies began in spring 2009 and are ongoing.
2. Develop performance labels and policy definitions.	Committees recommended performance categories, performance category labels, and general policy definitions for each performance category.	September 2010
3. Develop grade/course specific performance level descriptors (PLDs).	Committees consisting primarily of educators developed PLDs as an aligned system, describing a reasonable progression of skills within each content area (mathematics, English, science, and social studies).	November 2011
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. The STAAR EOC recommendations served as the foundation for decisions made regarding STAAR 3–8 and STAAR Alternate 2 assessments.	February 1-2, 2012
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.	Mathematics and English: February 22–24, 2012 Science and Social Studies: February 29–March 2, 2012
6. Review performance standards for reasonableness.	TEA reviewed the cut-score recommendations across content areas.	March 2012
7. Approve performance standards.	The Commissioner of Education approved performance standards.	April 2012

8. Implement performance standards.	Once established, performance standards were reported to students for the spring 2012 administration with phase-in standards applied.	May 2012
9. Review performance standards.	Performance standards are reviewed at least once every three years.*	Fall 2014

*In June 2013, the 83rd Texas Legislature enacted House Bill (HB) 5, which removed the requirement to convene standards review panels. However, TEA and the Commissioner of Education review statewide performance relative to the standards after each administration.

More details about each step in the STAAR standard-setting process are given in the "STAAR Standard-Setting Technical Report" available on the STAAR Performance Standards webpage of TEA's Student Assessment Division website.

STANDARD-SETTING FOR STAAR ALTERNATE 2

Standards were set for STAAR Alternate 2 in spring 2015. Standard-setting for STAAR Alternate 2 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 state assessments aligns with student performance on other assessments. TEA used an evidence-based standard-setting approach (O'Malley, Keng, & Miles, 2012) for the STAAR Alternate 2 program. Using this approach, TEA defined and implemented a nine-step process to establish performance standards for all the STAAR Alternate 2 grades 3–8 and EOC assessments. Table 3.2 provides high-level descriptions and timelines for the steps in the STAAR Alternate 2 standard-setting process.

Table 3.2 Overview of the STAAR Alternate 2 Standard-Setting Process

Standard-Setting Step	Description	Timeline
1. Conduct empirical studies.	Analyses of pilot data as well as analysis of score distributions were conducted.	Spring 2015
2. Develop performance labels and policy definitions.	A committee was convened jointly by TEA and the Texas Higher Education Coordinating Board (THECB) to recommend performance categories, performance category labels, and general policy definitions for each performance category. The STAAR Alternate 2 performance labels and policy definitions were adapted from those created for STAAR by the committee.	January 2015
3. Develop reasonable ranges for performance standards.	The same committee (from step 2) considered the policy implications of performance standards, empirical study results, and content recommendations to identify reasonable ranges for performance standards (neighborhoods).	January 2015
4. Develop grade and course PLDs.	TEA and Pearson created draft-specific PLDs, and educator committees reviewed and edited the PLDs. A goal of the development and review of the specific PLDs was to create an aligned system describing a reasonable progression of skills within each subject area (mathematics, reading, science, and social studies).	January 2015

5. Convene standard-setting committees.	Committees consisting of general education and special education experts with experience in grades 3–12 used performance labels, policy definitions, specific PLDs, and predetermined ranges within which to recommend cut scores for each STAAR Alternate 2 assessment. These committees also provided comments to assist TEA with finalizing the specific PLDs.	April 2015
6. Review performance standards for reasonableness.	TEA reviewed the recommendations across subject areas.	April 2015
7. Approve performance standards.	The Commissioner of Education approved the STAAR Alternate 2 performance standards.	April 2015
8. Implement performance standards.	Once established, performance standards were reported to students for the spring 2015 administration.	May 2015
9. Review performance standards.	Performance standards are reviewed at least once every three years.**	If applicable

**In June 2013, the 83rd Texas Legislature enacted HB 5, which removed the requirement to convene standards review panels. However, TEA and the Commissioner of Education review statewide performance relative to the standards after each administration.

More detailed information about the standard-setting process is provided in the STAAR Alternate 2 Standard-Setting Technical Report available on the STAAR Alternate 2 Resources page of TEA's Student Assessment Division website.

STANDARD-SETTING FOR TELPAS

TELPAS grades 2–12 reading proficiency level standards were established in 2008 when the Texas Assessment of Knowledge and Skills (TAKS) was the academic assessment in Texas. A two-phase approach was used to set the 2008 proficiency level standards. 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 or 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 first-phase recommendations were based, and impact data.

The move from TAKS to STAAR in 2011–2012 made it necessary to review the original TELPAS reading proficiency level standards so that performance on TELPAS could still be a meaningful indicator of the level of English language proficiency required to access the language in STAAR assessments. In August 2013, a standards review was conducted with committees of educators. TEA used an evidence-based standard-setting approach to determine the cut scores for the four proficiency level categories. As with STAAR standard-setting, the item mapping with external data method (Ferrara, Lewis, Mercado, D'Brot, Barth, & Egan, 2011; Phillips, 2012) was used for TELPAS, along with validity study information, to recommend the performance standards. The Commissioner of Education approved the new performance standards, which were first implemented during the 2014 spring administration of TELPAS reading.

The change to the TELPAS reading test design in spring 2018, in addition to the first-time administration of an online test for the listening and speaking domains, required establishing new cut scores for TELPAS proficiency levels. A test-centered, criterion-referenced method was used to guide panelists as they determined their proficiency level cut score recommendations. The applied method was a hybrid of the Angoff method (Angoff, 1971) and Extended Modified Yes/No Angoff method (Davis & Moyer, 2015; Plake, Ferdous, Impara, & Budkendahl, 2005). These new performance standards were approved by the Commissioner of Education in summer 2018 and applied for the first time to the scores from the spring 2018 TELPAS administration. More detailed information about the standard-setting process is available in the TELPAS Standard-Setting Technical Report on TEA's Student Assessment Division website.

STANDARD-SETTING FOR TELPAS ALTERNATE

In 2019, student proficiency for each language domain (listening, speaking, reading, and writing) on the TELPAS Alternate assessment was classified into one of five English language proficiency levels, or stages of increasing proficiency in English. The five levels are Awareness, Imitation, Early Independence, Developing Independence, and Basic Fluency.

The cut scores recommended by the standard-setting committees represent the proficiency students are expected to demonstrate to be classified into each proficiency level. To establish the proficiency levels for each domain, a test-centered, criterion-referenced method was used to guide the panelists. The procedure implemented was a hybrid of the Extended Modified Yes/No Angoff method (Davis & Moyer, 2015; Plake, Ferdous, Impara, & Buckendahl, 2005). The hybrid standard-setting procedure is a systematic method that combines various considerations into the process of recommending cut scores for the different proficiency levels.

The following steps were used for the TELPAS Alternate standard-setting process:

- **Pre-meeting development —** In anticipation of the standard-setting meetings, various tasks were completed, including the development of alternate proficiency level descriptors (PLDs) by teacher committees, draft borderline descriptions for each domain assessed, the development of materials for the panelists, preparation of the Pearson Standard Setting website for panelists and facilitators, presentation materials for the facilitators, and development of data analysis sources and procedures.
- **Standard-setting meetings —** Committees of panelists referenced the domain-specific borderline descriptions to make recommendations for cut scores that define the different proficiency levels for each assessment.
- **Composite score review —** The rules to determine the TELPAS Alternate composite score were established using the domain scores for proficiency level.

- **Reasonableness review —** TEA conducted a reasonableness review of the TELPAS Alternate cut score recommendations.

Item Analyses

Several statistical analyses, based on both classical test theory and item response theory (i.e., the Rasch measurement model), 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 to help determine which items may be included as operational test items in a future test administration.

Statistics generated for each item include p-value, point-biserial correlation, Rasch item difficulty, Rasch fit statistic, and response/score point distribution. An analysis of group differences in performance is also conducted. The following sections provide descriptions of each statistic.

P-Value

The p-value indicates the proportion 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. An item with a high p-value, such as 0.90 (meaning that 90 percent of students correctly answered the item), is a relatively easy item. An item with a low p-value, such as 0.30 (meaning that only 30 percent of students correctly answered the item), is a relatively difficult item.

Point-Biserial Correlation

The point-biserial correlation describes the relationship between a student's performance on a multiple-choice or gridded-response item (scored correct or incorrect) and performance on the assessment as a whole. A high point-biserial correlation indicates that students who answered the item correctly tended to score higher on the entire test than those who missed the item. In general, point-biserial correlations less than 0.20 indicate a potentially weaker-than-desired relationship.

Note that the point-biserial correlation may be weak on items with very high or very low p-values. For example, if nearly all students get an item correct (or incorrect), that item does not provide much useful information for distinguishing between students with higher performance and students with lower performance on the entire test.

Rasch Item Difficulty

The Rasch item difficulty estimate is another indicator of item difficulty. In contrast to p-values, which are influenced by the ability level of the students who took the item, Rasch item difficulties can be compared across test forms and across different samples of students taking an item across test administrations. Items with low Rasch item difficulty values (e.g., –1.5) are relatively easy, and items with higher values (e.g., +1.5) are relatively difficult.

RASCH FIT

The Rasch fit statistic indicates the extent to which student performance on a multiple-choice or gridded-response item is similar to what would be expected under the Rasch measurement model. Specifically, items with good Rasch fit have relatively few unexpected 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 fits the Rasch model poorly.

RESPONSE/SCORE POINT DISTRIBUTION

The response/score point distribution represents the percentage of students responding to each of the answer choices (i.e., A, B, C, or D) for a multiple-choice item, the percentage of students who responded correctly or incorrectly for a gridded-response item, or the percentage of students who received each of the score points for a written composition prompt (i.e., 0, 1, 2, 3, and 4). Response/score point distributions are provided for the entire group of students and for various demographic groups (e.g., gender and ethnicity for STAAR) or for proficiency level groups (e.g., Beginning, Intermediate, Advanced, and Advanced High for TELPAS).

GROUP DIFFERENCE ANALYSIS

Statistics from a group difference analysis provide information about how different student groups (e.g., male, female, African American, Hispanic, or white students) performed on an item. Such analyses help identify items on which a group of students performed unexpectedly well or poorly. This is referred to as differential item functioning (DIF). Two statistical indicators of DIF are used in the Texas assessment program: the Mantel-Haenszel (MH) alpha and the ABC DIF classification (also known as the ETS DIF classification; Petersen, 1987; Zieky, 1993).

MANTEL-HAENSZEL ALPHA

To calculate Mantel-Haenszel alpha, students are first divided into categories of similar proficiency. An odds ratio is calculated for each of those proficiency categories, where the odds ratio equals the odds of answering correctly for the designated reference group (e.g., males) divided by the odds of answering correctly for the focal group (e.g., females). These odds ratios are combined across proficiency categories to obtain a common odds ratio, known as the MH alpha. If the value of MH alpha is 1, students of similar proficiency, regardless of group membership (e.g., males or females), are equally likely to answer the item correctly. If the MH alpha value is statistically significantly greater than 1, the chance of success on the item is better for the reference group (e.g., males) than for the focal group (e.g., females) when comparing students of similar proficiency. Statistically, a MH alpha value significantly less than 1 indicates the item is easier for the focal group compared to similarly proficient students in the reference group.

ABC DIF CLASSIFICATION

The ABC DIF classification is based on MH alpha, but it considers 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 DIF, “B” means moderate DIF, and “C” means large DIF (refer to Zieky, 1993, for more information). Plus and minus signs (+/–) indicate the direction of DIF. A plus sign indicates that the item is unexpectedly easy for the focal group (e.g., females), and a minus sign indicates that the item is unexpectedly easy for the reference group (e.g., males). The ABC DIF classification is currently used as the DIF indicator for items on the STAAR, STAAR Alternate 2, TELPAS, and TELPAS Alternate assessments.

USE OF DIF ANALYSIS RESULTS

It should be noted that DIF analyses merely serve to identify test items that have unusual statistical characteristics related to student group performance. The DIF analyses alone do not prove that specific items are biased. Such judgments 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, STAAR Alternate 2, TELPAS, and TELPAS Alternate assessments: the raw score scale, the Rasch scale, and the reporting scale.

- The raw score scale is defined as the number of items answered correctly regardless of difficulty and includes weighting of written compositions, if applicable.
- The Rasch scale is a transformation of the raw scores onto a scale that considers the difficulty of the items and is comparable across different test forms and test administrations.
- The reporting scale is a linear transformation of the Rasch scale, through scaling constants, onto a user-friendly scale. Because the transformation is linear, the reporting scale also considers the difficulty of the items. The reported scale scores are comparable and maintain performance standards across test forms and test administrations.

The following sections detail the scaling process in terms of establishing the Rasch scale and transforming the scores on 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 locate students on a scale according to the amount of knowledge each student possesses. By comparison, stimulus-centered approaches 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 approaches. 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. Scores on the Rasch scale are then transformed to more user-friendly scale scores to facilitate interpretation.

RASCH PARTIAL-CREDIT MODEL (RPCM)

Test items (whether multiple-choice, gridded-response, or written composition) for all 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 & Stone (1979), Masters (1982), Wright & 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., written composition scored from zero to four 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 over the raw score 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 function:

$$p_{im}(\theta) = \frac{\exp\left[\sum_{k=0}^{m}(\theta - \delta_{ik})\right]}{\sum_{v=0}^{M_i-1}\exp\left[\sum_{k=0}^{v}(\theta - \delta_{ik})\right]}, \qquad (1)$$

where M_i is the number of score categories of item i; θ is a student's proficiency (ability) score; $m(=0,1,\ldots M_i-1)$ is a raw score of item i; $p_{im}(\theta)$ is the probability of getting score m on item i conditional on θ; δ_{ik} is the step difficulty parameter of score k on item i; and denote $\theta - \delta_{i0} \equiv 0$.

The RPCM provides the probability of scoring each value of *m* on item *i* as a function of a student's proficiency score *θ*, and the step difficulties δ_{ik}, which indicate the proficiency score at which the probability of scoring *k* equals the probability of scoring *k*-1 (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 this 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 as follows.

- All items, regardless of type, are placed on the same common Rasch scale.
- Students' achievement results are placed onto the same scale as the items, so it is possible to make inferences about which items a student is likely to get correct or incorrect based on the student's proficiency. 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 assessment. 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 and previous test forms. Linking the current test form to previous test forms enables comparisons of test difficulties and passing rates across test forms given in different administrations. Because both pre-equated and post-equated item difficulty estimates are available, any drift in scale or difficulty can be quantified.

Student test scores on the Rasch scale are converted using a linear transformation to a more user-friendly reporting scale.

HORIZONTAL SCALING

The STAAR scale scores (SS_θ) represent linear transformations of Rasch-based proficiency 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 operation is represented by the following equation:

$$SS_\theta = A \times \theta + B \quad (2)$$

The slope and intercept in Equation (2) are called scaling constants, and they are derived using a method described by Kolen and Brennan (2004). For STAAR and STAAR Alternate 2, 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:

$$A = \frac{\sigma_{SS}}{\sigma_\theta} \quad (3)$$

In Equation (3), σ_{SS} represents the desired standard deviation of the scale, and σ_θ represents the standard deviation of Rasch-based θ values among a sample group. For example, the standard deviation σ_θ was established for each STAAR EOC assessment using all students who took that assessment in spring 2011 (or spring 2013 in the case of English I and English II). For the STAAR grades 3–8 horizontal scales, the sample group for a given assessment consisted of all students who took that assessment in

spring 2012. For the STAAR Alternate 2 horizontal scales, the sample group for a given assessment consisted of all students who took that assessment in spring 2015. The *B* scaling constant is calculated for STAAR as follows in equation 4 and for STAAR Alternate 2 in equation 5:

$$B = SS_{Meets} - \frac{\sigma_{ss}}{\sigma_{\theta}} \times \theta_{Meets} \quad (4)$$

$$B = SS_{Level\ II} - \frac{\sigma_{ss}}{\sigma_{\theta}} \times \theta_{Level\ II} \quad (5)$$

Because each assessment's horizontal scale is derived using its own sample group, σ_{θ} varies across assessments. Likewise, each assessment has a unique Meets Grade Level performance standard on STAAR - in Rasch units, so θ_{Meets} varies across assessments. SS_{Meets} and σ_{SS} are set to be consistent within academic content areas but not across all assessments. Similarly, the STAAR Alternate 2 Level II: Satisfactory performance standards are also unique for each assessment, with $\theta_{Level\ II}$ varying across assessments, and $SS_{Level\ II}$ and σ_{SS} are set to be consistent within academic content areas. Once these constants are established, the same transformations are applied each year to the Rasch proficiency estimates derived from performance on that year's test questions.

VERTICAL SCALING

A vertical scale score system 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 in the same content area 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 mathematics in spring 2015, the STAAR grades 3–8 reading assessments and the STAAR Spanish grades 3–5 reading assessments in spring 2012. For these studies, embedded field-test positions from several regular field-test forms (refer to the Field-Test Equating section of this chapter) included vertical linking items instead of field-test 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 adjacent grade-level tests, allowing for direct comparison of item difficulties across grade levels. By embedding vertical linking items across grade levels, it is possible to calculate linking constants equal to the average differences in item difficulties of vertical linking items between adjacent grade pairs. These linking constants are used to create a vertical scale.

For detailed information about vertical scaling studies, refer to the Assessment Reports and Studies webpage on TEA's Student Assessment Division website.

Similar to the horizontally scaled assessments, vertically scaled scale scores also reflect linear transformations of Rasch-based proficiency scores (*θ*). Vertically scaled

scores, however, include an extra scaling constant (V_g) that varies across each grade (g). This is given by the equation below,

$$SS_\theta = A \times (\theta - V_g) + B \ , \qquad (6)$$

where SS_θ is the scale score for a Rasch proficiency score (θ). The scaling constants A and B in Equation (6) 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., Meets Grade Level for STAAR) is fixed only for one of the assessments in the vertical scale (e.g., STAAR grade 8 mathematics for the STAAR mathematics vertical scale), and the standard deviation is calculated across all of the assessments (e.g., all STAAR grades 3–8 mathematics assessments). The A scaling constant is calculated as follows:

$$A = \frac{\sigma_{SS}}{\sigma_\theta} \qquad (7)$$

In Equation (7), σ_{SS} represents the desired standard deviation of the scale across all assessments, while σ_θ represents the standard deviation of Rasch-based θ values for a sample group. The STAAR grades 3–8 reading vertical scale sample group consisted of all students who took a test form with embedded vertical scale items in spring 2012. For the STAAR grades 3–8 mathematics vertical scale, the sample group consisted of all students who took a test form with embedded vertical scale items in spring 2015. Like field-test items, vertical scale items are not used to calculate student scores.

The B scaling constant is calculated as follows:

$$B = SS_{Meets} - \frac{\sigma_{SS}}{\sigma_\theta} \times \theta_{Meets} \qquad (8)$$

In Equation (8), SS_{Meets} represents the desired scale score at the STAAR Meets Grade Level cut for the final assessment in the vertical scale, and θ_{Meets} represents the approved STAAR Meets Grade Level performance standard in Rasch units for the final assessment in the vertical scale.

Equating

Texas uses the common-item nonequivalent groups design 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 is 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 three stages in the item and test development process related to equating.

1. Pre-equating test forms that are under construction

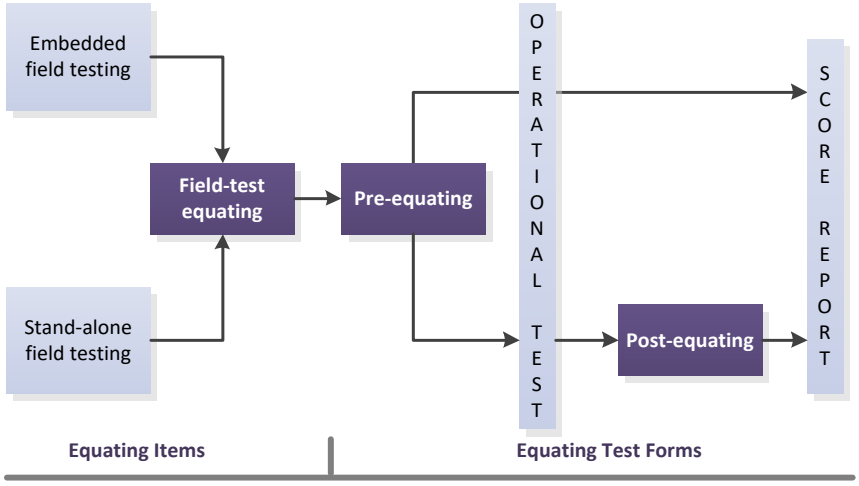
2. Post-equating operational test forms after administration

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 assessment performance standards for Algebra I, Biology, and U.S. History were approved by the Commissioner of Education in April 2012, and those STAAR EOC assessments were administered for the first time in spring 2012. Thus, the scale-score systems for those STAAR EOC assessments were 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. All Texas assessments are using one or more of these three types of 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 form is placed onto the Rasch scale prior to administration. 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 was established in advance (the items appeared previously on one or more test forms as field-test or operational items), the difficulty level of newly developed test forms can be estimated, 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 indicates the performance level cut scores.

The pre-equating process involves the following steps.

1. Select items that have been equated to the Rasch scale from the item bank.
2. Construct a new test form that meets the content specifications and statistical guidelines.
3. Evaluate the test form under construction against 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 in the following section) is also conducted. For some assessments, however, post-equating is not conducted, and the pre-equated RSSS conversion tables are 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 form or when faster reporting of scores is a priority.

POST-EQUATING

Post-equating might be preferred when changes in item presentation (e.g., position, formatting) or instructional practice have occurred since the time an item was field tested because those changes might impact the estimated difficulty of the item. Post-equating in the Texas assessment program employs conventional common-item nonequivalent groups equating design, whereby an equating constant is calculated and used 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:

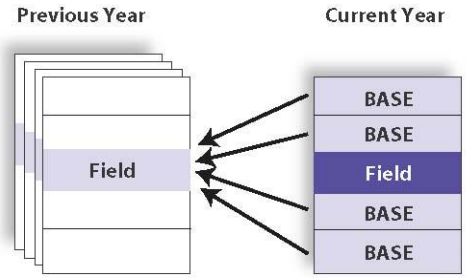
$$t_{a,b} = \frac{\sum_{i=1}^{k}(d_{i,a} - d_{i,b})}{k}, \quad (9)$$

where $t_{a,b}$ is the equating constant; $d_{i,a}$ is the Rasch difficulty of item i on the current form a; $d_{i,b}$ is the Rasch difficulty of item i on the item bank scale; and k is the number of common items (Wright, 1977). Once the equating constant is calculated, it is applied to all item difficulties, transforming them so they are on the item bank scale. After this transformation, the item difficulties from the current administration of the test are directly comparable to the item difficulties from all past administrations of the test (because

equating was also performed on those items). These updated item difficulty estimates are then used to create the RSSS conversion table that is used to report scale scores. 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.

Equating items are identified differently for STAAR Alternate 2 and TELPAS. For STAAR Alternate 2 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 STAAR Alternate 2 and TELPAS post-equating design. The arrows in figure 3.4. 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 their appearance in previous assessments.

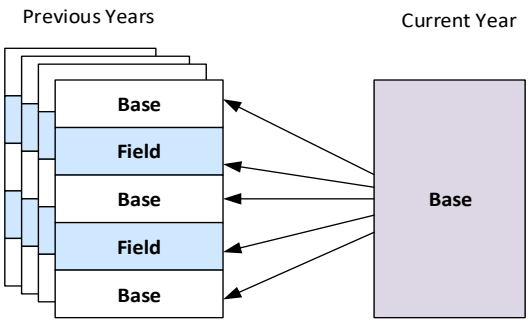
Figure 3.4. STAAR Alternate 2 and TELPAS Common-Item Post-Equating Design



STAAR Alternate 2 and 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 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. Prior to applying the final equating constant, the number of items in the equating set is compared to the base test, and the content representation of the equating item set is compared to the base test to verify that the test content is appropriately represented in the equating item set.

STAAR assessments use all base-test items as the equating item set. Figure 3.5 illustrates the source of the common item sets for these tests. The base-test items in the current year form could be field-test items or operational items from forms used in previous years.

Figure 3.5. STAAR Common-Item Post-Equating Design



STAAR post-equating is conducted on a sample of students in order to provide score reporting in a timely manner. The requirements for the sample include a minimum sample size of 100,000 students (except for STAAR Spanish, Algebra II, and English III), regional representation similar to the student population, ethnic distribution similar to the student population, and a representative 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. 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. If an item's Rasch item difficulty is less stable than expected, the item will be excluded from the equating item set during the stability check. Prior to applying the final equating constant, the number of items in the equating set is compared to the base test, and the content representation of the common item set is compared to that of the base test to verify that the reporting categories are appropriately represented.

The post-equating procedure involves the following 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 and released, newly developed items must be field tested and equated to the Rasch scale of the assessment. The STAAR (except tests with written compositions), STAAR Alternate 2, and TELPAS assessments use embedded field-test designs to collect data on field-test items. Additionally, grade 4 writing, Spanish grade 4 writing, grade 7 writing, English I, and English II use standalone field-test designs to collect data on field-test writing prompts.

In field-test designs, after a newly constructed item has cleared the review process, it is embedded in a test form along with 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 or in the case of STAAR Alternate 2, one cluster) 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 base-test items as the common items, calculate an equating constant equal to the difference between the mean Rasch item difficulty estimates for the common items on the base Rasch scale and for 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 base scale, the equated field-test items are also on the base scale.

MATCHED SAMPLE COMPARABILITY ANALYSIS

When the same assessment is administered in paper and online delivery modes, studies can be conducted to determine whether using the same RSSS conversion table for both delivery modes is warranted. Texas uses a comparability methodology known as Matched Samples Comparability Analysis (MSCA; Way, Davis, & Fitzpatrick, 2006). In this design, a bootstrap sampling approach, described in the Sampling section (see *Resampling and Replication Methods: Bootstrap*) of this chapter, is used to select online and paper student samples wherein each selected online test taker is matched to a paper test taker with the same demographic variables and similar performance on previous tests. 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 replicated or

repeated many times. RSSS conversion tables are retained and aggregated across replications, and the mean and standard deviation of the scale scores are determined 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 is 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.

Reliability

The concept of reliability is based on the idea that repeated administrations of the same test should generate consistent results. The degree to which results are consistent is assessed using a reliability coefficient. Reliability is a critical technical characteristic of any measurement instrument because unreliable scores cannot be interpreted in a meaningful way.

INTERNAL CONSISTENCY ESTIMATES

Reliability coefficients 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 (e.g., for low-stakes or high-stakes purposes). Two types of internal consistency measures used to estimate the reliability of Texas assessments are described below.

- Kuder-Richardson 20 (KR_{20}) 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.

KR_{20} is a mathematical expression of the classical test theory definition of test score reliability as the ratio of true score variance (i.e., no measurement error) to observed score variance (i.e., measurement error included). The classical test theory concept of reliability, in general, can be expressed as:

$$P'_{XX} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2+\sigma_E^2}\,, \tag{10}$$

where the reliability P'_{XX} of test *X* is a function of the ratio between true score variance σ_T^2 and observed score variance σ_X^2, which is further defined as the sum of the true score variance and error variance $\sigma_T^2 + \sigma_E^2$. As error variance is reduced, reliability increases (that is, students' observed scores are more precise estimates of their true scores). KR_{20} can be represented mathematically as:

$$KR_{20} = \left[\frac{k}{k-1}\right]\left[\frac{\sigma_X^2-\sum_{i=1}^{k} p_i(1-p_i)}{\sigma_X^2}\right], \tag{11}$$

where KR_{20} is a lower-bound estimate of the true reliability; k is the number of items in test X; σ_X^2 is the observed score variance of test X; and p_i is the proportion of students who answered item i correctly. This formula is used when test items are scored dichotomously.

Coefficient alpha (also known as Cronbach's alpha) is an extension of KR_{20} to cases where items are scored polytomously (in more than two possible score categories) and is computed as follows:

$$\alpha = \left[\frac{k}{k-1}\right]\left[1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_X^2}\right], \quad (12)$$

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

The stratified coefficient alpha is an 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 written composition) is treated as a subtest. A separate measure of reliability is computed for each component and combined as follows:

$$Stratified\ \alpha = 1 - \frac{\sum_{j=1}^{c}\sigma_{X_j}^2(1-\alpha_j)}{\sigma_X^2}, \quad (13)$$

where c is the number of item-type components; α_j is the estimate of reliability for each item-type component; $\sigma_{X_j}^2$ is the observed score variance for each item-type component j; and σ_X^2 is the observed score variance for the total score. For components consisting of multiple-choice or short answer items, coefficient alpha is used as the estimate of component reliability. The correlation between ratings of the first two raters (i.e., interrater reliability) is used as the estimate of component reliability for written compositions.

INTERRATER RELIABILITY

Assessments that are not composed of multiple-choice and gridded-response items might require different types of reliability evidence than those described above. For example, TELPAS writing involves teachers evaluating students based on their recent demonstrations of English language proficiency in the classroom. As part of the process for evaluating the reliability of such assessments, TEA provides evidence that the teacher observation and resulting evaluation of student performance were appropriately conducted.

To gather such evidence of interrater reliability, two evaluators observe the same student performance and then independently provide ratings of that performance. These ratings can then be analyzed, and the extent of agreement (or correlation) between the two sets of ratings can be calculated. 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 is not associated with 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, and they affect scores in a consistent manner (i.e., making them lower or higher). An example of a systematic error would be a language barrier that caused a student to incorrectly answer questions to which he or she knew the answer. By contrast, random errors are chance occurrences that may increase or decrease test scores. 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 consistency and accuracy for the purpose of estimating the amount of random error in test scores.

STANDARD ERROR OF MEASUREMENT (SEM)

The SEM reflects the amount of random variance in a score resulting from factors other than what the assessment is designed to measure. Because underlying traits such as academic achievement cannot be measured with perfect precision, the SEM is used to quantify the margin of uncertainty in test scores. For example, factors such as chance error and differential testing conditions can cause a student's observed score (the score achieved on a test) to fluctuate above or below his or her true score (the student's expected score). The SEM is calculated using both the standard deviation and the reliability of test scores, as follows:

$$SEM = \sigma_X \sqrt{(1 - P'_{XX})}\ , \qquad (14)$$

where P'_{XX} is the reliability estimate (for example, KR_{20}, coefficient alpha, or stratified alpha), and σ_X is the standard deviation of raw scores on test *X*. A standard error provides some sense of the uncertainty or error in the estimate of the true score using the observed score. For example, suppose a student achieves a raw score of 50 on a test with a SEM of 3. Placing a one-SEM band around this student's score would result in a raw score range of 47 to 53. If the student took the test 100 times, about 68 of those test raw scores will fall into the range of 47 to 53. In other words, the student's true score has 68% probability to be in this range.

It is important to note that the SEM provides an estimate of the average test score error for all students regardless of their individual proficiency scores. 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)

Like the SEM, the CSEM reflects the amount of variance in a score resulting from random 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. The CSEM is usually smallest, and thus scores are most reliable, near the middle of the score distribution because achievement tests typically include a relatively large number of moderately difficult items (compared to easy or difficult items), and such items provide more precise information about student proficiency near the middle of the score distribution.

IRT methods for estimating score-level CSEM are used because test- and item-level difficulties for STAAR, STAAR Alternate 2, and TELPAS are calibrated using the Rasch measurement model, as described 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 CONSISTENCY AND ACCURACY

Test scores are used to classify students into performance levels. Because all test scores contain errors, the classifications have errors too. Usually, there are two indicators to evaluate the quality of classifications: consistency and accuracy. Consistency refers to the percentage of students who are classified into the same performance levels if they took two parallel forms of a test, while accuracy refers to the percentage of students who are correctly classified into their true performance levels based on their observed scores on a test. Classification consistency and accuracy are two related but different concepts; high consistency does not necessarily lead to high accuracy, and vice versa. To better understand the classification quality, TEA conducts an analysis of the consistency and accuracy of student classifications into performance levels based on results of tests for which performance standards have been previously established.

The classification consistency index developed for IRT models (Lee, 2010) is used here. The basic idea is to estimate the probability (P1) of classifying into each performance level conditional on each test raw score based on an IRT model. For a performance level and a raw score, the probability (P2) that the raw score is classified into the same performance level on two parallel forms is just the square of the above probability for one test (P1). Across all performance levels, the probability (P3) that a raw score is consistently classified on two parallel forms is the sum of the above probabilities for two tests and one performance level (P2). The consistency index for a test is then the sum of the above probabilities (P3) over all raw scores weighted by the observed percentages of students on each raw score. The mathematical formula of consistency index can be expressed as:

$$\hat{\varphi} = \sum_{r=0}^{r_5} \left\{ \sum_{l=1}^{3} \left[\sum_{x=r_l}^{r_{l+1}-1} \hat{p}(x \mid \hat{\theta}_r) \right]^2 + \left[\sum_{x=r_4}^{r_5} \hat{p}(x \mid \hat{\theta}_r) \right]^2 \right\} f_r \,, \qquad (15)$$

where l is the performance level (for STAAR tests 1 = Does Not Meet, 2 = Approaches, 3 = Meets, 4 = Masters); r_l and r_{l+1} are the raw score cuts for level l and $l+1$, respectively, with r_1 = 0 and r_5 = maximum possible test raw score; $\hat{\theta}_r$ is the estimated proficiency score associated with raw score r; $\hat{p}(x|\hat{\theta}_r)$ is the estimated probability of getting raw score x conditional on $\hat{\theta}_r$; and f_r is the percentage of students with raw score r. The probability, $\hat{p}(x|\theta_r)$, can be estimated based on the following recursive algorithm:

$$\hat{p}(x|\hat{\theta}_r) = \sum_{m=0}^{M_i-1} \hat{p}_{i-1}(x-m|\hat{\theta}_r)\hat{p}_{im}(\hat{\theta}_r),$$

where i refers to the i-th item in a test; x is a raw score in a performance level which is between the minimum ($\min_i$) and maximum ($\max_i$) scores after adding the i-th item; M_i is the number of score categories for item i; $\hat{p}_{im}(\hat{\theta}_r)$ is the estimated probability of getting score m on item i conditional on $\hat{\theta}_r$, which is calculated based on the Rasch partial-credit model (Equation 1); $\hat{p}_i(x|\hat{\theta}_r)$ is the estimated probability of getting score x conditional on $\hat{\theta}_r$ after adding the i-th item. Note that $\hat{p}_0(x|\hat{\theta}_r)=1$, and when $x-m < \min_{i-1}$ or $x-m > \max_{i-1}$ for $i>1$, then define $\hat{p}_{i-1}(x-m|\hat{\theta}_r)=0$.

The method recommended by Rudner (2001, 2005) is adapted here for computing classification accuracy. Under an IRT model, for an estimated proficiency score $\hat{\theta}_r$ associated with raw test score r, the true proficiency score θ_r is expected to be normally distributed with a mean of $\hat{\theta}_r$ and an estimated standard deviation of $\hat{\sigma}_{\theta_r}$ (the CSEM). The estimated proficiency score cut $\hat{\theta}_l$ for each performance level l is also available. Then, for each raw score point in a performance level, the probability of correctly classifying into this level can be estimated. The accuracy index is just the sum of these probabilities across all test raw scores weighted by the observed percentages of students on each raw score point, f_r. In particular, the estimation formula is written as:

$$\hat{\psi} = \sum_{l=1}^{3}\sum_{r=r_l}^{r_{l+1}-1}\left[\phi\left(\frac{\hat{\theta}_{l+1}-\hat{\theta}_r}{\hat{\sigma}_{\theta_r}}\right)-\phi\left(\frac{\hat{\theta}_l-\hat{\theta}_r}{\hat{\sigma}_{\theta_r}}\right)\right]f_r + \sum_{r=r_4}^{r_5}\left[\phi\left(\frac{\hat{\theta}_{l+1}-\hat{\theta}_r}{\hat{\sigma}_{\theta_r}}\right)-\phi\left(\frac{\hat{\theta}_l-\hat{\theta}_r}{\hat{\sigma}_{\theta_r}}\right)\right]f_r, \qquad (16)$$

where ϕ is cumulative standard normal distribution function, and θ_l is the proficiency score cut for level l with $\theta_{l=1}$ = -10 and $\theta_{l=5}$ = 10.

Note that each STAAR EOC assessment has two different Approaches level cuts for students who first took an EOC assessment before the December 2015 administration and students who first took an EOC assessment on or after the December 2015 administration. Therefore, for each EOC assessment, first estimate the classification consistency/accuracy for each group of students who have the same Approaches cut (i.e., “Approaches 2012-15” or “Approaches”), and then sum the classification

consistency/accuracy indexes weighted by proportion of students in each group as the overall classification consistency/accuracy estimate for a test.

Validity

In the Texas assessment program, validity refers to the extent to which test scores help educators make appropriate inferences about student achievement. The concepts described here are not types of validity, but types of validity evidence. Validity evidence can be organized into five categories (described in detail below): test content, response processes, internal structure, relations to other variables, and consequences of testing (AERA/APA/NCME, 2014; Schafer, Wang, & Wang, 2009). Such evidence supports the valid interpretation and use of test scores. It must be acknowledged, however, that validation is a matter of degree and is an ongoing process.

Evidence Based on Test Content

Validity evidence based on test content supports the assumption that the content of the test adequately reflects the intended construct. For example, the STAAR test scores are designed to help make inferences about students' knowledge and understanding of the statewide curriculum standards, the TEKS. Therefore, evidence supporting the content validity of the STAAR assessments, including STAAR Spanish, maps the test content to the TEKS. Validity evidence supporting Texas' test content comes from the established test development process and the judgments of content experts about the relationship between the items and the test construct.

The test-development process started with a review of the TEKS by Texas educators. The educators then worked with TEA to define the readiness and supporting standards in the TEKS and helped determine how each standard would best be assessed. A test blueprint was developed with educator input, which maps the items in future development to the reporting categories they are intended to represent. Items were then developed based on the test blueprints. 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.

- Develop items based on the TEKS curriculum standards and item guidelines.
- Review items for appropriateness of item content and difficulty and to eliminate potential bias.
- Collect and review data on field-test items to determine appropriateness for inclusion on a test.
- 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 showing that the manner in which students are required to respond to test items supports an 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 reflects an appropriate manner for students to demonstrate their writing abilities. Student response processes on Texas' assessments differ by both item type and administration mode.

The STAAR program requires students to respond to three item types: multiple-choice, gridded-response, and written compositions. STAAR Alternate 2 items involve test administrators observing students as they respond to standardized items and scoring them based on item-specific rubrics. TELPAS online assessments require students to respond to multiple-choice items, technology enhanced items, and performance-based speaking tasks scored by an automated-scoring engine and/or human scorers. TELPAS holistic and TELPAS Alternate assessments do not contain traditional items and students are instead evaluated by either holistic ratings based on ongoing classroom observations and students interactions, or by specific descriptions of behaviors, called Observable Behaviors. Texas gathers evidence to support validity based on response processes from several sources. First, when new item types or changes to the format of existing item types are considered for any of the Texas assessments, cognitive labs are used to study the way students engage with the various item presentations. In cognitive labs, students "think aloud" while responding to assessment items, and this can provide evidence that students' cognitive processes are consistent with those expected of a given item type and reflect the knowledge and skills described in the TEKS. Next, test items are pilot-tested with a larger sample of students to gather information about performance on new item types and formats. After new 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, point-biserial correlations, and differential item functioning. The evidence is then submitted to content expert review.

The process used to score items can provide validity evidence related to response processes. For assessments with constructed-response items, such as written compositions, rubrics are used by human readers to score student responses. For TELPAS speaking assessments, the speaking responses are scored by an automated scoring process. The validity of student scores is supported if such rubrics accurately describe the characteristics of student responses on a continuum from low to high quality. All rubrics for the STAAR assessments have been validated by educator committees and content experts. In addition, TEA has implemented a rigorous scoring process for the constructed-response items that includes training and qualification requirements for readers; ongoing monitoring during scoring; adjudication and resolution processes for student responses that do not meet the perfect/adjacent scoring requirements; and rescoring of responses as needed. A more comprehensive

description of the scoring process for constructed-response items is available in chapter 2, "Building a High-Quality Assessment System."

When students are given the option to take tests either on paper or online, evidence is necessary to indicate 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 paper and online test score interpretations. Separate conversion tables are used for tests when evidence suggests that the paper and online forms are not comparable.

Evidence Based on Internal Structure

When a test is designed to measure a single construct, the internal components of the test should exhibit a high level of homogeneity that can be quantified in terms of the internal consistency reliability coefficients, as described in the Reliability section of this chapter. Internal consistency estimates are evaluated for Texas assessments for reported student groups, including all students as well as 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 English and Spanish 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, predictive, convergent, or discriminant.

- Concurrent: the performance on two measures taken at the same time are correlated.
- Predictive: the current performance on one measure predicts performance on a future measure.
- Convergent: the performance on two measures that are meant to assess the same or similar construct should be strongly correlated; or
- Discriminant: the performance on two measures that are meant to assess unrelated constructs should have a weak correlation or no correlation.

A large number of research studies have been and continue to be conducted to evaluate the relationship between performance on the STAAR assessments and performance on other related tests or criteria. The studies include the following:

- STAAR-to-TAKS comparison studies, which link performance on the STAAR assessments to performance on TAKS assessments (e.g., the STAAR grade 7 mathematics to the TAKS grade 7 mathematics)
- STAAR linking studies, which link performance on the STAAR assessments across grade levels or courses in the same content areas (e.g., grade 4 reading to grade 5 reading, and English I to English II)
- STAAR inter-correlation estimates, which evaluate the strength of the relationship (or lack thereof) among scores on the STAAR assessments across different content areas (e.g., grade 4 mathematics to grade 4 reading, and English I to biology)
- 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 (e.g., Scholastic Aptitude Test [SAT] and American College Testing [ACT])
- college students taking STAAR studies, which link performance on the STAAR EOC assessments to college course grades

For detailed descriptions and results of such studies, refer to the STAAR Performance Standards webpage of TEA's Student Assessment Division website.

Like STAAR, STAAR Alternate 2 inter-correlation estimates are calculated, which evaluate the strength of the relationship between scores on the STAAR Alternate 2 assessments across different content areas. Results from all these analyses are provided in Appendix C.

To examine validity evidence based on external measures for TELPAS, an annual analysis is conducted of the relationship between TELPAS reading/writing performance and STAAR reading/writing or STAAR English EOC assessment performance. For each grade level and TELPAS proficiency level breakout group, two types of performance data are examined:

- average STAAR scale scores
- STAAR passing rates (Approaches Grade Level Performance)

See chapter 6, "TELPAS" for more details and the results of these studies. The same analysis is also conducted for the TELPAS Alternate assessments and the relationship to STAAR Alternate reading, writing and English performance. See chapter 7, "TELPAS Alternate" for more details and the results of these studies.

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 inferences based on test scores. For example, the STAAR assessments are intended to

have an effect on 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.

Given the important stakes associated with the Texas assessment program, the validity of interpretations and uses of test scores are critical. The intended interpretations of test results are stated in the policy definitions of the performance levels, which are provided on the STAAR Performance Standards webpage of TEA's Student Assessment Division website.

Measures of Student Progress

Measures of student progress express a comparison between current and previous student performance. Student progress information provides essential context to understanding students' current performance. For example, consider a student who achieves Approaches Grade Level on a STAAR assessment. The interpretation of Approaches Grade Level performance would depend on the performance the student achieved in the current year. If the student achieved Did Not Meet Grade Level performance in the previous year, then the student made notable progress this year by advancing a performance level. However, if the student had achieved Meets Grade Level performance in the previous year, then the interpretation of Approaches Grade Level performance this year would be quite different because the student regressed a performance level.

Student progress information can also provide insight to help set future performance goals. For example, one goal would be for all students to achieve at or above Meets Grade Level performance on the STAAR assessments. When considered together, student progress measures and current performance can be used to set reasonable, individual goals. For those students who have not yet reached Meets Grade Level performance, progress measures can be used to evaluate whether a student is on track to meet the Meets Grade Level performance in a future year. To that end, TEA calculates a STAAR on-track measure, which provides information about whether a student is on track to be at or above the Meets Grade Level performance standard in a future target year. Using gain scores, individual students are categorized as *Not On Track* or *On Track* toward the target year. On-track measures are available for STAAR reading in grades 4–7, STAAR Spanish reading in grade 4, and STAAR mathematics in grades 4–8. Details about the calculation of STAAR on-track measures are provided in the "STAAR On-Track Measure Q&A" available on the Progress Measures webpage of TEA's Student Assessment Division website.

Types of Student Progress Measures

Given the value of progress information, student progress measures are calculated and reported for STAAR and STAAR Alternate 2. Several types of progress measures were considered for use with STAAR and STAAR Alternate 2, including student growth

models based on Regression, Student Growth Percentile, Growth to Proficiency, Value/Transition Tables, and Gain Scores.

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

Development of STAAR and STAAR Alternate 2 Progress Measures

As part of the development of STAAR and STAAR Alternate 2 progress measures, several factors were considered, including

- the suitability of different models for measuring student progress given the characteristics of the STAAR and STAAR Alternate 2 assessments,
- the appropriateness of progress measures given the content relationships among STAAR and STAAR Alternate 2 assessments,
- the usability of progress measures for accountability given federal and state requirements, and
- the effectiveness of communicating progress measure results given various reporting options.

Additionally, input was sought from a number of advisory groups regarding the development of the STAAR and STAAR Alternate 2 progress measures. Several options for progress measures were presented to the Texas Technical Advisory Committee (TTAC), a national group of educational measurement experts who provided recommendations and guidance. Progress measures were also discussed with the Accountability Technical Advisory Committee (ATAC) and the Accountability Policy Advisory Committee (APAC), which are groups consisting of educators from various Texas campuses, districts, and Education Service Centers (ESCs), as well as parents, higher education representatives, business leaders, and legislative representatives. Input from these groups was requested at several points during the development of progress measures for STAAR and STAAR Alternate 2.

Implementation

Based on the input and considerations described above, gain scores were selected as the progress measure for STAAR (refer to the STAAR Progress Measures Questions and Answers document for more information). The STAAR Progress Measure was implemented for the first time in 2012–2013. A progress measure was reported for English learners (ELs) for the first time in 2013–2014 and was discontinued in 2017–2018 due to TELPAS revisions. An EL performance measure was reported for qualifying ELs beginning 2018–2019. More information can be found TEA's Assessment Scoring and Reporting website.

The STAAR Alternate 2 progress measure employs a transition table approach and was reported for the first time in 2016. More details about the STAAR Alternate 2 progress

measure are available on the Progress Measures webpage of TEA's Student Assessment Division website.

STAAR progress measures are calculated and reported for grades 4–8 mathematics and reading (including Spanish), Algebra I, and English II. An EL performance measure was calculated and reported for all STAAR assessments except STAAR Spanish, Algebra II, and English III. For STAAR Alternate 2, progress measures are calculated and reported for grades 4–8 mathematics and reading, Algebra I, English I, and English II. Details about these progress measures can be found in chapter 4, "STAAR", and chapter 5, "STAAR Alternate 2." Due to the cancellation of STAAR and STAAR Alternate 2 testing in the spring and summer of 2020, neither progress measure nor on-track measure is available in 2020.

Sampling

Sampling is a procedure used to select a relatively small number of observations that are representative of the population from which they are drawn. In this case, sampling involves the selection of a set of Texas students that is representative of the entire body of Texas students. The results from well-drawn samples allow TEA to estimate characteristics of the 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 should generalize. For example, consider a study with the goal of understanding how grade 3 ELs perform on a set of test questions. In that case, the target population would be all grade 3 ELs in Texas. Careful consideration is given to defining the target population before sampling takes place.

SAMPLING, SAMPLES, AND OBSERVATION UNITS

Sampling is the process of selecting a subset of the target population to participate in a study. A well-drawn sample allows 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, consider a study designed to estimate student computer access in the

entire state, and each of the same 20 sampled campuses is requested to report student data on how many students have computer access at home. In that case, 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 reflect student characteristics).

REASONS FOR SAMPLING

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

- **Accessibility —** There are situations where collecting data on every member of the target population is not feasible.
- **Burden —** Sampling minimizes the participation requirements for the campus and district, thereby reducing the testing burden.
- **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.
- **Size —** It is more efficient to examine a representative sample when the size of the target population is very large.
- **Time —** Using sampling to study the target population is less time-consuming. Sampling might be needed when the speed of the analysis is important.

SAMPLING DESIGNS

The Texas assessment program uses the following sampling designs to collect data for the purpose of field testing, audits, and research studies (e.g., linking studies, cognitive labs, 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. 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 sampling units are first grouped (i.e., stratified) according to variables of interest such as gender and ethnicity; then, a random sample is selected from each group.
- **Cluster sampling —** The sampling units are first grouped into clusters according to variables of interest. Then, unlike stratified sampling, a predetermined number of clusters is randomly selected. All sampling units within the selected clusters are observed.

Regardless of the type of probability sampling design used, a decision about whether to sample with or without replacement must be made. To help clarify this distinction, consider simple random sampling with replacement and simple random sampling without replacement. First, suppose that a simple random sample of size *n* with replacement is drawn from a population of size *N*. In this case, when a sampling unit is randomly selected, that unit remains eligible to be selected again. In other words, after the sampling unit 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 its data would be duplicated in the resulting sample of size *n*.

By comparison, suppose that a simple random sample of size *n* without replacement is drawn from a population of size *N*. In this case, once a sampling unit is chosen, it is ineligible to be selected again. In other words, after the sampling unit 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 add 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 (see *Resampling and Replication Methods: Bootstrap*).

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

Bootstrapping is one of the resampling and replication methods, which treat the sample like a population. These methods repeatedly draw 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 & Tibshirani (1993). Texas uses bootstrapping methods when conducting comparability studies that compare online and paper versions of a test form.