TECHNICAL DIGEST 2022–2023

Chapter 1

Historical Overview of Assessment in Texas

Assessment Timeline

Texas has a long history of student assessment dating back to 1979, when its first statewide testing program was required by statute. Over the years, changes in legislation and policy have impacted the size and scope of the Texas Assessment Program. This chapter provides an overview of these changes, including an assessment timeline and a description of changes to the assessment program over time.

—1979–1980

The Texas Assessment of Basic Skills (TABS) was administered for the first time in February 1980. TABS included mathematics, reading, and writing assessments in grades 3, 5, and 9. The final administration of TABS was in fall 1985.

—1986–1987

The Texas Educational Assessment of Minimum Skills (TEAMS) was first administered in fall 1986 and included mathematics, reading, and writing assessments in grades 1, 3, 5, 7, 9, and 11. TEAMS represented the first time that Texas students were required to pass a state assessment to be eligible to receive a high school diploma; students had to pass the TEAMS grade 11 exit-level tests in mathematics and reading to graduate. The final administration of TEAMS was in fall 1989. After that, students who were required to meet TEAMS graduation requirements had to take the Texas Assessment of Academic Skills (TAAS) exit-level assessments with adjusted performance standards.

—1990–1991

First administered in fall 1990, TAAS shifted the focus of assessment from minimum skills to academic skills and included mathematics, reading, and writing assessments in grades 3, 5, 7, 9, and 11. Students had to pass the TAAS grade 11 exit-level assessments in mathematics, reading, and writing to receive their high school diploma.

—1993–1994

Administration of TAAS moved to the spring, and the grades and subjects assessed were reconfigured. From 1994 to 2002, TAAS was administered every spring to students in grades 3–8 and 10 in mathematics and reading; grades 4, 8, and 10 in writing; and grade 8 in science and social studies. Students had to pass the TAAS grade 10 exit-level assessments in mathematics, reading, and writing to be eligible to graduate.

The final administration of TAAS for grades 3–8 was in spring 2002. Because TAAS remained the graduation requirement for students in grade 9 or above on January 1, 2001, exit-level TAAS tests continued to be administered through July 2009. Subsequently, students who were required to meet TAAS graduation requirements were able to take Texas Assessment of Knowledge and Skills (TAKS) exit-level assessments with adjusted performance standards.

—1995–1996

Spanish-language TAAS mathematics and reading assessments were incorporated into the testing program for grades 3 and 4.

Algebra I and Biology end-of-course (EOC) assessments were administered for the first time to students who completed these courses.

—1996–1997

Spanish-language TAAS mathematics and reading assessments were incorporated into the testing program for grades 5 and 6.

—1998–1999

English II and U.S. History EOC assessments were administered for the first time to students who completed these courses. Through spring 2002, the four EOC assessments—Algebra I, English II, Biology, and U.S. History—were administered as state-mandated assessments and as an option for meeting graduation requirements.

—1999–2000

Reading Proficiency Tests in English (RPTE)

The Reading Proficiency Tests in English (RPTE) were first administered in spring 2000 to emergent bilingual (EB) students in grades 3–12.

—2000–2001

State-Developed Alternative Assessment (SDAA)

The State-Developed Alternative Assessment (SDAA) was first administered in spring 2001 to eligible students receiving special education services. SDAA included assessments in mathematics and reading for kindergarten through grade 8 and in writing for kindergarten through grade 7. The final administration of SDAA was in spring 2004.

—2002–2003

TAKS Texas Assessment of Knowledge and Skills

To satisfy legislative requirements, TAKS was designed to be more comprehensive than its predecessors and to measure more of the state-mandated curriculum known as the Texas Essential Knowledge and Skills (TEKS). TAKS was first administered in spring 2003 and included assessments in mathematics in grades 3–11; reading in grades 3–9; writing in grades 4 and 7;

English language arts (ELA) in grades 10 and 11; science in grades 5, 10, and 11; and social studies in grades 8, 10, and 11. Spanish versions of TAKS were administered in grades 3–6. Students had to pass the TAKS grade 11 exit-level tests in mathematics, ELA, science, and social studies to receive a high school diploma.

In compliance with the Student Success Initiative (SSI), satisfactory performance on TAKS grade 3 reading, grade 5 mathematics and reading, and grade 8 mathematics and reading assessments were requirements for promotion to the next grade level. These requirements became effective for grade 3 in the 2002–2003 school year, grade 5 in the 2004–2005 school year, and grade 8 in the 2007–2008 school year. The TAKS grade 3 reading promotion requirements were removed beginning with the 2009–2010 school year.

The final administration of TAKS for grades 3–10 was in spring 2011. Because TAKS remained the graduation requirement for students in grade 9 or above in the 2011–2012 school year, exit-level TAKS tests continued to be administered through June 2017. After that, students who were required to meet TAKS graduation requirements could take the State of Texas Assessments of Academic Readiness (STAAR®) EOC assessments with adjusted performance standards.

—2003–2004

To fulfill requirements of the federal No Child Left Behind Act (NCLB), the Texas Observation Protocol (TOP) was developed to assign holistic English language proficiency ratings for students based on observations during instruction. Holistic ratings were developed in the language domains of listening, speaking, and writing in kindergarten through grade 12 and in reading in kindergarten through grade 2.

TELPAS
Texas English Language Proficiency Assessment System

Together, TOP and RPTE formed the Texas English Language Proficiency Assessment System (TELPAS).

—2004–2005

LAT
Linguistically Accommodated Testing

In response to NCLB regulations, a linguistically accommodated testing (LAT) process was added to TAKS grades 3–8 and 10 mathematics for eligible EB students.

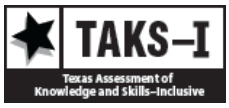
State-Developed Alternative Assessment II (SDAA II)

SDAA was replaced with SDAA II in spring 2005 to better align the alternate assessment to TAKS. SDAA II was available for students who received special education services in mathematics in kindergarten through grade 10, reading in kindergarten through grade 9, writing in kindergarten through grade 9, and ELA in grade 10. The final administration of SDAA II was in spring 2007.

In response to the 2004 Algebra Incentive Program and Executive Order RP53, the Algebra I EOC assessment was revised and made available online in spring 2005.

—2005–2006

Based on legislative requirements, TAKS grade 8 science was added to the testing program.



To meet the requirements of the Individuals with Disabilities Education Act (IDEA) of 2004, TAKS–Inclusive (TAKS–I) was added to the assessment program in spring 2006. TAKS–I was available for eligible students receiving special education services and included science in grades 5, 8, 10 and 11; science in Spanish in grade 5; social studies in grades 8, 10, and 11; and mathematics and ELA in grade 11. The final administration of TAKS–I was in spring 2007.

—2006–2007

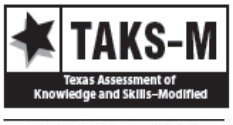
LAT administrations of TAKS grades 3–8 reading and grade 10 ELA were implemented in spring 2007 for eligible EB students.

—2007–2008

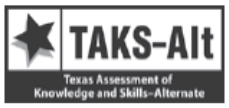
LAT administrations of TAKS grades 5, 8, and 10 science were implemented in spring 2008 for eligible EB students.

TAKS
(Accommodated)

TAKS (Accommodated) replaced TAKS–I for students receiving special education services who met the eligibility requirements for specific accommodations. TAKS (Accommodated) was available for mathematics in grades 3–11; reading in grades 3–9; writing in grades 4 and 7; ELA in grades 10 and 11; science in grades 5, 8, 10, and 11; and social studies in grades 8, 10, and 11. The final administration of TAKS (Accommodated) was in spring 2011.



TAKS–Modified (TAKS–M) was an alternate assessment based on modified academic achievement standards and was first administered in spring 2008. TAKS–M was available for eligible students receiving special education services and included mathematics in grades 3–11; reading in grades 3–9; writing in grades 4 and 7; ELA in grades 10 and 11; science in grades 5, 8, 10, and 11; and social studies in grades 8, 10, and 11. The final administration of TAKS–M was in spring 2011.



To fulfill federal requirements, TAKS–Alternate (TAKS–Alt) was first administered in spring 2008. It was developed for students with significant cognitive disabilities and was based on alternate achievement standards. TAKS–Alt included mathematics in grades 3–11; reading in grades 3–9; writing in grades 4 and 7; ELA in grades 10 and 11; science in grades 5, 8, 10, and 11; and social studies

in grades 8, 10, and 11. The final administration of TAKS-Alt was in spring 2011.

Based on NCLB requirements, TELPAS reading for grades 2–12 was redesigned and administered as an online testing program beginning in spring 2008.

EOC assessments in Geometry and Biology were first administered on a voluntary basis.

—2008–2009

Based on legislation, TAKS grade 6 assessments in Spanish were administered for the final time in spring 2009.

EOC assessments in Chemistry and U.S. History were first administered on a voluntary basis.

—2009–2010

EOC assessments in Physics and World Geography were first administered on a voluntary basis.

—2010–2011

EOC assessments in Algebra II and English I were first administered on a voluntary basis.

—2011–2012

STAAR replaced TAKS as the state academic assessment program beginning in spring 2012. STAAR included mathematics and reading in grades 3–8, writing in grades 4 and 7, science in grades 5 and 8, and social studies in grade 8. For high school, grade-specific assessments were replaced with 15 STAAR EOC assessments: Algebra I, Geometry, Algebra II, English I reading, English I writing, English II reading, English II writing, English III reading, English III writing, Biology, Chemistry, Physics, World Geography, World History, and U.S. History. STAAR Spanish was administered in grades 3–5.

In compliance with SSI, satisfactory performance on STAAR grades 5 and 8 mathematics and reading were requirements for promotion to the next grade level through spring 2021.

Depending on their graduation program, high school students were required to meet the passing standard (or achieve a predetermined minimum score) on at least 11 of the 15 STAAR EOC assessments. Additionally, students needed to meet a cumulative score requirement in each content area.

STAAR L, a linguistically accommodated English version of STAAR, was first administered online in spring 2012. STAAR L was available for EB students for

grades 3–8 and EOC assessments in mathematics, science, and social studies. The final administration of STAAR L was in fall 2016.

STAAR Modified replaced TAKS–M beginning in spring 2012. STAAR Modified originally included mathematics and reading in grades 3–8, writing in grades 4 and 7, science in grades 5 and 8, and social studies in grade 8. The final administration of STAAR Modified was in spring 2014.

STAAR Alternate replaced TAKS–Alt in spring 2012. STAAR Alternate included mathematics and reading in grades 3–8, writing in grades 4 and 7, science in grades 5 and 8, social studies in grade 8, and EOC assessments in Algebra I, Geometry, English I reading, English I writing, English II reading, English II writing, English III reading, English III writing, Biology, World Geography, World History, and U.S. History. The final administration of STAAR Alternate was in spring 2014.

—2012–2013

Based on legislative changes, spring 2013 was the final administration of STAAR Geometry, Chemistry, Physics, World Geography, and World History EOC assessments. STAAR Algebra II and English III post-secondary readiness assessments became optional, and their administration was suspended until spring 2016.

STAAR Modified EOC assessments in Algebra I, Geometry, English I reading, English I writing, English II reading, English II writing, Biology, World Geography, and World History were added to the testing program.

—2013–2014

Based on legislative requirements, STAAR high school English assessments were redesigned to combine reading and writing into a single assessment. The redesigned STAAR English I and English II EOC assessments were first administered in spring 2014.

The STAAR Modified U.S. History EOC assessment was added to the testing program.

—2014–2015

STAAR A was administered online for the first time in spring 2015 with embedded accommodations designed to help students who met eligibility requirements access the content being assessed. STAAR A was available for mathematics and reading in grades 3–8, writing in grades 4 and 7, science in grades 5 and 8, social studies in grade 8, and EOC assessments in Algebra I, English I, English II, Biology, and U.S. History. The final administration of STAAR A was in fall 2016.

STAAR Alternate 2 was administered for the first time in spring 2015 to eligible students with the most significant cognitive disabilities. STAAR Alternate 2 includes assessments for mathematics and reading in grades 3–8, science in grades 5 and 8, social studies in grade 8, and EOC assessments in Algebra I, English I, English II, Biology, and U.S. History.

—2015–2016

STAAR Algebra II and English III post-secondary readiness assessments were administered as optional assessments from spring 2016 through spring 2021.

—2016–2017

STAAR online with embedded supports replaced STAAR A and STAAR L beginning with the spring 2017 administration. This change allowed for a wider range of accessibility features and accommodations, including content supports and language and vocabulary supports, based on each student's needs.

—2017–2018

TELPAS listening and speaking holistic assessments for grades 2–12 were combined and made into standardized item-based assessments to be administered online. In addition, the blueprint for TELPAS reading was shortened.

New optional STAAR Interim Assessments were offered for grades 3–8 mathematics and reading, Spanish grades 3–5 mathematics and reading, and EOC assessments in Algebra I, English I, and English II.

—2018–2019

In the 2018–2019 school year, STAAR was administered online for the first time in American Sign Language (ASL) and refreshable braille. STAAR Spanish grades 3–5 assessments were offered online for the first time.

TELPAS Alternate was first administered in spring 2019 to EB students in grades 2–12 with the most significant cognitive disabilities. TELPAS Alternate is a holistic assessment process that includes the English language domains of listening, speaking, reading, and writing.

—2019–2020

In response to the COVID-19 pandemic, the Texas Education Agency (TEA) launched optional end-of-year (EOY) assessments that school systems and parents could choose to administer free of charge in the absence of STAAR to evaluate the academic progress students made.

—2020–2021

TEA launched optional beginning-of-year (BOY) assessments that school systems could choose to administer free of charge to evaluate the academic progress students made. BOY assessments were available each fall through the 2022–2023 school year.

—2021–2022

A braille version of TELPAS reading was available for the first time for students with visual impairments.

—2022–2023

STAAR transitioned to a primarily online assessment program beginning with the December 2022 administration.

Spring 2023 marked the launch of the STAAR redesign. New non-multiple-choice question types were present across all grades, subjects, and courses. STAAR reading language arts (RLA) assessments included reading and writing components.

TELPAS writing in grades 2–12 moved from a holistic assessment to a standardized item-based assessment administered online and was combined with the reading assessment beginning in spring 2023.

Changes to the Assessment Program Over Time

The Texas Assessment Program must comply with federal regulations and state statutes concerning student assessment. Federal regulations are mandated by NCLB, the Elementary and Secondary Education Act (ESEA), and the Every Student Succeeds Act (ESSA). The majority of state law pertaining to the statewide student assessment program is found in Texas Education Code (TEC) Chapter 39, Subchapter B.

The Texas Assessment Program measures students' understanding of the statewide curriculum. When the statewide curriculum is revised, changes are subsequently made to the assessment program to maintain a strong, direct, and effective link between the statewide curriculum and the state assessments.

The following provides a summary of changes in law and in the statewide curriculum that affected the Texas Assessment Program.

1979

The Texas Assessment Program began in 1979 when the 66th Texas Legislature enacted Senate Bill (SB) 350, which required basic skills competencies in mathematics, reading, and writing for grades 3, 5, and 9. As a result of SB 350, TABS was implemented in 1980.

1981

House Bill (HB) 246, passed by the 67th Texas Legislature, Regular Session, 1981, made changes to the state curriculum. As a result, the State Board of Education (SBOE) adopted the Essential Elements in 1984.

1984

HB 72, passed by the 68th Texas Legislature, Second Called Session, 1984, called for accountability provisions in exit-level testing requirements. HB 72 also led to the implementation of TEAMS, which replaced TABS in 1986.

1991

In 1991, the 72nd Texas Legislature passed SB 7, which required statewide testing of students in grades 3–8 and exit-level tests in high school. As a result, TEAMS was replaced by TAAS, which was administered from 1990 to 2002.

1995

Enacted by the 74th Texas Legislature in 1995, SB 1 overhauled the TEC and required the development of four EOC assessments. Students could use satisfactory performance on the Algebra I, the English II, and either the Biology or the U.S. History EOC assessment in place of TAAS to meet graduation assessment requirements.

1997

In July 1997, the SBOE replaced the Essential Elements with the TEKS. Implemented as the statewide curriculum for Texas in the 1998–1999 school year, the TEKS were developed to be more specific and focused than the Essential Elements, with emphasis placed on the knowledge and skills students were expected to learn rather than on the delivery standards expected of teachers.

1999

In 1999, the 76th Texas Legislature passed SB 103, which required the development of TAKS to replace TAAS. SB 103 also required the development of a system to assess the reading proficiency and language acquisition of EB students, resulting in the development of RPTE.

SSI, enacted by the Texas Legislature in 1999, made satisfactory performance on the grade 3 reading assessment, the grade 5 mathematics and reading assessments, and the grade 8 mathematics and reading assessments a promotion requirement for Texas students. The first cohort of students affected by this law was the grade 3 class of 2002–2003. Passing the grade 5 mathematics and reading assessments was a promotion requirement for the first time in the 2004–2005 school year. Grade 8 promotion requirements became effective in the 2007–2008 school year. In 2009, the Texas Legislature amended SSI to remove the grade 3 promotion requirement.

2005

In response to the governor's 2004 Algebra Incentive Program, the Algebra I EOC assessment was revised and administered online in spring 2005 on a voluntary basis to students who completed Algebra I coursework.

Executive Order RP53, issued by the governor in December 2005, called for increased college readiness programs in Texas schools and authorized the development of a series of EOC assessments in subjects assessed by TAKS in grade 11, including Algebra I, Geometry, Biology, Chemistry, Physics, and U.S. History.

2007

In May 2007, the 80th Texas Legislature enacted SB 1031, which required the implementation of an EOC assessment program. With the expanded role of the EOC assessment program, SB 1031 phased out TAKS grade level–based testing in high school and replaced it with EOC assessments as a component of the new high school graduation requirements that would apply beginning with the incoming freshman class of 2011–2012. The bill required the development of EOC assessments for Algebra I, Geometry, Algebra II, English I, English II, English III, Biology, Chemistry, Physics, World Geography, World History, and U.S. History.

HB 1, also passed in 2007, required TEA and the Texas Higher Education Coordinating Board (THECB) to develop the College and Career Readiness Standards (CCRS). After the CCRS were developed, TEA and THECB linked the CCRS to the TEKS in mathematics, RLA, science,

and social studies. Finally, as part of the TEKS review process, the SBOE incorporated the CCRS into the TEKS, making Texas the first state in the country to adopt a curriculum aligned to college and career readiness.

The SBOE adoption of new English Language Proficiency Standards (ELPS) for EB students in kindergarten through grade 12 was effective in December 2007. Beginning in 2008, TELPAS was aligned to the new ELPS.

2009

In 2009, the 81st Texas Legislature, Regular Session, enacted HB 3, which made further changes to the assessment program. HB 3 required that the performance standards for mathematics and reading assessments in grades 3–8 be linked from grade to grade to the college readiness performance standards for the Algebra II and English III assessments. The required vertical linking, along with the replacement of exit-level TAKS with EOC assessments, necessitated the design of a new series of assessments to indicate college readiness. As a result, STAAR was developed to encompass the EOC assessments mandated by SB 1031 in 2007 and the grades 3–8 assessments mandated by HB 3.

HB 3 also required the commissioner of education, rather than the SBOE, to determine performance levels for assessments and eliminated the exit-level TAAS assessments. As a result, students who had been required to meet TAAS graduation requirements could take TAKS exit-level assessments with adjusted performance standards.

2010

In 2010, the SBOE adopted revised social studies TEKS; alignment with those TEKS was reflected in the 2011–2012 STAAR social studies assessments.

2011

In 2011, the 82nd Texas Legislature, Regular Session, passed HB 2135, which impacted students receiving above-grade-level instruction. The bill allowed students who were enrolled in and taking the assessment for an above-grade-level course to not take the grade-level assessment. Additionally, the bill indicated that a student in an SSI grade could not be denied promotion based on performance on an assessment if the student was taking an above-grade-level assessment instead of the grade-level assessment.

2012

In 2012, the SBOE adopted new mathematics TEKS; alignment with the new TEKS was reflected in the spring 2015 STAAR grades 3–8 mathematics assessments and in the spring 2016 STAAR Algebra I and Algebra II assessments.

2013

In 2013, the 83rd Texas Legislature, Regular Session, enacted HB 5, which reduced the number of STAAR EOC assessments required for graduation from 15 to five: Algebra I, English I, English II, Biology, and U.S. History. The administration of Algebra II and English III was suspended until the 2015–2016 school year and became optional for districts. In addition, the separate reading and writing assessments for the high school English courses were required to be combined into a single assessment for each course with a single reported score. HB 5 removed the requirement to provide a cumulative and minimum score and to include the STAAR EOC assessment results as 15 percent of a course grade.

HB 5 also required changes to the administration of STAAR Alternate, and SB 906 required changes to the performance standards for STAAR Alternate. Based on both bills, STAAR Alternate was redesigned, and STAAR Alternate 2 was administered for the first time in spring 2015.

2015

In 2015, the 84th Texas Legislature passed several bills that affected the assessment program. SB 149 allowed students to qualify for graduation through an individual graduation committee (IGC) beginning in the 2014–2015 school year.

As required by HB 1164 that year, TEA completed a pilot study to examine alternative methods of assessing writing. The pilot study included the collection and scoring of a range of student writing samples produced throughout the school year.

Also passed in 2015, HB 743 required that STAAR be designed so that 85 percent of students taking an assessment in grades 3–5 could complete a test in two hours and 85 percent of students taking an assessment in grades 6–8 could complete the assessment in three hours. In response to HB 743, TEA redesigned the grades 3–8 assessments by reducing the total number of questions and the number of field-test questions on most assessments and redesigned the two-day grades 4 and 7 writing tests as single-day tests that could be completed in a four-hour administration.

The legislature also passed HB 2349, which revised the state's assessment requirements for graduation. Effective beginning with the 2015–2016 school year, a student who earned high school credit for a course for which there was an EOC assessment prior to enrolling in a Texas public school and for which a Texas public school district accepted the credit was not required to take that EOC assessment to receive a Texas diploma. Additionally, HB 2349 required a school district or charter school to report to TEA whether a student assessed with STAAR transferred into a Texas school or district from out of state during the current school year so those students could be excluded in accountability calculations.

2017

In 2017, the SBOE adopted new English and Spanish RLA TEKS, which were implemented in the STAAR RLA assessments beginning in spring 2022. The SBOE also adopted streamlined

TEKS for science, which were first reflected in the STAAR science assessments in December 2018.

2018

In 2018, the SBOE adopted streamlined TEKS for social studies. The streamlined TEKS were first reflected in the 2019–2020 STAAR social studies assessments.

2019

In 2019, the 86th Texas Legislature passed HB 3906, which addressed several components of the assessment program. The bill's key measures included eliminating the STAAR grades 4 and 7 writing assessments, developing a transition plan to administer all STAAR assessments online by the 2022–2023 school year, establishing a cap of no more than 75 percent multiple-choice questions on any STAAR assessment, codifying STAAR Interim Assessments, and developing an integrated formative assessment pilot.

Additionally, HB 1244 required that the STAAR U.S. History EOC assessment include 10 questions randomly selected from the civics test administered by the United States Citizenship and Immigration Services (USCIS). The 10 questions selected were required to align with the TEKS for United States History Studies Since 1877 and were added in the 2019–2020 school year.

2020

In response to the COVID-19 pandemic in spring 2020, the governor used his statutory authority to suspend annual academic assessment requirements for the remainder of the 2019–2020 school year. Therefore, STAAR was not administered in spring or summer 2020, and specific STAAR EOC assessment requirements for graduation were waived for students enrolled in and completing the corresponding course. STAAR Alternate 2 was not administered. Since the administrations of TELPAS and TELPAS Alternate had already begun, completion of these assessments was made optional for districts. TEA received approval from the U.S. Department of Education to waive statewide assessment and accountability requirements for the 2019–2020 school year.

In addition, SSI requirements were waived for the 2020–2021 school year, so retests for STAAR grades 5 and 8 mathematics and reading were not administered.

2021

In 2021, the 87th Texas Legislature, Regular Session, passed HB 4545, which eliminated the grade retention and retesting requirements associated with SSI and established new requirements for accelerated instruction for students who do not pass STAAR.

2023

The 88th Texas Legislature, Regular Session, 2023, passed HB 1225, which permitted districts to provide paper administrations of STAAR to any student whose parent, guardian, or teacher requests it. Requests must be submitted to the district by September 15 for fall administrations and December 1 for spring administrations. The number of students who are administered paper by request is limited to three percent of the total number of students enrolled in the district and is separate and distinct from the students who are eligible to receive a special paper administration of STAAR.

Also passed in 2023, HB 1883 allowed districts to consider the dates of religious holy days likely to be observed by their students when establishing district calendars and days within the testing windows on which students are administered state assessments. For the bill's purposes, holy days were defined as those observed by a religion whose places of worship are exempt from property taxation under Section 11.20 of the Tax Code. HB 1883 required districts to provide alternative dates within the testing window for students who are absent from school on scheduled testing dates to observe a religious holy day. As a result, districts are required to provide make-up testing opportunities for religious holy days observed by students.

TECHNICAL DIGEST 2022–2023

Chapter 2

Building a High-Quality Assessment System

Test Development Activities

Texas educators, including kindergarten through grade 12 classroom teachers, higher education representatives, curriculum specialists, administrators, and Education Service Center (ESC) staff, play a vital role in every phase of the test development process. Thousands of Texas educators have served on one or more of the educator committees involved in the development of the Texas Assessment Program, including STAAR, STAAR Alternate 2, TELPAS, and TELPAS Alternate. These committees are intended to represent the state geographically, ethnically, by gender, and by type and size of school district. While there are slight differences in the processes for developing different assessments, Table 2.1 outlines the procedures used to develop a test framework and provide for ongoing development of test items for the Texas Assessment Program.

Table 2.1. Test Development Process

Step	Process
1	Committees of Texas educators review the state-mandated curriculum, the TEKS, or the ELPS to develop appropriate assessment categories for a specific grade and subject, course, or domain that is assessed. For each grade and subject, course, or domain, educators provide advice on an assessment model that aligns with best practices in classroom instruction.
2	Educator committees work with TEA both to prepare draft test reporting categories and to determine how these categories would best be assessed. These preliminary recommendations are reviewed by classroom teachers, higher education representatives, curriculum specialists, and assessment specialists.
3	A draft of the reporting categories and TEKS or ELPS student expectations to be assessed is refined based on input from Texas educators. TEA begins to gather statewide opportunity-to-learn information.
4	Prototype test questions are written to measure each reporting category and, when necessary, are pilot-tested with Texas students from volunteer classrooms.
5	Educator committees assist in developing guidelines for assessing each reporting category. These guidelines outline the eligible test content and test question formats and include sample items.
6	With educator input, a preliminary test blueprint is developed that sets the number of questions on the test and the number of test questions measuring each reporting category.
7*	Professional item writers develop test items based on the reporting categories, the TEKS or ELPS student expectations, and the item guidelines.
8*	TEA content specialists review and revise the proposed test items.
9*	Item review committees composed of Texas educators review the revised test items to judge the appropriateness of item content and difficulty and to eliminate potential bias.
10*	Test questions are revised again based on input from Texas educator committees and are then field-tested with large representative samples of Texas students.
11*	Technical processes are used to analyze field-test data for reliability, validity, and possible bias.
12*	Data reviews are held to determine whether items are appropriate for inclusion in the item bank from which test forms are built.
13	A final blueprint for each test is developed to establish the number of questions on the test and the number of test questions measuring each reporting category.

Step	Process
14*	All accepted field-test items and data are entered into an item bank. Tests are built from the item bank so that the tests are comparable in difficulty and content from one administration to the next.
15*	Content validation panels composed of university-level experts in each content area review the EOC assessments for accuracy because of the advanced level of content being assessed.
16*	Tests are administered to Texas students.
17*	Stringent quality control (QC) measures are applied to all stages of developing, administering, scoring, and reporting for both online and paper assessments. Test results are reported at the student, campus, district, regional, and state levels.
18	In accordance with state law, the Texas Assessment Program releases tests to the public.
19	In accordance with state law, the commissioner of education uses impact data, study results, and statewide opportunity-to-learn information, along with recommendations from standard-setting panels, to set a passing standard for state assessments.
20	A technical digest is developed and published annually to provide verified technical information about the tests.

*For STAAR, STAAR Alternate 2, and TELPAS, these steps are repeated annually to ensure that tests of the highest quality are developed.

Groups Involved

The entities described below perform crucial functions in the test development process, and their collaborative efforts significantly contribute to the quality of the Texas Assessment Program.

Assessment Development Division

The TEA Assessment Development Division is composed of content experts and psychometricians. The content experts provide content expertise during the item development and test development processes for all statewide assessments. The psychometricians are responsible for ensuring that assessments meet reliability and validity requirements for a sound assessment system.

Student Assessment Division

The TEA Student Assessment Division is responsible for implementing the provisions of federal and state law for the state assessment program. The Student Assessment Division oversees the planning, scheduling, administration, scoring, and reporting of all major assessment activities. TEA staff members in this division conduct QC activities for the administration, scoring, and reporting of the assessment program.

Cambium Assessment, Inc.

Cambium Assessment, Inc. (CAI) is the test administration, scoring, and reporting contractor for the Texas Assessment Program. CAI also serves as the program integration contractor. This

role includes working with Pearson to make sure that the Texas Assessment Program is managed in accordance with TEA requirements.

Pearson

Pearson is TEA's primary item development contractor. Due to the diverse nature of the services required, Pearson employs highly qualified assessment specialists and independent contractors with diverse experience teaching and assessing students.

Texas Educators

When a new assessment is developed, committees of Texas educators review the state-mandated curriculum, help determine appropriate reporting categories, and provide input on the appropriate alignment of the assessment items to the curriculum standards.

Teachers, curriculum specialists, assessment specialists, and administrators review draft reporting categories with the corresponding TEKS or ELPS student expectations. Texas educator committees assist in the review and revision of the eligible TEKS or ELPS documents that outline the student expectations eligible for assessment. TEA staff members then revise and finalize these draft reporting categories and eligible TEKS or ELPS documents based on input from Texas educators.

Following the development of test items by professional item writers, committees of Texas educators review the items to ensure appropriate content alignment and level of difficulty and to eliminate potential bias. Items are revised based on this input and then field-tested.

Item Development and Review

Pearson assumes the major role for STAAR (including STAAR Spanish), STAAR Alternate 2, and TELPAS item development, and TEA personnel are involved throughout the item development process.

Item Guidelines

Item and performance task specifications provide guidance to item writers on how to translate the TEKS or ELPS into assessment items. Item writers strictly follow these guidelines to ensure the accurate measurement of the TEKS or ELPS student expectations. In addition, guidelines for universal design, bias and sensitivity, accessibility and accommodations, and style help item writers and reviewers establish consistency in the development of test items.

Item Writers

Pearson and its subcontractors employ item writers with extensive experience developing items for standardized achievement tests, large-scale criterion-referenced measurements, and English language proficiency tests. These individuals are selected based on their content-area knowledge, their teaching or curriculum development experience in the relevant grades, or their experience teaching EB students or students with special needs.

For each STAAR (including STAAR Spanish), STAAR Alternate 2, and TELPAS assessment, TEA receives an item inventory indicating the number of test items to be developed for each reporting category and TEKS student expectation (for STAAR and STAAR Alternate 2) or ELPS student expectation (for TELPAS). Item inventories are used throughout the item review process. If necessary, Pearson develops additional items to provide the requisite number of items per student expectation.

For TELPAS Alternate, Texas educators developed the Observable Behaviors during a series of TEA-led meetings. Guided by TEA and Pearson staff, the educators created an inventory of items that align to the ELPS and cover the alternate proficiency level descriptors (PLDs).

Training

Pearson provides extensive training for item writers. Before item development begins, Pearson reviews in detail the content expectations and item specifications for the applicable assessment program and discusses the scope of the testing program, security issues, adherence to the measurement specifications, and avoidance of possible economic, regional, cultural, gender, or ethnic bias.

Contractor Review

Pearson staff members who are content experts in the grades and subject areas for which items are developed participate in the review of each set of newly developed items. The review includes a check for content accuracy and item fairness for various demographic groups. Pearson reviewers also consider the alignment between the items and the reporting categories, range of difficulty, clarity, accuracy of correct answers, and plausibility of incorrect answer choices (or distractors) as well as the more global issues of universal design, passage appropriateness, passage difficulty, readability measures, interactions among items, and appropriateness of artwork, graphics, or charts. Pearson editorial staff members examine the items before submission to TEA for review.

TEA Review

TEA Assessment Development Division staff members who are content experts in the grades and subject areas for which items are developed review each item to verify alignment to a particular student expectation in the TEKS or ELPS, grade appropriateness, clarity of wording, content accuracy, plausibility of the distractors, accessibility, and identification of any potential economic, regional, cultural, gender, or ethnic bias. TEA staff members provide edits and meet with Pearson to discuss the progress of the reviews before each item review committee meeting.

Item Review Committee

Each year the TEA Assessment Development Division convenes committees composed of Texas classroom teachers (including general education teachers, special education teachers, and bilingual and English as a second language [ESL] teachers) and curriculum specialists to work with TEA staff in reviewing newly developed test items.

TEA seeks recommendations for item review committee members from superintendents and other district administrators, district curriculum specialists, ESC executive directors and staff members, and staff from other agency divisions. In addition, TEA has developed an educator committee application database where educators can self-nominate to participate on TEA educator committees. Item review committee members are selected based on their established expertise in a content area or in second-language acquisition. Committee members are selected to represent the 20 ESC regions and various types of districts (e.g., urban, suburban, rural, large, small) in Texas, as well as the major ethnic groups in the state.

TEA staff works with Pearson and its subcontractors to train committee members on the proper procedures and criteria for reviewing newly developed items. Committee members judge each item for alignment, appropriateness, adequacy of student preparation, and any potential bias. Committee members discuss each test item and recommend whether it should be field-tested as written or revised, recoded to a different TEKS or ELPS student expectation, or rejected. In their reviews, committee members consider the effect any item may have on various student populations and work toward eliminating potential bias against any group. Table 2.2 shows the guidelines that item review committee members follow.

Table 2.2. Item and Passage Review Guidelines

Category	Guidelines
Reporting Category/Student Expectation Item Match	• The item measures what it is supposed to assess. • The item poses a clearly defined problem or task.
Appropriateness (Interest Level)	• The item or passage is well written and clear. • The point of view is relevant to students taking the test. • The subject matter is of fairly wide interest to students at the grade being tested. • The artwork is clear, correct, and appropriate.
Appropriateness (Format)	• The format is appropriate for the intended grade. • The format is interesting to the student. • The item is formatted so it is not unnecessarily difficult.
Appropriateness (Answer Choices)	• The answer choices are reasonably parallel in structure. • The answer choices are worded clearly and concisely. • The answer choices do not eliminate each other. • There is only one correct answer.
Appropriateness (Difficulty of Distractors)	• Each distractor is plausible. • There is a rationale for each distractor. • Each distractor is relevant to the knowledge and understanding being measured. • Each distractor is at a difficulty level appropriate for both the objective and the intended grade.
Opportunity to Learn	• The item is a good measure of the curriculum. • The item is suitable for the grade or course.

Category	Guidelines
Sensitivity Concerns and Freedom from Bias	• The item or passage does not assume racial, class, or gender values or suggest such stereotypes. • The item does not provide an advantage or disadvantage to any group of students because of their personal characteristics, such as race, gender, socioeconomic status, or religion. • The item or passage avoids needless reference to topics that are extremely controversial or upsetting. • The item or passage addresses sensitive topics in a careful, fair, and balanced way. • The item fairly represents cultural, ethnic, social, and political diversity.

TEA field-tests the recommended items to collect student responses from representative samples across the state. Items rejected by the item review committee are not considered for field testing.

Annual item review committees are not convened for TELPAS Alternate because the TELPAS Alternate Observable Behaviors that were written and revised by educators during the development of the assessment are used every year.

Pilot Testing

The purpose of pilot testing is to gather information about test item prototypes and administration logistics for a new assessment and to refine item development guidelines as needed. Pilot testing can be conducted to accomplish varying objectives. If the purpose is to gather information about test items of differing types and ranges of difficulty, the pilot test might occur before the extensive item development process described above. If the purpose is to gather information about test administration logistics, the pilot test might occur after major item development but before field testing.

Field Testing and Data Review

Field-Test Procedures

Items are field-tested before they are used on an operational test form. Whenever possible, TEA conducts field tests of new items by embedding them in spring operational tests so that the field-test items are randomly distributed across the state. This procedure ensures that a large representative sample of responses is gathered on each item. Experience has shown that embedded field testing yields sufficient data for precise item evaluation and allows for the collection of statistical data on a large number of items in a realistic testing situation. (Performance on field-test items does not affect students' scores on the operational tests.) TEA also periodically conducts stand-alone field tests of new items (e.g., extended constructed-response items) by administering them to a purposefully selected representative student sample. Refer to Chapter 4, "STAAR," for detailed information about stand-alone field testing.

Typically, for STAAR grades 3-8, six field-test questions are embedded in each form for mathematics, RLA, science, and social studies. For spring STAAR EOC assessments, 13 field-

test questions are embedded in each English I and English II form, eight are embedded in each Algebra I and Biology form, and four are embedded in the U.S. History form.

For STAAR Alternate 2, at least four field-test questions are embedded in each form for all grades and subjects and courses assessed.

For TELPAS, at least seven field-test questions are embedded in each form for listening and speaking and for reading and writing.

The Observable Behaviors for TELPAS Alternate are the same each year; therefore, TELPAS Alternate does not include field-test questions.

To ensure that each item is examined for potential ethnic bias, the sample selection is designed so that the proportions of African American and Hispanic students in the samples are representative of their respective total student populations in Texas. Data obtained from the field test include:

- the number of students by ethnicity and gender in each sample;
- the percentage of students choosing each response for multiple-choice questions or obtaining each score point for non-multiple-choice questions;
- the percentage of students, by gender and by ethnicity, choosing each response for multiple-choice questions or obtaining each score point for non-multiple-choice questions;
- point-biserial correlations to determine the relationship between a student's score on a particular test item and the score obtained on the total assessment;
- Rasch statistical indices to determine the relative difficulty of each test item; and
- Mantel-Haenszel statistics for dichotomous items and standardized mean difference (SMD) for constructed-response items to identify, by gender and ethnicity, greater-than-expected differences in group performance on any single item.

Data Review Procedures

After field testing, TEA content development specialists provide feedback to Pearson on each test item and its associated data regarding reporting category and student expectation match, appropriateness, level of difficulty, and potential gender, ethnic, or other bias. They then recommend acceptance or rejection of each field-test item. Items that pass all stages of development—item review, field testing, and data review—are placed in the item bank and become eligible for use on future test forms. Rejected items are marked as such and eliminated from consideration for use on any summative assessment.

Item Bank

The item bank maintained by CAI for the Texas Assessment Program stores each test item, accompanying artwork, and item data such as the unique item number (UIN), grade and subject or course, reporting category, TEKS or ELPS student expectation measured, dates the item was

administered, and item statistics. The item bank also contains information obtained during data review meetings that specifies whether a test item is acceptable for use. During the test construction process, TEA, CAI, and Pearson use item statistics and other item information to maintain consistent test difficulty and adjust tests for content coverage and balance.

Test Construction

Each grade and subject and course assessment is based on a specific test blueprint that guides how each test is constructed. Test blueprints delineate the number of items and points from each reporting category that will appear on a given test. STAAR, including STAAR Spanish, focuses on the TEKS that are most critical by incorporating readiness and supporting standards into the test blueprints. Readiness standards are emphasized annually; supporting standards, while eligible for assessment as an important part of instruction, may not be tested each year. All decisions about the relative emphasis of each reporting category are based on feedback from Texas educators and are indicated in the assessed curriculum and blueprint documents found on the STAAR Resources webpage. General characteristics of readiness and supporting standards are shown in Table 2.3.

Table 2.3. Comparison of Readiness and Supporting Standards

Readiness Standards	Supporting Standards
• are essential for success in the current grade or course • are important for preparedness for the next grade or course • support college and career readiness • necessitate in-depth instruction • address broad and deep ideas	• may be introduced in the current grade or course and emphasized in a subsequent year • may be reinforced in the current grade or course and emphasized in a previous year • play a role in preparing students for the next grade or course, but not a central role • address more narrowly defined ideas

STAAR Alternate 2 provides access to the grade-level TEKS through vertical alignment and curriculum framework documents. These documents, along with the blueprints for STAAR Alternate 2, can be found on the STAAR Alternate 2 Resources webpage.

TELPAS is based on the ELPS. TELPAS assessed curriculum and blueprints can be found on the TELPAS Resources webpage.

Overall, each STAAR, STAAR Alternate 2, and TELPAS assessment is designed to reflect:

- problem-solving and complex thinking skills,
- the range of content represented in the TEKS or ELPS,
- the level of difficulty of the skills represented in the TEKS or the range of English proficiency represented in the ELPS, and
- the application of content and skills in different contexts, both familiar and unfamiliar.

Tests are constructed from the bank of items determined to be acceptable after data review. Field-test data are used to place the item difficulty values on a common Rasch scale. This scale allows for the comparison of the difficulty of each item with that of all other items in the bank.

Consequently, items are selected not only to meet sound content and test construction practices but also to ensure that tests are approximately comparable in difficulty from one administration to the next. Refer to Chapter 3, “Standard Technical Processes,” for detailed information about Rasch scaling.

Tests are constructed to meet a blueprint for the required number of items and points on the overall test and for each reporting category. For STAAR, including STAAR Spanish, blueprints indicate the number of dichotomous and polytomous items and the number of extended constructed-response items. In addition, blueprints for STAAR, including STAAR Spanish, list a specific number of readiness and supporting standards. Items that test each reporting category are included for every administration, but the array of TEKS student expectations represented might vary from one administration to the next. Although the STAAR, including STAAR Spanish, assessments are constructed to emphasize the readiness standards, they still measure a variety of TEKS student expectations and represent the range of content eligible for each reporting category being assessed.

Before test construction is completed for the STAAR EOC assessments, panels made up of university-level experts review the content to ensure that each assessment is of the highest quality. A content-validation review is critical to the development of the EOC assessments because of the advanced level of content being assessed. Committee members note any issues of concern, and when necessary, replacement items are chosen and reviewed. STAAR Alternate 2, TELPAS, and TELPAS Alternate do not have content validation reviews.

After test construction for STAAR, including STAAR Spanish, is complete, TEA and Pearson work together to develop content and language supports for students who meet eligibility criteria. Content and language supports allow for various types of assistance (e.g., scaffolded directions, assistance with tracking, graphic organizers, simplified language, graphic representations of vocabulary and concepts) to support a student’s understanding of passages, test questions, and answer choices and are mainly in the form of pop-ups, rollovers, prereading text, and supplementary materials. These embedded supports are available for all online STAAR test forms.

For STAAR Alternate 2, accommodations and supports are included as part of the test design. For TELPAS, embedded accommodations are available on writing questions for students who meet eligibility criteria. Embedded accommodations are not provided on TELPAS Alternate.

All test content, including embedded supports, is reviewed and approved by TEA, after which the assessments are ready to be administered.

TELPAS Alternate is a holistic inventory that contains the same Observable Behaviors every year. Thus, there is no annual test construction process. Blueprints for TELPAS Alternate are available on the TELPAS Alternate Resources webpage.

Security

TEA prioritizes test security and confidentiality for all aspects of the Texas Assessment Program, from development and construction to administration and reporting. TEA ensures that every allegation of cheating or breach of confidentiality is properly investigated.

Maintaining the security and confidentiality of the Texas Assessment Program is critical for ensuring valid test scores and providing standardized and comparable testing opportunities for all students. TEA has implemented numerous measures to strengthen test security and confidentiality, including the development of various administrative procedures and manuals to train and support district testing personnel.

Test Administration Materials

The *District and Campus Coordinator Resources* and assessment-specific test administrator manuals provide guidelines on training testing personnel, administering tests, creating secure testing environments, and properly storing test materials. They also instruct testing personnel on how to report any confirmed or alleged testing irregularities that might have occurred. The manuals include information on the test security oaths that all personnel with access to secure test materials are required to sign as well as specific details about the possible penalties for violating test procedures. In addition, Texas Administrative Code (TAC) §101.3031 addresses test administration procedures and includes specific language detailing the requirements of school districts and charter schools to maintain security and confidentiality of assessment instruments, including a list of violations and their consequences.

Training

TEA training materials cover test administration best practices, including test security issues. All district and campus personnel who participate in state-mandated testing or handle secure test materials and content are required to be trained in test security and administration procedures. In addition to this required training, TEA provides optional online training modules. It is strongly recommended that districts and charter schools use these modules to help supplement the mandatory training required of all personnel involved in testing. Trainings are posted on the Learning Management System (LMS).

Security Violations

In accordance with test administration procedures, any person who violates, solicits another to violate, or assists in the violation of test security or confidentiality, and any person who fails to report such a violation, could be penalized. An educator involved with a testing irregularity might face:

- restrictions on the issuance, renewal, or holding of a Texas educator certificate, either indefinitely or for a set term;
- issuance of an inscribed or non-inscribed reprimand;
- suspension of a Texas educator certificate for a set term; or
- revocation or cancellation of a Texas educator certificate without opportunity for reapplication for a set term or permanently.

Students involved in a violation of test security could have their test results invalidated.

Incident Tracking

TEA regularly monitors and tracks testing irregularities and reviews all incidents reported from districts and campuses.

Processes that have been developed to assist in test administration and security include:

- an internal database that allows TEA to track reported testing irregularities and security violations,
- a system to review and respond to each reported testing irregularity, and
- a resolution process that tracks missing secure test materials after each administration and provides suggested best practices that districts can implement for proper handling and return of secure materials.

Quality-Control Procedures

The data provided by the Texas Assessment Program plays an important role in decision-making about student performance and public education accountability. Individual student test scores are used for accelerated instruction and graduation. In addition, the aggregated student performance results from the Texas Assessment Program are a major component of state and federal accountability systems used to rate individual public schools and school districts in Texas. The data are also used in education research and in the establishment of public policy. Therefore, it is essential that assessments are scored correctly and that scores are reported accurately.

TEA uses a comprehensive QC system to review work produced by the testing contractors. This section describes the procedures used to confirm the validity of scoring, reporting, and test development.

Data and Report Processing

TEA undertakes an extensive and comprehensive QC process to verify the quality and accuracy of final Texas Assessment Program results before reporting them. Begun months in advance of an assessment date, the QC process involves internal steps taken by CAI and the implementation of a joint process supported by TEA. This QC process is applied to every operational assessment administered in the school year.

CAI executes an internal QC system for the reporting of test results. QC at the unit level confirms that software modules associated with various business processes—such as online test delivery, scoring, and reporting—are properly developed and that they operate to meet program requirements. Performed by a group that is independent from the software development group, system QC confirms that all the modules work together so that outputs from one module in the system match the proper inputs for the next module. This process allows for independent verification and interpretation of project requirements. Once the independent testing group has completed and approved the test, the system is moved into production mode.

The joint QC process involves a complete scoring and reporting test run. For each test administration, TEA prepares response data for thousands of hypothetical students who serve as test cases. The test-run processing includes scoring the responses and generating student- and district-level reports and data files, and TEA independently verifies information during every step. Reports are not sent to districts until all discrepancies in the QC data set are resolved and the reports generated by TEA and the contractor match. Details of the QC process can be found in Appendix A.

Technical Processing

In addition to the processing of data and generation of reports, psychometric or technical processing of the data also occurs before and after each test administration and includes additional QC measures.

Each technical procedure requires calculations or transformations of the data to be completed and verified by multiple psychometricians and testing experts at CAI and Pearson; TEA also verifies these calculations.

Each year's calculations are also compared to historical values to further validate the reasonableness of the results. Comparisons of technical procedures and assessment results from year to year help verify the quality of the assessments and inform TEA of the program's impact on student achievement.

For more information about the standard technical processes of the Texas Assessment Program, see Chapter 3, "Standard Technical Processes."

Performance Assessments

STAAR, including STAAR Spanish, and TELPAS contain constructed-response items, which require scoring by trained human raters, on the following operational assessments:

- STAAR grades 3–8 RLA, grades 5 and 8 science, grade 8 social studies, English I, English II, Biology, and U.S. History
- STAAR Spanish grades 3–5 RLA and grade 5 science
- TELPAS grades 2–12 speaking and grades 2–12 writing

STAAR, including STAAR Spanish, uses extended constructed responses, which measure the student's ability to synthesize the component skills of writing; that is, the extended constructed-response task requires the student to express ideas effectively in writing for a specified purpose.

The types of writing required vary by grade and subject or course and represent the learning progression evident in the TEKS. RLA assessments include short constructed-response questions as well as an extended constructed-response question at every grade level. Science and social studies assessments include short constructed-response questions.

Extended constructed responses for STAAR, including STAAR Spanish, are evaluated using a holistic scoring process, meaning that the student response is evaluated as a whole according to pre-established criteria. These criteria, which are explained in detail in the scoring rubrics for

each type of writing, are used to determine the effectiveness, and thus the score, of each response.

The 5-point rubric for extended constructed responses includes two main components: 1) development and organization of ideas and 2) conventions. A student response may receive up to 3 points for development and organization of ideas and up to 2 points for the use of writing conventions. The constructed responses are scored independently by two scorers, and the scores are added to create a final score; therefore, a student may receive up to 10 points for his or her essay. Short constructed responses in the reading domain are scored using a 2-point prompt-specific rubric, and short constructed responses in the writing domain are scored using a 1-point rubric. Responses deemed nonscorable are assigned a condition code and receive 0 points. The STAAR writing rubrics for extended constructed responses can be found on the STAAR Resources webpage. Rubrics for the short constructed responses are included in the STAAR constructed-response scoring guides found on the same webpage.

TELPAS grades 2–12 reading and writing assessments include constructed-response and sentence-rewrite items, the scoring rubrics for which are found on the TELPAS Resources webpage. TELPAS writing items are evaluated using a holistic scoring process. Sentence-rewrite items receive a score of 0 or 1 based on the criteria defined in the rubric. Scorers use a 4-point writing rubric to evaluate constructed responses at grades 2 and 3, with two scorers independently evaluating student responses and those scores added together to calculate the students' raw score (from 2 to 8). For grades 4 through 12, a 12-point rubric is used to evaluate constructed responses, which are scored for three traits: vocabulary, usage, and completeness. Each trait is worth 1 to 4 points, and trait scores are added together to calculate a raw score of 3 to 12 points. A second scorer scores 25 percent of constructed-response items and 5 percent of sentence-rewrite items. These QC measures ensure the validity and reliability of scores.

The TELPAS speaking assessment consists of prompts that elicit student speaking responses recorded using a headset with a microphone. Speaking responses are scored according to a 2-point or 4-point rubric, depending on the item type. An automated scoring engine scores all student responses for TELPAS speaking. To ensure continued validity, reliability, and calibration of the assessment scoring process, a second scorer scores 10 percent of engine-scored responses. Data from these two methods are continuously compared to ensure the process is reliable.

Human scorers train the automated scoring engine by assigning points to the responses gathered during field testing. For operational items, human scorers score any responses that are considered uncertain cases or are part of a backread to examine the inter-rater reliability of the automated scoring engine. The TELPAS 2-point and 4-point speaking rubrics can be found on the TELPAS Resources webpage. Human scoring also takes place for responses that the automated engine identifies as nonscorable. These responses often have a unique characteristic—including, for example, background noise, mumbled or unclear speech, or low volume—that makes them appropriate for scoring by a human scorer. All scorers undergo the same extensive training process using the same materials and rubrics. Refer to Chapter 6, "TELPAS," for detailed information about the TELPAS speaking scoring process.

Scoring Staff

All test scorers have at least a four-year college degree and must undergo rigorous TEA-approved training before they are allowed to begin scoring. As part of this training, applicants must review an anchor set, score practice sets, and pass qualification testing. Scorers are monitored daily to produce scores that are accurate and reliable.

Pearson's training and monitoring of scorer performance is conducted by content specialists, supervisors, directors, and program managers, all of whom have demonstrated expertise with scoring constructed responses. Content specialists build the training materials from field-test responses to represent a full range of scores and train scoring leadership on both content and job expectations before scorer training. During operational scoring sessions, supervisors guide, support, and monitor scorers, and directors guide, support, and monitor supervisors; both roles share responsibility for monitoring and managing scoring quality by answering scorers' questions and reviewing scoring reports. Supervisors and directors apply all condition codes and reach out to content specialists when they need guidance. Program managers monitor all aspects of scoring for STAAR, including STAAR Spanish, and TELPAS, specify the configuration of training materials, and oversee the schedule and process for performing the work.

Distributed Scoring

Distributed scoring is used for STAAR, including STAAR Spanish, to allow scorers to participate in the scoring process from any location, provided they qualify and meet strict requirements. Distributed scoring is a secure, web-based model that incorporates several innovations and includes the following benefits:

- The number of scorers available locally can be augmented by other highly credentialed scorers from across the state and country.
- More teachers across the state can participate in the scoring process.
- Paper handling and associated costs and risks are reduced.
- Scorers are trained and qualified using comprehensive, self-paced online training modules that allow them to manage their training more efficiently.
- Distributed scoring uses state-of-the-art approaches to monitor scoring quality and communicate feedback to distributed scorers.

The ePEN Scoring System

STAAR, including STAAR Spanish, and TELPAS constructed responses are scored using the Pearson ePEN system. Scorers have access to TEA-approved rubrics and anchor papers during training, qualification, and operational scoring, and once they have completed training and qualification, they have secure access to students' constructed responses. The ePEN response viewer renders scanned images of students' constructed responses. Scorers can adjust contrast, color, and magnification to improve readability and reduce fatigue.

All constructed responses from a particular student and test are linked throughout Pearson scoring and reporting processes via a unique identifier. To protect student anonymity and prevent potential bias, student identifiers and other demographic information are not visible to scorers in ePEN.

Responses are grouped by grade and subject or course and are stored on the ePEN server. As scorers score the responses, more responses are routed into their scoring queues. Each scorer independently reads a response and selects a score from a menu on the computer screen. Scoring supervisors, scoring directors, and content specialists can identify which scorer reads each response.

Although the automated scoring engine scores most TELPAS speaking responses, sometimes responses require review by a human scorer. Pearson scores these items through ePEN, providing secure access to the students' audio files and scoring reports for content supervisory staff. Each scorer independently listens to a response and selects the appropriate score in the scoring grid. The system provides numerous tools and reports to help supervisory staff monitor scoring, and the rubric and training can be reinforced through qualification sets delivered regularly or when needed to address a scoring issue.

Scorer Training Process

All scorers who work on the STAAR and TELPAS performance task scoring projects receive extensive training through Pearson's online modules. This training covers the materials associated with the performance questions for each assessment and includes orientation in the ePEN system. Scorers receive training on the scoring guide that provides the rubric and anchor sets of each rubric score point for a particular assessment item. Additionally, scorers score training sets and have an opportunity to explain and discuss the scores. Scorers are required to demonstrate a complete understanding of the rubrics and to pass a qualification set before being allowed to score operational student responses.

Extended Constructed Responses

Training materials are selected to clearly differentiate student performance at the different rubric score points and to help scorers learn the difference between score points. To help scorers refine their understanding of differences between adjacent score points, training materials also include responses determined to be on the borderline between two adjacent score points. Supervisors are available during scorer training to assist and answer questions.

Once scorers complete the training sets, they are administered qualification sets of student responses. These student responses have already been scored by TEA and Pearson staff, and scorers must accurately assign scores to the student responses. Scorers are given two opportunities to qualify, with a different set of responses in each set. Any scorer who cannot meet the standards established by TEA and Pearson is dismissed from scoring.

Ongoing Training

After initial training, ongoing training is available to ensure scoring consistency and high scorer agreement. Supervisors and scoring directors monitor scoring and provide mentoring continually

during operational scoring. The ePEN scoring system includes a comprehensive set of scoring and monitoring tools that help identify areas for additional training.

Scoring Process

STAAR, including STAAR Spanish, constructed responses are scored using a holistic approach in which scores can be exact (scorer 1 and scorer 2 agree) or adjacent (scores by scorer 1 and scorer 2 differ by no more than 1 point). During scoring, two scorers independently assign a score from 1 to 4 to each student response. The scores are summed and weighted, if applicable, and the performance is reported to districts on both the STAAR Student Report Card for individuals and on the Constructed Responses Summary Report for campuses and districts.

In instances in which the scores are discrepant (i.e., scores from scorer 1 and scorer 2 differ by more than 1 point), the student response is routed to a resolution queue. A supervisor or scoring director reviews the response and applies a third score, invalidating the two initial scores. This score is then doubled and becomes the reported score.

Throughout scoring, TEA staff members are consulted on responses that are highly unusual or require a policy decision from TEA.

Nonscorable Responses

Only a scoring director may determine if a constructed response is nonscorable. Before a constructed response can be given a nonscorable designation, the supervisor or scoring director thoroughly reviews the response. If the scoring director determines that the response is scorable, it is assigned a score and routed to a second content scoring leader. If the scoring director determines that the response is nonscorable, a nonscorable code is applied, and the response is routed to another scoring director for confirmation. While the response is under review, it is held in a review queue that prevents it from being distributed to other scorers.

Monitoring of Scorer Quality

Scorers can defer student responses to their supervisor, who will provide feedback on how to score the response or pass the question to the scoring director. This allows scorers to receive feedback regularly on their performance. If a scorer is identified as having difficulty applying the criteria, the responses they scored are invalidated and rescored, and that scorer must then complete targeted qualification training. Any scorer who cannot pass the targeted qualification training set is dismissed from scoring.

Validity responses are student responses that have already been assigned a score during anchor approval meetings and are presented to scorers throughout the operational scoring process to monitor their scoring quality. TEA approves all validity responses before they are introduced into the scoring systems. Indistinguishable from operational responses, validity responses are inserted randomly into the scoring queue. Scorers' accuracy is evaluated based on how often their scores on validity responses agree with the scores that have been assigned to them.

For TELPAS, a supervisor using ePEN can back-listen to responses scored and send that scorer feedback through the ePEN messaging system. Scorers can also submit responses for

review so that a scoring supervisor or scoring director can listen and provide feedback. Validity responses with TEA-approved scores are delivered randomly to scorers throughout the project. Scorers failing to meet the standard for validity after remediation are dismissed from the project, and their work is reset and scored again.

Anchor Sets

In addition to the scoring that field-test scorers perform, TEA and Pearson staff members independently score samples of the field-test responses that will be used on the operational assessments. TEA and Pearson content and management staff and Texas educators participate in a series of anchor approval meetings to analyze these responses and assign scores. Assessment specialists select responses from the anchor approval meetings to be included in each scoring guide. Scoring directors then assign the remaining pre-scored responses from the meetings to training sets and qualifying sets for use in scorer training. Educators assist in the review and make recommendations to reach a consensus on the scores. Before scoring, TEA staff members review and approve all scoring guides and training sets.

Score Reliability and Validity Information

TEA regularly reports on the reliability and validity of the performance scoring process. Reliability is expressed in terms of scorer agreement (percentage of exact agreement between scorers' scores) and correlation between first and second scores. Validity is assessed by the inclusion of validity responses throughout the operational scoring process and expressed in terms of exact agreement between the score assigned by the scorer and the score assigned by TEA and Pearson.

Appeals

If a district has questions about the score assigned to a response, a rescore can be requested in TIDE. If the score changes, CAI provides rescore results by posting an updated student report card to the TIDE secure inbox and to the Family Portal. In instances where a rescore improves scores, the fee associated with the rescore request is waived. If the score does not change, the district pays the associated fee. If a district files a formal appeal with TEA related to scores reported on the consolidated accountability file, an analysis of the response in question is provided to explain the final outcome of the appeal and whether the score was changed.

TECHNICAL DIGEST 2022–2023

Chapter 3

Standard Technical Processes

Overview

The Standards for Educational and Psychological Testing, developed jointly by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME), provides guidelines for evaluating the quality of testing practices. TEA applies these standards to all aspects of the Texas Assessment Program to ensure its assessments are technically defensible and appropriate for the purposes for which they are used.

To promote fairness, accuracy, reliability, and validity in the Texas Assessment Program, TEA uses the following technical concepts, which are discussed in detail in this chapter:

- performance standards
- item analysis
- scaling
- equating
- reliability
- validity
- measures of student progress
- sampling

Program-specific technical processes are covered in subsequent chapters.

Technical Details and Procedures

Performance Standards

A critical aspect of any statewide testing program is the establishment of performance standards that provide a frame of reference for interpreting test scores. 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 demonstrate on an assessment.

STAAR, including STAAR Spanish, has three cut scores that identify the following four performance levels:

- Did Not Meet Grade Level
- Approaches Grade Level

- Meets Grade Level
- Masters Grade Level

STAAR Alternate 2 has two cut scores that identify the following three performance levels:

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

TELPAS has three cut scores that identify the following four English proficiency levels:

- Beginning
- Intermediate
- Advanced
- Advanced High

TELPAS Alternate has four cut scores that identify the following 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. The standard-setting framework and process for the STAAR, STAAR Alternate 2, TELPAS, and TELPAS Alternate programs are described below.

Standard Setting for STAAR

Performance standards for STAAR were originally established in 2012 using an evidence-based standard-setting approach (O'Malley, Keng, & Miles, 2012). Standard setting for STAAR involved a process of combining policy considerations, the PLDs derived from 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 largely of diverse groups of educators, considered the interaction of all these elements for each STAAR assessment.

In 2014, standard-setting committees reset performance standards for the STAAR English I, English II, and English III assessments, which combined the reading and writing components into a single assessment. In 2015, standard-setting committees reset the STAAR grades 3–8 mathematics performance standards due to changes in the TEKS. With the STAAR redesign in the 2022–2023 school year, performance standards for all STAAR assessments were updated using the Modified Angoff (Angoff, 1971) standard-setting method.

Refer to the *STAAR Standard Setting Technical Report* available on the Assessment Reports and Studies webpage for more detailed information.

Standard Setting for STAAR Alternate 2

Performance standards for STAAR Alternate 2 were originally established in spring 2015 using an evidence-based standard-setting approach (O'Malley, Keng, & Miles, 2012). This 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 aligned with student performance on other assessments.

Due to changes in the STAAR Alternate 2 RLA assessments, performance standards for these assessments were reset in spring 2023 using the Modified Angoff (Angoff, 1971) method. This content- and item-based method led panelists through a standardized process in which they considered student expectations, as defined by the PLDs, and the individual items that were administered to students to recommend cut scores for each performance level.

Refer to the *STAAR Alternate 2 Standard Setting Technical Report* available on the Assessment Reports and Studies webpage for more detailed information.

Standard Setting for TELPAS

TELPAS grades 2–12 reading proficiency standards were originally established in 2008. The method consisted of a two-phase process in which an internal work group made initial recommendations and then an external committee of state educators recommended specific cut scores after reviewing the recommendations, the test forms on which the recommendations were based, and impact data.

During the 2013–2014 school year, TEA convened educator committees to review the proficiency standards for TELPAS grades 2–12 reading to align the program with STAAR. TEA used an evidence-based standard-setting approach to determine the cut scores. 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 updated proficiency standards.

The TELPAS grades 2–12 reading test redesign in spring 2018 and the first-time administration of an online test for the grades 2–12 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 the Extended Modified Yes/No Angoff method (Davis & Moyer, 2015; Plake, Ferdous, Impara, & Buckendahl, 2005).

Proficiency standards were established for TELPAS grades 2–12 writing in spring 2012 as the assessment transitioned to a standardized online assessment. The standard-setting methodology used was a modification of the well-known Body of Work method (Kingston, Kahl, Sweeney, & Bay, 2001; Kingston & Tiemann, 2012), which has been used to recommend proficiency level cut scores for various large-scale state assessments.

Refer to the *TELPAS Standard Setting Technical Report* available on the Assessment Reports and Studies webpage for more detailed information.

Standard Setting for TELPAS Alternate

The proficiency standards for TELPAS Alternate were established in 2019. To establish the proficiency levels for each domain, a test-centered, criterion-referenced method was used to guide the panelists. The implemented procedure 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.

Refer to the *TELPAS Alternate Standard Setting Technical Report* available on the Assessment Reports and Studies webpage for more detailed information.

Item Analysis

Statistical analyses are conducted on student performance data 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. Several statistical analyses, based on both classical test theory and item response theory (IRT), are used to analyze the data collected annually for operational items. Item analyses are also conducted annually for the purpose of reviewing the quality of newly field-tested items to help determine which items may be included as operational items in future test administrations. Statistics generated after each administration of STAAR (including STAAR Spanish), STAAR Alternate 2, and TELPAS include *p*-value, point-biserial correlation, Rasch item difficulty, Rasch fit, and response or score point distribution. In addition, group difference analyses, also known as differential item functioning (DIF), are conducted using the Mantel-Haenszel (MH) alpha and ABC DIF classification.

p-Value

The *p*-value indicates the proportion of the total group of students answering a multiple-choice or dichotomous item correctly. For polytomous items, the *p*-value indicates the average score obtained by students divided by the number of points possible. An item's *p*-value shows how difficult the item was for the students who were administered the item. An item with a high *p*-value, such as 0.90, is a relatively easy item. An item with a low *p*-value, such as 0.30, is a relatively difficult item.

Point-Biserial Correlation

The point-biserial correlation describes the relationship between a student's performance on the item 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 answered the item incorrectly. 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 perform well (or poorly), that item does not provide useful information for distinguishing between those students with higher performance from those students with lower performance on the entire assessment.

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 were administered the item, Rasch item difficulties can be compared across test forms and 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 an 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, high-scoring students answering easy items incorrectly). In general, a Rasch fit value lower than 0.7 or greater than 1.3 may indicate that the item fits the Rasch model poorly.

Response or Score Point Distribution

The response or 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 dichotomous item, or the percentage of students who received each of the score points for a polytomous item. Response or score point distributions are provided for the entire group of students and for various demographic groups (e.g., gender, ethnicity for STAAR) or for proficiency level groups (e.g., Beginning, Intermediate, Advanced, 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, White) performed on an item. Such analyses help identify items on which a group of students performed unexpectedly well or poorly. Both the MH alpha and the ABC DIF classification, also known as the Educational

Testing Service (ETS) DIF classification (Petersen, 1987; Zieky, 1993), are used for the Texas Assessment Program.

It should be noted that DIF analyses serve to merely 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.

Mantel-Haenszel Alpha

To calculate the MH 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 the MH alpha is 1, students of similar proficiency, regardless of group membership (e.g., males, 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, an 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 the 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).

Scaling

Scaling associates numbers with characteristics of interest to provide information about measurable quantities for those characteristics. For example, temperature can be described using the Fahrenheit scale or the Celsius scale. Different numbers refer to the same temperature, but they describe it using different scales. Similarly, test scores can also be reported using more than one scale.

The number of items that a student answers correctly on a given test is known as the raw score, and this raw score is interpreted in terms of the specific set of answered test questions. In general, raw scores from different test forms are not comparable. For example, suppose there are two forms of an assessment that are not equally difficult: Form A is harder than Form B. One student takes Form A and earns a raw score of 34 out of 50, while another takes Form B

and also earns a raw score of 34 out of 50. Here, the first student's performance on the harder test reflects greater achievement than the second student's performance on the easier one, 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 be compared directly across test administrations using raw scores. To facilitate comparisons, raw scores from different test forms and administrations are placed onto a common scale resulting in scale scores. Unlike raw scores, scale scores allow for direct comparisons of student performance across separate test forms and different test administrations. A scale score considers the difficulty level of the specific set of questions on a test form, and it describes students' performance relative to each other and relative to the performance standards across separate test forms.

Three scales underlie the STAAR (including STAAR Spanish), STAAR Alternate 2, TELPAS, and TELPAS Alternate assessments: the raw score scale, the Rasch scale, and the reporting scale. The scales are defined as follows:

- The raw score scale is defined as the number of items answered correctly, regardless of difficulty.
- 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 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 item difficulty. The reported scale scores are comparable and maintain performance standards across test forms and 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 demonstrates, while 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 simultaneously locate students and items on a scale based on how students respond to the items and how difficult the items are and can be thought of as a combination of subject-centered and stimulus-centered approaches; therefore, they are the most complex approaches.

TEA scales 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, the Texas Assessment Program uses 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

Test items (whether dichotomous or polytomous) for the Texas Assessment Program are scaled and equated using the RPCM. The RPCM is an extension of the Rasch one-parameter IRT model attributed to Georg Rasch (1966), and extended by Wright and Stone (1979), Masters (1982), Wright and Masters (1982), and Linacre (2018). The RPCM was selected because of its flexibility in accommodating dichotomous or polytomous items. 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 equation:

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

where M_i is the number of score categories of item i, θ is a student's proficiency (ability) score, m=(0, 1,… 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 θ–δ_{i0} =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 dichotomous technology-enhanced items, 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 respond to

correctly or incorrectly based on the student's proficiency. This facet 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 in test forms given across different administrations. Because both pre-equated and post-equated item difficulty estimates are available, any drift in scale or difficulty can be quantified.

The Texas Assessment Program uses two types of scale scores—horizontal and vertical. Horizontal scale scores are used for STAAR grades 5 and 8 science (including STAAR Spanish grade 5 science), STAAR grade 8 social studies, STAAR EOC assessments, STAAR Alternate 2, TELPAS, and TELPAS Alternate. Vertical scale scores are used for STAAR grades 3–8 mathematics, STAAR grades 3–8 RLA, STAAR Spanish grades 3–5 mathematics, and STAAR Spanish grades 3–5 RLA.

Horizontal Scaling

Scale scores (SS_θ) for the Texas Assessment Program 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 scaling constants, and they are derived using a method described by Kolen and Brennan (2004). For STAAR grades 5 and 8 science, STAAR grade 8 social studies, STAAR EOC assessments, TELPAS, and TELPAS Alternate, two scale score values at two specific standards were established in advance. These standards are Meets Grade Level and Approaches Grade Level for STAAR, Advanced and Advanced High for TELPAS, and Early Independence and Developing Independence for TELPAS Alternate. The A scaling constant is calculated as follows:

$$A = \frac{SS_2 - SS_1}{\theta_2 - \theta_1} \quad (3)$$

In equation 3, SS_2 represents the desired scale score at the higher of the two standards desired to be fixed, and SS_1 represents the desired scale score at the lower standard, where θ_2 and θ_1

are the corresponding Rasch-based proficiency estimates at the selected standards. The *B* scaling constant is calculated as follows:

$$B = SS_2 - A \times \theta_2 \quad (4)$$

For STAAR Alternate 2, the scale score value at the passing standard (Satisfactory) and the standard deviation of the reportable scale score were established in advance. The *A* scaling constant is calculated as follows:

$$A = \frac{\sigma_{SS}}{\sigma_\theta} \quad (5)$$

In equation 5, σ_{SS} represents the desired standard deviation of the scale score, and σ_θ represents the standard deviation of the Rasch-based θ values among a sample group. For all STAAR Alternate 2 assessments except grades 3–8 RLA, English I, and English II, the horizontal scales sample group comprised all students who took that assessment in spring 2015. For grades 3–8 RLA, English I, and English II, the sample group comprised all students in the spring 2023 administration. The *B* scaling constant is calculated as follows:

$$B = SS_{Satisfactory} - A \times \theta_{Satisfactory} \quad (6)$$

In equation 6, $SS_{Satisfactory}$ and $\theta_{Satisfactory}$ represent the selected scale score to be fixed at the passing standard and its corresponding Rasch-based proficiency estimate, respectively.

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 content areas but not across all assessments. Similarly, the STAAR Alternate 2 Level II: Satisfactory performance standards are also unique for each assessment; $\theta_{Level\ II}$ varies across assessments, and $SS_{Level\ II}$ and σ_{SS} are set to be consistent within 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. Such studies were conducted for STAAR grades 3–8 mathematics and RLA and STAAR Spanish grades 3–5 RLA in spring 2023. For these studies, embedded field-test positions (refer to the Field-Test Equating section) were also used to administer vertical linking items. The studies assumed a common-item nonequivalent groups design (refer to the Equating section) 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.

Similar to the horizontally scaled assessments, vertically scaled 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:

$$SS_\theta = A \times \left(\theta + V_g\right) + B, \quad (7)$$

where SS_θ is the scale score for a Rasch proficiency score (θ). The scaling constants *A* and *B* in equation 7 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) is fixed only for one of the assessments in the vertical scale (e.g., STAAR grade 3 mathematics for the STAAR mathematics vertical scale), and the standard deviation is calculated using the calibration sample of the base grade. The *A* scaling constant is calculated as follows:

$$A = \frac{\sigma_{SS}}{\sigma_\theta} \quad (8)$$

In equation 8, 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 the calibration sample in the base grade. The STAAR grades 3–8 mathematics, grades 3–8 RLA, and Spanish grades 3–5 RLA vertical scale sample group comprised all students who took a test form with embedded vertical scale items in spring 2023. Like field-test items, these vertical scale items are not used to calculate student scores.

The *B* scaling constant is calculated as follows:

$$B = SS_{Approaches} - \frac{\sigma_{SS}}{\sigma_\theta} \times \theta_{Approaches} \quad (9)$$

In equation 9, $SS_{Approaches}$ represents the desired scale score at the STAAR Approaches Grade Level cut score for the final assessment in the vertical scale, and $\theta_{Approaches}$ represents the approved STAAR Approaches Grade Level performance standard in Rasch units for the final assessment in the vertical scale.

Equating

Used in conjunction with the scaling process, equating is the process that considers the differences in difficulty across test forms and administrations and allows scores to be placed onto a common scale. The Texas Assessment Program uses the common-item nonequivalent groups design to equate most assessments 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 student sample 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 for the Texas Assessment Program are described in this section. Refer to Kolen and Brennan (2004) or Petersen, Kolen, and Hoover (1989) for a more detailed explanation of equating designs and methods.

With the spring 2023 administrations, new scales were established for STAAR; STAAR Alternate 2 grades 3–8 RLA, English I, and English II; and TELPAS writing. When new scales are created, there is no equating methodology employed since there is no link to a previous scale necessary during calibrations. All items are freely calibrated to establish the new scale.

During test construction, pre-equating based on the previous scales for STAAR and STAAR Alternate 2 and the field-test results for TELPAS were used as a guide. However, these scales were not used for reporting. Thus, for these assessments, equating was mainly employed for field-test analyses.

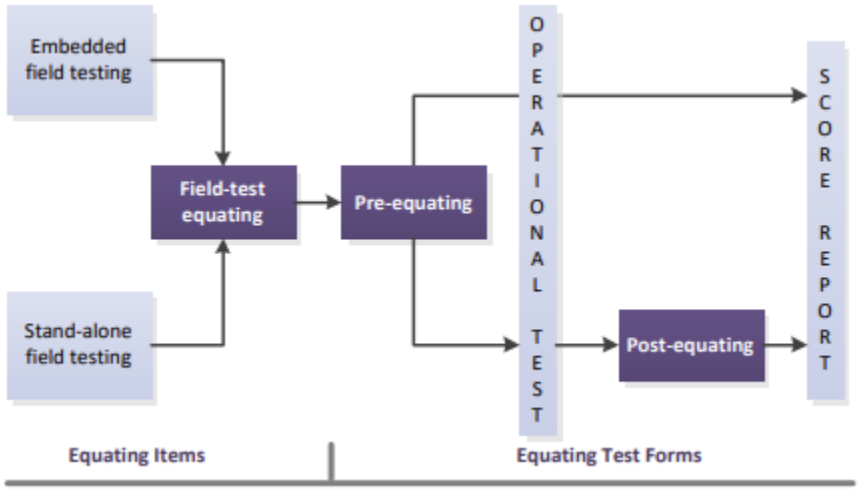
Types of Equating

The following are the three types of equating used in the item and test development process:

1. pre-equating test forms that are under construction
2. post-equating operational test forms after administration
3. equating field-test items after administration

One or more of these three types of equating is used on each component of the Texas Assessment Program, allowing the established performance standards for the assessments to be maintained on all subsequent test forms. Figure 3.1 illustrates the three types of equating used for the Texas Assessment Program. 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 3.1. Three Types of Equating Used



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 is 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 four steps:

1. Items are selected that have been equated to the Rasch scale from the item bank.
2. A new test form is constructed that meets the content specifications and statistical guidelines.
3. The test form under construction is evaluated against Rasch-based difficulty targets.
4. An RSSS conversion table for the operational test form is developed using the Rasch-based item difficulties.

Pre-equating is conducted as part of the test construction process for all assessments for which scale scores are reported (i.e., STAAR, STAAR Spanish, STAAR Alternate 2, TELPAS grades 2–12). In many cases, post-equating 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 an item was field-tested because those changes might impact the estimated difficulty of the item. STAAR, STAAR Spanish, STAAR Alternate 2, and TELPAS grades 2–12 are post-equated. Post-equating in the Texas Assessment Program employs a 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:

$$t_{a,b} = \frac{\sum_{i=1}^{k}(d_{i,a} - d_{i,b})}{k}, \quad (10)$$

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 to 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 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 student proficiency are on the same scale under the Rasch model. Therefore, the resulting scale scores are comparable from year to year.

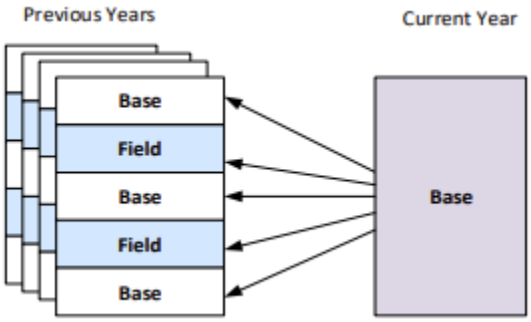
The post-equating procedure for STAAR involves the following steps:

1. Tests are assembled and evaluated using Rasch-based difficulty targets.
2. Data from the test administrations are sampled (where applicable).
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 redesigned STAAR assessments were first administered in spring 2023 with updated performance standards. Subsequent STAAR test forms will be equated to the spring 2023 administration. However, the June 2022 and December 2022 STAAR EOC assessments followed the previous scale and employed the post-equating steps listed above. For these English I and English II assessments, all multiple-choice items on each assessment were used as the equating item set, and post-equating was conducted on the entire population to ensure

representativeness. Figure 3.2 illustrates the source of the common item sets for these tests. The base-test items in the current year form were field-test items in previous years.

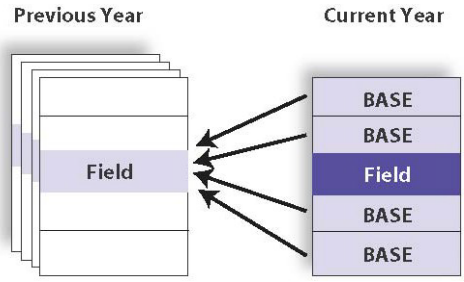
Figure 3.2. STAAR Common-Item Post-Equating Design



The initial equating item set comprised all multiple-choice items. However, the stability of the Rasch item difficulty estimates for the equating items is monitored from year to year. If a 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.

For STAAR Alternate 2 and TELPAS, the equating item set comprises all the base-test items, and the base-test items' Rasch difficulty values from field testing are compared to their values from operational testing to calculate the equating constant. Figure 3.3 illustrates the source of the equating items for STAAR Alternate 2 and TELPAS. The arrows in Figure 3.3 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 3.3. 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. 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 and the

content representation of the equating item set are compared to the base test to verify that the test content is appropriately represented in the equating item set.

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. STAAR (including STAAR Spanish), STAAR Alternate 2, and TELPAS use embedded field-test designs to collect data on field-test items. A stand-alone field test is occasionally conducted for STAAR.

After a newly constructed field-test item has cleared the review process, it is embedded in a test form along with operational items. There are two ways in which field-test items may be embedded.

STAAR (including STAAR Spanish) field-test items are randomly administered to students using a linear-on-the-fly test (LOFT) design in which all students are presented the same set of operational items that count toward their score. The LOFT design also achieves a representative sample of test takers for each item while eliminating the need for spiraling of forms.

STAAR Alternate 2 and TELPAS field-test items are placed on fixed forms along with operational items. Each field-test item appears on only a small number of test forms (typically one form) and does not count toward students' scores. Test forms containing field-test items are distributed so that a representative sample of test takers responds to the field-test items.

Regardless of which method is used to field-test items, all items are combined into a single data matrix, and a calibration of the Rasch item difficulties for both the operational items and the field-test items is conducted.

STAAR and TELPAS use a fixed common-items parameter approach to place the field-test items on the same Rasch scale as the operational items. In this procedure, all operational or base-test items are anchored to their bank values, and field-test items are calibrated and equated to the bank scale in a single step. STAAR Alternate 2 uses Wright's (1977) common-items equating procedure to transform the Rasch difficulty of the field-test items to the same Rasch scale as the common items. 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.

Reliability

Reliability indicates the precision of test scores, which also reflects the consistency of test results across testing conditions. The degree to which results are consistent is assessed using a reliability coefficient. 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 methods for estimating test score reliability, including some that require multiple assessments to be administered to the same sample of students. Because obtaining these types of reliability estimates is burdensome on schools and students, 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.

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). The following types of internal consistency measures are used to estimate the reliability of the components of the Texas Assessment Program:

- Kuder-Richardson 20 (KR20) is used for tests with only dichotomously scored items.
- Stratified coefficient alpha is used for tests containing a mixture of dichotomously scored and polytomously scored 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:

$$P'_{XX} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2+\sigma_E^2}\,, \tag{11}$$

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 (i.e., students' observed scores are more precise estimates of their true scores). KR_{20} can be mathematically represented as:

$$KR_{20} = \left[\frac{k}{k-1}\right]\left[\frac{\sigma_X^2-\sum_{i=1}^{k} p_i(1-p_i)}{\sigma_X^2}\right], \tag{12}$$

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 dichotomously scored.

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

$$\alpha = \left[\frac{k}{k-1}\right]\left[1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_X^2}\right], \quad (13)$$

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 is treated as a subtest. Given the small N counts for non-multiple-choice items, items are subset by multiple choice versus non-multiple choice. A separate measure of reliability is computed for each component and combined as follows:

$$Stratified\ \alpha = 1 - \frac{\sum_{j=1}^{c}\sigma_{X_j}^2(1-\alpha_j)}{\sigma_X^2}, \quad (14)$$

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 comprising multiple-choice and non-multiple-choice items, coefficient alpha is used as the estimate of component reliability. The correlation between ratings of the first two raters (i.e., inter-rater reliability) is used as the estimate of component reliability for written responses.

Inter-Rater Reliability

Some assessments require different types of reliability evidence than those described above. For example, STAAR RLA assessments include an extended constructed-response question at all grade levels. As part of the process for evaluating the reliability of such assessments, TEA provides evidence that the evaluation of student performance is appropriately conducted.

To gather such evidence of inter-rater reliability, two evaluators independently score the same student response. If the scores from the two scorers differ by more than one point, then a third evaluation is conducted by a supervisor or scoring director to resolve the discrepancy. These scores can then be analyzed, and the extent of agreement (or correlation) between the two sets of scores 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

Test scores for the Texas Assessment Program are typically highly reliable; however, each test score contains an associated measurement error, which 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 scores lower or higher). An example of a systematic error would be a language barrier that caused a student to incorrectly answer questions to which the student 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. TEA computes the classical standard error of measurement (SEM), the conditional standard error of measurement (CSEM), and classification consistency and classification accuracy for the purpose of estimating the amount of random error in test scores.

Standard Error of Measurement

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:

$$SEM = \sigma_X \sqrt{(1 - P'_{XX})}\,, \tag{15}$$

where P'_{XX} is the reliability estimate (e.g., KR_{20}, coefficient alpha, 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 takes 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 a 68 percent probability of being 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 (e.g., refer to 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 CSEMs.

Conditional Standard Error of Measurement

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. 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 also have errors. 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 in this section. The basic idea is to estimate the probability 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 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. Across all performance levels, the probability 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. The consistency index for a test is then the sum of the above probabilities 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:

$$\hat{\varphi} = \sum_{r=0}^{r_5} \left\{ \sum_{l=1}^{3} \left[\sum_{x=r_l}^{r_{l+1}-1} \hat{p}(x \mid \hat{\theta}_r) \right]^2 + \left[\sum_{x=r_4}^{r_5} \hat{p}(x \mid \hat{\theta}_r) \right]^2 \right\} f_r \,, \tag{16}$$

where l is the performance level (for STAAR, 1 = Did Not Meet, 2 = Approaches, 3 = Meets, 4 = Masters); r_l and r_{l+1} are the raw score cut scores 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 ; $\hat{p}(x \mid \hat{\theta}_r)$ 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, $\hat{p}(x \mid \hat{\theta}_r)$, can be estimated based on the following recursive algorithm:

$$\hat{p}(x \mid \hat{\theta}_r) = \sum_{m=0}^{M_i-1} \hat{p}_{i-1}(x-m \mid \hat{\theta}_r)\hat{p}_{im}(\hat{\theta}_r) \,, \tag{17}$$

where i refers to the ith 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 ith 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 RPCM (equation 1); and $\hat{p}_i(x \mid \hat{\theta}_r)$ is the estimated probability of getting score x conditional on $\hat{\theta}_r$ after adding the ith item.

Note that $\hat{p}(x \mid \hat{\theta}_r) = 1$, and when $x - m < \min_{i-1}$ or $x - m < \max_{i-1}$ for $i > 1$, then define $\hat{p}_{i-1}(x-m \mid \hat{\theta}_r) = 0$.

The method recommended by Rudner (2000, 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 $\hat{\sigma}_{\theta_r}$ (i.e., the CSEM). The estimated proficiency score cut score $\hat{\theta}_l$ for each performance level l is also available. For each raw score point in a performance level, the probability of correctly classifying into this level can then be estimated. The accuracy index is the sum of these probabilities across all raw scores weighted by the observed percentages of students on each raw score point, f_r. In particular, the estimation formula is written as:

$$\hat{\psi} = \sum_{l=1}^{3} \sum_{r=r_l}^{r_{l+1}-1} \left[\phi\left(\frac{\hat{\theta}_{l+1} - \hat{\theta}_r}{\hat{\sigma}_{\theta_r}} \right) - \phi\left(\frac{\hat{\theta}_l - \hat{\theta}_r}{\hat{\sigma}_{\theta_r}} \right) \right] f_r + \sum_{r=r_4}^{r_5} \left[\phi\left(\frac{\hat{\theta}_{l+1} - \hat{\theta}_r}{\hat{\sigma}_{\theta_r}} \right) - \phi\left(\frac{\hat{\theta}_l - \hat{\theta}_r}{\hat{\sigma}_{\theta_r}} \right) \right] f_r , \quad (18)$$

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

Note that each STAAR EOC assessment has three different Approaches level cut scores: one for students who first took an EOC assessment before the December 2015 administration, one for students who first took an EOC assessment on or after the December 2015 administration and before spring 2023, and one for students who first took an EOC assessment in spring 2023. Therefore, for each EOC assessment, first the classification consistency and accuracy for each group of students who have the same Approaches cut score (i.e., "Approaches 2012–2015," "Approaches 2016–2022," or "Approaches") are estimated, and then the classification consistency and accuracy indexes weighted by proportion of students in each group as the overall classification consistency and accuracy estimate for a test are summed.

Validity

Validity refers to the extent to which test scores accurately measure what the test is intended to measure. The results of STAAR, including STAAR Spanish, and STAAR Alternate 2 are used to make inferences about how well students know and understand the TEKS curriculum. Similarly, TELPAS and TELPAS Alternate test results are used to make inferences regarding English language acquisition aligned with the ELPS. When test scores are used to make inferences about student achievement, it is important that the assessment supports those inferences. In

other words, the assessment should measure what it was intended to measure in order for inferences about test results to be valid.

Validity evidence can be organized into five categories: 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; however, 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, STAAR and STAAR Spanish 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 STAAR maps the test content to the TEKS. Validity evidence supporting test content comes from the established test development process and the judgment of content experts about the relationship between the items and the test construct.

The test development process starts with a review of the TEKS by Texas educators. The educators then work with TEA to identify the readiness and supporting standards in the TEKS and help determine how each standard can best be assessed. A test blueprint developed with educator input maps the items to the reporting categories they are intended to represent. Items are then developed based on the test blueprints.

The steps in the test development process followed each year to support the validity of test content for the Texas Assessment Program are:

- Items are developed based on the TEKS curriculum standards and item guidelines.
- Items are reviewed for appropriateness of item content and difficulty, for alignment to the TEKS, and to eliminate potential bias.
- Data on field-test items is collected and reviewed to determine appropriateness for inclusion on a test.
- Tests are built to pre-defined criteria.
- 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, STAAR RLA assessments include extended constructed-response items because requiring students to respond to open-ended writing questions reflects an appropriate manner for students to demonstrate their writing abilities. Student response processes for the components of the Texas Assessment Program differ by item type.

STAAR requires students to respond to various item types, including multiple choice, technology enhanced, short constructed response, and extended constructed response. STAAR Alternate 2 involves test administrators observing students as they respond to standardized items and scoring the items based on item-specific rubrics. TELPAS grades 2–12 requires students to respond to multiple-choice items, technology-enhanced items, and performance-based speaking tasks. Holistic assessment for TELPAS kindergarten and grade 1 and TELPAS Alternate do not contain traditional items; instead, students are evaluated and assigned holistic ratings based on ongoing classroom observations.

TEA gathers evidence to support validity based on response processes from several sources. When new item types or changes to the format of existing item types are considered for any assessments, cognitive labs are used to study the way students engage with the various item presentations. In this setting, students “think aloud” while responding to assessment items. This can provide evidence that students’ cognitive processes are consistent with those expected for a given item type and that they reflect the knowledge and skills described in the TEKS. After evaluation in the cognitive lab setting, test items are pilot-tested with a larger sample of students to gather information about performance on new item types and formats. Once new item types and formats are determined to be appropriate, evidence including statistical information (e.g., item difficulty, point-biserial correlations, DIF) is gathered about student responses through field testing. The evidence is then submitted for content expert review.

The process used to score items can provide validity evidence related to response processes. For assessments with constructed-response items, human scorers use rubrics to score student responses. For TELPAS speaking, the responses are scored by an automated scoring engine. 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 STAAR, including STAAR Spanish, have been validated by educator committees and content experts. In addition, TEA has implemented a rigorous scoring process for constructed-response items that includes training and qualification requirements for human scorers, ongoing monitoring during scoring, adjudication and resolution processes for student responses that do not meet the perfect or 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.”

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. Internal consistency estimates are evaluated for reported groups, including all students, female students, male students, Black or African American

students, Hispanic or Latino students, 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 the components of the Texas Assessment Program 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 referred to as criterion-related validity. The relationship can be concurrent, predictive, convergent, or discriminant:

- Concurrent—The performances 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 performances on two measures that are meant to assess the same or similar construct should be strongly correlated.
- Discriminant—The performances on two measures that are meant to assess unrelated constructs should have a weak correlation or no correlation.

Several past and current research studies have been designed to evaluate the relationship between performance on STAAR and performance on other related tests or criteria, including the following:

- STAAR to TAKS comparison studies, which link performance on STAAR to performance on TAKS (e.g., STAAR grade 7 mathematics to TAKS grade 7 mathematics)
- STAAR linking studies, which link performance on STAAR across grade levels or courses in the same content areas (e.g., grade 4 RLA to grade 5 RLA, English I to English II)
- STAAR intercorrelation estimates, which evaluate the strength of the relationship (or lack thereof) among scores on STAAR across different content areas (e.g., grade 4 mathematics to grade 4 RLA, English I to Biology)
- grade correlation studies, which link performance on STAAR EOC assessments to course grades
- validity studies, which link performance on STAAR to other measures (e.g., Scholastic

Aptitude Test [SAT], American College Testing [ACT], Lexiles, Quantiles, STAAR Interim Assessments)

- college students taking STAAR studies, which link performance on STAAR EOC assessments to grades in college courses

For detailed descriptions and results of such studies, refer to the Assessment Reports and Studies webpage.

STAAR Alternate 2 intercorrelation estimates are calculated to evaluate the strength of the relationship between scores on STAAR Alternate 2 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 on the relationship between TELPAS reading and writing performance and STAAR RLA performance. For each grade level and TELPAS proficiency level breakout group, the following two types of performance data are examined:

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

Refer to Chapter 6, “TELPAS,” for more details. The same analysis is also conducted on the relationship between TELPAS Alternate and STAAR Alternate 2. Refer to Chapter 7, “TELPAS Alternate,” for more details.

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, STAAR is intended to have an effect on instructional content and delivery strategies; however, an unintended consequence could be the narrowing of instruction, a phenomenon sometimes referred to as "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.

Measures of Student Progress

Measures of student progress describe changes in student performance across time. 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. 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 that the student achieved in the previous year. If the student achieved Did Not Meet Grade Level in the previous year, then the student made notable progress this year by advancing a performance level. However, if the student achieved Meets Grade Level in the previous year, then the interpretation of Approaches Grade Level performance this year would be quite different because the student regressed.

Development of Progress Measures

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.

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 STAAR and STAAR Alternate 2,
- the appropriateness of progress measures given the content relationships among STAAR and STAAR Alternate 2,
- 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 progress measures for STAAR and STAAR Alternate 2. 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 composed of educators from various Texas campuses, districts, and 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 earlier, gain score was selected as the progress measure for STAAR. The STAAR Progress Measure was implemented for the first time in the 2012–2013 school year beginning with STAAR and STAAR Spanish mathematics

and reading. Since then, Algebra I, English I, and English II have been added to the STAAR Progress Measure, which has been reported every year except for the years when performance standards have been reestablished.

In addition to the STAAR Progress Measure, TEA also produces an 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 and STAAR Spanish mathematics and RLA.

The STAAR Alternate 2 Progress Measure employs a transition table approach and was reported for the first time in in 2016 with the mathematics and reading assessments. STAAR Alternate 2 progress measures were calculated and reported for mathematics and RLA assessments.

Details about these progress measures can be found in Chapter 4, "STAAR," and Chapter 5, “STAAR Alternate 2,” and on the Progress Measures webpage.

Sampling

Sampling is a procedure that is used to select and examine a small set that is representative of the population from which it is drawn. The results from well-drawn samples allow TEA to estimate characteristics of the Texas student population as a whole. Through the careful selection of student samples, TEA is able to make reliable and valid inferences about student performance on its assessments while minimizing the burden on campuses and districts.

Key Concepts of Sampling

A target population is the set of students to which the results should generalize, also known as the complete collection of objects of interest (Lohr, 1999). For example, consider a study with the goal of understanding how grade 3 EB students perform on a set of test questions. In that case, the target population would be all grade 3 EB students in Texas. Careful consideration is given to defining the target population before sampling takes place.

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 that each of 20 randomly selected campuses 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, in which 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 selected), the observation unit is the student (because the data being collected reflect student characteristics).

Reasons for Sampling

The Texas Assessment Program employs sampling instead of studying entire target populations for several reasons, including the following:

- 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 more cost efficient to obtain data for a carefully selected subset of a population than it is to collect the same data for the entire population.
- Size—It is more efficient to examine a representative sample when there is a large target population.
- Time—Using sampling to study the target population is less time consuming. Sampling might be needed when the timeline of the analysis is important.

Sampling Designs

The Texas Assessment Program uses sampling to collect data for the purpose of field testing, audits, and research studies (e.g., linking studies, cognitive labs, 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. Research studies generally involve assessing a sample of students under various testing conditions to collect evidence to support the technical quality of and make improvements to the Texas Assessment Program. TEA uses the following sample designs.

Probability Sampling

In a probability sample, all sampling units have a known probability of being selected. Probability sampling requires that the number of sampling units in the target population is known. 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 following are the major types of probability sampling designs:

- Simple Random Sampling—All sampling units in the target population have the same probability of being selected.
- Stratified Sampling— First the sampling units are 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— First the sampling units are 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 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 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 comprises *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 important in re-sampling and replication methods, such as bootstrapping.

Re-Sampling and Replication Methods: Bootstrapping

Bootstrapping is one of the re-sampling and replication methods that treats 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 and Tibshirani (1993). The Texas Assessment Program uses bootstrapping methods when conducting comparability studies that compare online and paper versions of a test form.

Convenience (Nonprobability) Sampling

A sample that is created without the use of random selection is a convenience (or nonprobability) 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.

TECHNICAL DIGEST 2022–2023

Chapter 4

State of Texas Assessments of Academic Readiness

Overview

TEA, in collaboration with THECB and Texas educators, developed the STAAR program in accordance with educational requirements set forth by the Texas legislature in 2007 and 2009.

STAAR was implemented in the 2011–2012 school year and included Spanish versions of the assessments for grades 3–5.

STAAR is designed to measure the extent to which a student has learned and is able to apply the knowledge and skills defined in the TEKS. Every item is directly aligned to the TEKS currently in effect for the tested grade and subject or course. STAAR includes the following assessments:

- grades 3–8 mathematics,
- grades 3–8 RLA,
- grades 5 and 8 science,
- grade 8 social studies, and
- EOC assessments for:
    - Algebra I,
    - English I,
    - English II,
    - Biology, and
    - U.S. History.

Based on legislation passed in 2019, STAAR was redesigned to align more closely with effective classroom instruction. The redesign was implemented in the 2022–2023 school year and included:

- the addition of new non-multiple-choice questions that give students more ways to show their understanding and better reflect questions teachers ask in the classroom,
- the addition of a writing component to reading assessments for grades 3–8 to better support the interconnected way these subjects are taught, and
- the incorporation of more cross-curricular passages into the new RLA assessments so that test questions can reference topics students have learned about in other classes.

STAAR Spanish

STAAR Spanish is administered to eligible students for whom the language proficiency assessment committee (LPAC) determines that STAAR Spanish is the most appropriate way to measure those students' mastery of skills and is also available for students who receive academic instruction in Spanish while they learn English. The STAAR Spanish assessments are offered for grades 3–5 mathematics and RLA and for grade 5 science. The English and Spanish

versions of STAAR have the same test blueprint and assess the same TEKS student expectations for mathematics and science and similar student expectations in RLA.

STAAR Interim Assessments

STAAR Interim Assessments are a set of optional online assessments aligned to the TEKS; the purpose of the interim assessments is to monitor student progress and predict student performance on STAAR summative assessments. The interim assessments are available at no cost to districts and are not tied to accountability. More information is available on the STAAR Interim Assessments webpage.

Testing Requirements

All students enrolled in Texas public schools and open-enrollment charter schools are required to take STAAR unless the student meets the participation requirements for STAAR Alternate 2.

Students enrolled in grade 9 or below for the first time in the 2011–2012 school year or later are required to meet STAAR EOC assessment graduation requirements.

In 2015, legislation revised the state's assessment graduation requirements to allow an eligible student to receive a Texas high school diploma by means of an IGC if the student fails to pass no more than two STAAR EOC assessments. Eligibility criteria for an IGC can be found in TEC §28.0258.

The admission, review, and dismissal (ARD) committee makes educational decisions, including decisions related to state assessments and graduation requirements as described in TAC §89.1070, for students receiving special education services.

Due to the impact of the COVID-19 pandemic, STAAR testing was suspended for spring and summer 2020, and a STAAR EOC assessment waiver reduced the number of EOC assessments that a student was required to pass to meet assessment graduation requirements. To qualify for the waiver, a student must have:

- been enrolled in the course during spring or summer 2020,
- completed the full course by the end of spring or summer 2020, and
- earned full course credit by the end of the spring or summer 2020.

Test Development

Maintaining a high-quality student assessment program involves a complex and detailed test-development process, and TEA relies on input from educators to ensure that all measures of learning for Texas public school students are equitable and accurate. Test items for STAAR, including STAAR Spanish, are developed annually, reviewed by educator committees, field-tested, reviewed with their data, and, if approved, added to the STAAR item bank. In most cases, newly developed items are embedded in STAAR operational assessments each spring. However, stand-alone field tests are periodically required and have been administered in 2011, 2015, 2019, and 2022. For more information regarding each step of the STAAR test-development process, refer to Chapter 2, "Building a High-Quality Assessment System,"

which outlines the processes used to develop each STAAR assessment's framework and explains ongoing test development.

STAAR English-Spanish Alignment

TEA staff, Texas educators, and Spanish-language experts collaborate to develop STAAR Spanish test materials. STAAR Spanish RLA assessments are composed entirely of passages and items in Spanish. This development approach allows the Spanish RLA curriculum to be assessed in a more authentic and meaningful manner. Items for STAAR Spanish mathematics and science are transadapted. Transadaptation involves translating items from English and adapting them as necessary to ensure cultural and linguistic accessibility. Spanish bilingual educators then review all original and transadapted test items in accordance with the educator review process described in Chapter 2, "Building a High-Quality Assessment System."

The following practices reinforce alignment of the STAAR English and Spanish assessments:

- When the performance standards for STAAR were established, standard-setting panels reviewed both the English and Spanish grades 3–5 RLA assessments to establish comparable performance standards.
- The development and review processes for the RLA assessments in English and Spanish are parallel (i.e., item reviews for English and Spanish include judgments related to each item's alignment to the TEKS). Field-test data reviews for English and Spanish items also include item statistics reviews based on actual student performance. These safeguards ensure that only psychometrically sound items are selected for inclusion in the STAAR item banks.
- Each year, STAAR development staff review the newly developed test items, focusing on the best ways to assess the TEKS and further enhancing the alignment between the English and Spanish assessments.
- The RLA assessments in English and Spanish are constructed concurrently and in coordination, and they adhere to the same test construction guidelines regarding the range of item content and cognitive complexity.
- The Spanish mathematics and science assessments are transadapted from the corresponding English assessments. The item-writing and review processes for transadapted items ensure that the Spanish items are linguistically and culturally appropriate and that the interpretations of grade-level performance expectations are the same for English and Spanish.
- The test blueprints for the English and Spanish assessments are the same, including the number of items that assess each reporting category and the number of items on the test.

Accommodations

The goal of STAAR accommodations is to ensure that each student can interact appropriately with the content, presentation, and response modes of the state assessments. To meet this goal, STAAR accommodations are designed to allow all students to demonstrate their

knowledge of the content being assessed without the format of the assessment, the non-tested language, or the type of response needed to answer the questions being barriers. The various accommodations made available on STAAR are also designed to be the same or similar to those accommodations commonly used during classroom instruction.

Accommodation policies for STAAR, including STAAR Spanish, are divided into three main categories: accessibility features, locally-approved designated supports, and designated supports requiring TEA approval. More information is available on the Accommodations page of the *Coordinator Resources*.

Accessibility Features

Accessibility features may be provided to students based on their needs. In general, these procedures and materials are available to any student who regularly benefits from their use during instruction; however, a student cannot be required to use them during STAAR. Coordinators are responsible for ensuring that test administrators understand the proper implementation of these procedures and use of these materials. In some cases, a student may need to complete the test in a separate setting to eliminate distractions to other students and to ensure that the security and confidentiality of the test are maintained.

Locally-Approved Designated Supports

Locally-approved designated supports include accommodations that may be made available to students who meet eligibility criteria. The appropriate team of people at the campus level (e.g., Response to Intervention [RtI] team; LPAC; Section 504 committee; ARD committee) determines eligibility as indicated in each policy document. The decision to allow the use of a designated support during STAAR should be made on an individual student basis, taking into consideration the needs of the student and whether the student routinely receives the support during classroom instruction and classroom testing. In addition, the support has to have been proven effective in meeting the student's specific needs, as evidenced by student scores or teacher observations.

Designated Supports Requiring TEA Approval

These designated supports require the submission of an Accommodation Request Form to TEA. The appropriate team of people at the campus level, as indicated in each policy document, determines whether the student meets all the specific eligibility criteria and, if so, submits an Accommodation Request Form to TEA. TEA must receive Accommodation Request Forms according to the posted deadlines. Late requests will not be processed unless circumstances involving the student change after the deadline (e.g., newly enrolled student, medical emergency, updated ARD committee decision). The request must be approved by TEA before a student can use the designated support on STAAR.

Training

TEA develops instructional materials, including manuals, guides, presentations, online modules, and videos, to support the training of all testing personnel on test security and administration procedures. Preparation for test administration begins every year with a TEA-provided training-of-trainers session for testing coordinators from each of the 20 Texas regional ESCs as well as district testing coordinators from the state's 25 largest districts. Using materials and information

provided in the TEA training session, ESC regional testing coordinators train the district coordinators in their respective regions. District coordinators then train their campus testing coordinators, who are responsible for training test administrators.

Test security and administration procedures provided in the *Coordinator Resources* and the *STAAR Test Administrator Manual* must be followed so that all students have an equal opportunity to demonstrate their academic knowledge and skills. The *Coordinator Resources* guide district and campus coordinators through their responsibilities as they oversee the administration of the Texas Assessment Program. This online resource contains preparation and administration procedures for each state-required assessment and is available prior to the annual ESC training.

Test Administration

All STAAR assessments—grades 3–8 mathematics, grades 3–8 RLA, grades 5 and 8 science, grade 8 social studies, Spanish grades 3–5 mathematics, Spanish grades 3–5 RLA, Spanish grade 5 science, Algebra I, English I, English II, Biology, and U.S. History—are administered online in the spring. STAAR EOC assessments are also administered online in June and December. A paper version of STAAR is available for students with a special circumstance. The number of students tested for each STAAR assessment is shown in Table 4.1.

Table 4.1. STAAR Assessments Administered in 2022–2023

STAAR Assessment	Assessments Administered
Grade 3 mathematics	370,006
Grade 3 RLA	356,558
Grade 3 Spanish mathematics	16,454
Grade 3 Spanish RLA	30,213
Grade 4 mathematics	373,988
Grade 4 RLA	365,035
Grade 4 Spanish mathematics	11,497
Grade 4 Spanish RLA	21,694
Grade 5 mathematics	378,663
Grade 5 RLA	372,677
Grade 5 science	378,742
Grade 5 Spanish mathematics	8,483
Grade 5 Spanish RLA	15,991
Grade 5 Spanish science	9,775
Grade 6 mathematics	384,766
Grade 6 RLA	391,376
Grade 7 mathematics	331,698
Grade 7 RLA	400,416

STAAR Assessment	Assessments Administered
Grade 8 mathematics	364,110
Grade 8 RLA	410,472
Grade 8 science	407,847
Grade 8 social studies	414,692
Algebra I	608,559
English I	712,965
English II	612,128
Biology	544,720
U.S. History	427,007

NOTE: For the STAAR EOC assessments, the table includes the sum total of the December, spring, and June administrations.

The Test Delivery System

STAAR online assessments are administered using the Test Delivery System (TDS). TDS includes the Test Administrator Interface, which is used for scheduling test sessions; the Student Interface, which allows students to participate in testing; and the Secure Browser application, which provides a secure online environment for testing. TDS allows for the secure transfer and storage of test data while remaining scalable to support the student testing population. The TDS architecture has demonstrated stability and efficiency by seamlessly handling over 1.2 million concurrent users.

Make-up Testing

Make-up testing opportunities for students who are absent on the day of testing are available during the STAAR testing window for all grades, subjects, and courses.

Out-of-District Testing

For STAAR EOC assessments, students who are unable to test in their home district are allowed to test out-of-district (OOD). For example, a student from Houston who spends the summer in Dallas could register to test in Dallas. OOD students are required to complete registration within a set window so that receiving districts are aware of the student's intent and have the resources to administer the assessment. Students must present photo identification at the test administration site on the day of the test.

Out-of-School Testing

Examinees who have not passed a STAAR EOC assessment and are no longer enrolled in school but have otherwise completed requirements for graduation may take an assessment during a test administration at a participating district.

Performance Standards

Performance standards directly relate levels of test performance to what students are expected to learn, as defined in the statewide curriculum. Standard setting is the process of establishing cut scores that define the performance levels on an assessment.

Performance Levels and Policy Definitions

For STAAR, including STAAR Spanish, the performance levels and policy definitions are as follows:

Did Not Meet Grade Level

Performance in this category indicates that students are unlikely to succeed in the next grade or course without significant ongoing academic intervention. Students in this category do not demonstrate a sufficient understanding of the assessed knowledge and skills.

Approaches Grade Level

Performance in this category indicates that students are likely to succeed in the next grade or course with targeted academic intervention. Students in this category generally demonstrate the ability to apply the assessed knowledge and skills in familiar contexts.

Meets Grade Level

Performance in this category indicates that students have a high likelihood of success in the next grade or course but may still need some short-term, targeted academic intervention. Students in this category generally demonstrate the ability to think critically and apply the assessed knowledge and skills in familiar contexts.

Masters Grade Level

Performance in this category indicates that students are expected to succeed in the next grade or course with little or no academic intervention. Students in this category demonstrate the ability to think critically and apply the assessed knowledge and skills in varied contexts, both familiar and unfamiliar.

Standard Setting

The STAAR program's goal was to have a comprehensive assessment system with curriculum standards and performance standards that were vertically aligned within a content area (i.e., the curriculum and performance standards linked from the high school courses back to the middle school and elementary school grades and subjects). Standard setting for STAAR took into consideration a variety of factors, such as policy, 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 performance on other assessments. Standard-setting committees comprised diverse groups of stakeholders who carefully considered the interaction of these elements for each STAAR assessment. The task of each standard-setting committee was to recommend cut scores that would define the performance levels for each STAAR assessment.

Initial performance standards for all STAAR assessments were established in 2012, and performance standards were reset in 2023 with the redesign of STAAR. The current performance standards for STAAR are provided in Tables 4.2 and 4.3.

Table 4.2. STAAR Grades 3–8 Performance Standards

Assessment	Approaches Grade Level (Scale Score)	Meets Grade Level (Scale Score)	Masters Grade Level (Scale Score)
Grade 3 Mathematics	1360	1471	1600
Grade 4 Mathematics	1462	1557	1690
Grade 5 Mathematics	1515	1634	1776
Grade 6 Mathematics	1616	1745	1889
Grade 7 Mathematics	1703	1793	1965
Grade 8 Mathematics	1754	1859	2009
Grade 3 RLA	1345	1467	1596
Grade 4 RLA	1414	1552	1663
Grade 5 RLA	1475	1592	1700
Grade 6 RLA	1535	1634	1749
Grade 7 RLA	1564	1669	1771
Grade 8 RLA	1592	1698	1803
Grade 5 Science	3550	4000	4380
Grade 8 Science	3550	4000	4619
Grade 8 Social Studies	3550	4000	4352
Grade 3 Spanish Mathematics	1360	1471	1600
Grade 4 Spanish Mathematics	1462	1557	1690
Grade 5 Spanish Mathematics	1515	1634	1776
Grade 3 Spanish RLA	1318	1447	1515
Grade 4 Spanish RLA	1408	1488	1581
Grade 5 Spanish RLA	1431	1556	1662
Grade 5 Spanish Science	3550	4000	4380

Table 4.3. STAAR EOC Assessments Performance Standards

Assessment	Approaches Grade Level 2012–2015 (Scale Score)	Approaches Grade Level 2016–2022 (Scale Score)	Approaches Grade Level (Scale Score)	Meets Grade Level (Scale Score)	Masters Grade Level (Scale Score)
Algebra I	3489	3541	3550	4000	4345
English I	3775	3775	3775	4000	4606
English II	3766	3775	3775	4000	4734
Biology	3516	3550	3550	4000	4531
U.S. History	3486	3536	3550	4000	4424

Refer to the STAAR standard-setting technical reports, which are available on the Assessment Reports and Studies webpage, for more information.

Scores and Reports

TEA publishes resources on both the TEA and Texas Assessment websites to assist school personnel in understanding and interpreting student performance data and to help parents understand their child's STAAR results. School personnel can access STAAR test results through the Centralized Reporting System (CRS), parents can access their child's STAAR results in the Family Portal, and the public can access STAAR statewide, region, district, and campus data using the Research Portal.

TEC §39.030 and TAC §101.3014 specify the requirements for maintaining the confidentiality of individual student results and for reporting district-level and campus-level results. The results of individual student performance on state assessments are confidential and may be released only in accordance with the Family Educational Rights and Privacy Act (FERPA). Districts must provide each student's state assessment results to the student, to his or her parent or guardian, and to his or her teacher for the applicable subject area. In addition, all state assessment results must be included in each student's academic achievement record.

Description of Scores

Scores for STAAR and STAAR Spanish include raw scores, scale scores, and the resulting performance level associated with the student's score. Additionally, percentiles, Lexiles, Quantiles, and English learner (EL) performance measures are provided.

The number of points that a student earns on a STAAR assessment is the student's raw score. A scale score is a conversion of the raw score onto a scale that is common to all test forms for that assessment. The scale score is used to determine whether a student achieved the Did Not Meet Grade Level, Approaches Grade Level, Meets Grade Level, or Masters Grade Level performance standard. Refer to Chapter 3, "Standard Technical Processes," for more information about raw scores and scale scores.

Percentiles represent the percentage of students across the state who took the assessment and received a scale score at or below the scale score of interest. Percentiles are calculated based on all students (except out-of-school [OOS] testers) who received valid scale scores on the assessment in the previous year's spring administration.

Students receive a Lexile Measure on the STAAR RLA assessments, including grades 3–8 RLA, Spanish grades 3–5 RLA, English I, and English II. Lexile measures indicate the level of difficulty of materials a student can read and range from below 0L for beginning readers to above 1600L. Similarly, students receive a Quantile Measure on STAAR mathematics assessments, including grades 3–8 mathematics, Spanish grades 3–5 mathematics, and Algebra I. Quantile measures indicate the mathematics concepts a student has learned and the concepts they are ready to learn next. These measures range from below 0Q to above 1400Q. More information about Lexiles and Quantiles is available on the Texas Assessment website.

Beginning in the 2018–2019 school year, qualifying EB students who tested in English also received an EL performance measure, which showed whether an eligible EB student was making sufficient progress on each STAAR content-area assessment based on predetermined

performance measure progress expectations. The EL performance measure was calculated and reported for all STAAR assessments except STAAR Spanish.

Assessment Reports

TEA provides reports of student performance on STAAR, including STAAR Spanish, to all Texas public school districts and open-enrollment charter schools. For each STAAR administration, student report cards, student labels, campus rosters, summary reports, and reporting data files are provided.

The spring administration of each assessment for STAAR, including STAAR Spanish, is released to the public through the Practice Test Site. To correspond with the released tests, TEA provides student item analysis reports and item analysis summary reports. These summary reports are available at the campus, district, region, and state level.

For more information about scoring and reporting for STAAR, refer to the Interpreting Results page of the *Coordinator Resources*.

Use of Test Results

Test results can be used to evaluate the performance of a group over time. Average scale scores and the percentage of students meeting the Approaches Grade Level, Meets Grade Level, and Masters Grade Level performance standards can be analyzed by grade and content area across administrations to provide insight into whether student performance is improving across years. For example, the average scale score for students who took the STAAR grade 4 RLA test can be compared over time.

Test results can also be used to compare the performance of different demographic or program groups. STAAR scores can be analyzed within the same content area of any single administration to determine, for example, which demographic or program group has the highest average scale score, which group has the lowest percentage meeting the Approaches Grade Level performance standard, and which group has the highest percentage achieving the Meets Grade Level performance standard. Other scores can be used to help evaluate the academic performances of demographic or program groups in core academic areas. For example, reporting category data can help districts and campuses identify areas of potential academic weakness for a group of students. The same methodology can be applied to an entire district or campus. Test results for groups of students can be used when evaluating instruction or programs that require average-score or year-to-year comparisons. The tests are designed to measure content areas within the required state curriculum, and so the consideration of test results by content area and reporting category might be helpful when evaluating curriculum and instructional programs. All test scores can be compared with statewide and regional performance within the same content area for any administration.

Test scores can also be used to identify where an individual student needs additional instruction or support in each subject. Other scores can provide information about a student's relative strengths or weaknesses in core academic areas. For example, reporting category–level data can provide information about a student's relative strengths or weaknesses and can be used to identify areas where a student might be having difficulty. This identification can help educators plan the most effective instructional intervention. Finally, individual student test scores are also used in conjunction with other performance indicators to assist in making placement decisions.

While scores can contribute to decisions regarding placement, educational planning for a student should take into account as much student information as possible.

Generalizations from test results can be made from the specific content area being measured on the test. However, because each test measures a finite set of skills with a limited set of items, any generalizations about student achievement derived solely from a particular test should be made with great caution and with full reference to the fact that the conclusions are based only on that test. Instruction and program evaluations should take into account as much information as possible, rather than relying on test scores alone, to provide a more complete picture of student performance.

Measures of Student Progress

Student progress measures provide information beyond performance levels by providing a comparison of performance over time. Whereas performance-level information describes students' current achievement, progress measures describe students' achievement across multiple years.

STAAR Progress Measure

The STAAR Progress Measure is legislatively mandated and was reported for the first time in the 2012–2013 school year. For STAAR, progress is measured as a student's gain score, which represents the difference between the scale score a student achieved in the prior school year and the scale score a student achieved in the current school year. These gain scores are then classified as Limited, Expected, and Accelerated in relation to progress targets. The progress targets define the expectation of annual progress for each grade and content area. These progress targets are grounded in the STAAR performance standards, the goal of having all students achieve Meets Grade Level or above, and having high-performing students maintain Masters Grade Level performance.

Specifically, for students who achieve Did Not Meet Grade Level, Approaches Grade Level, or Meets Grade Level performance standards in the prior year, the Expected progress target is defined as the distance between the Meets Grade Level performance standard on the prior year test and the Meets Grade Level performance standard on the current year test in the same content area. For students who achieve the Masters Grade Level performance standard in the prior year, the progress target is based on the distance between Masters Grade Level on the prior year test and Masters Grade Level on the current year test in the same content area.

The Accelerated progress classification is a designation reserved for those students who have demonstrated significant growth over the course of the year beyond that of the Expected progress target. The Accelerated progress target is defined as the distance between Meets Grade Level on the prior year test and Masters Grade Level on the current year test.

Students with gain scores less than the Expected progress target are classified as having achieved Limited progress. Students with gain scores greater than or equal to the Expected progress target and less than or equal to the Accelerated progress target are classified as having achieved Expected progress. Students with gain scores greater than the Accelerated progress target are classified as having achieved Accelerated progress.

At the extreme high and low ends of the scale, the application of the Limited, Expected, and Accelerated definitions would not be appropriate. At the extreme ends of the scale, unlike the rest of the scale, answering one more question correctly results in significant differences in scale scores. For this reason, several places on the scale have been identified as exceptions to the Limited, Expected, and Accelerated definitions.

Because the performance standards do not have the same numerical value across grades and content areas, the Expected and Accelerated progress targets differ from grade to grade and across content areas. Steps for calculating progress measures and progress targets for each STAAR grade and content area, including when students skip grade levels, can be found on the Progress Measures webpage.

Due to the STAAR redesign and updated performance standards, progress measures were not reported for STAAR assessments for the 2022–2023 school year.

STAAR On-Track Measure

While the STAAR Progress Measure accounted for performance from the prior year and the current year, it did not include any information about how the student was likely to perform in the future. Because this additional information may be helpful to students, teachers, and other stakeholders, TEA developed the STAAR on-track measure, which was reported for the first time in 2013–2014. The on-track measure used the STAAR Progress Measure and extrapolated performance into future years to determine if a student was on-track to achieve Meets Grade Level in a later grade or course. To calculate the STAAR on-track measure, three assessments covering the same content area must be available (i.e., previous year, current year, and target year). For example, the on-track measure can be calculated for STAAR grade 7 RLA (current year assessment) because the previous year assessment was STAAR grade 6 RLA and the target year assessment will be STAAR grade 8 RLA. Additional information about on-track measures can be found on the Progress Measures webpage.

Due to the STAAR redesign and updated performance standards, on-track measures were not reported for STAAR assessments for the 2022–2023 school year.

Scaling

Scaling is a statistical procedure that places raw scores on a common scoring metric to make test scores comparable across test administrations. Scaling associates numbers with characteristics of interest to provide information about measurable quantities for those characteristics. STAAR, including STAAR Spanish, uses the RPCM to place test items on the same Rasch scale across administrations for a given STAAR assessment. Once performance standards have been set for an assessment, the Rasch scale is then transformed to a more user-friendly metric to ease interpretation of the test scores. Details of the RPCM scaling method are provided in Chapter 3, “Standard Technical Processes.”

Reporting Scales

STAAR scale scores are reported on either a horizontal scale or a vertical scale. Horizontal scale scores allow for direct comparisons of student performance between specific sets of test items from different test administrations. Vertical scale scores allow for direct comparisons of student scores across grades within a content area. Student increases in vertical scale scores

provide information on the year-to-year growth of students. Refer to Chapter 3, “Standard Technical Processes,” for detailed information about the scaling process.

Horizontal Reporting Scales

The following STAAR assessments are reported on horizontal scales:

- grade 5 science
- grade 8 science
- grade 8 social studies
- Spanish grade 5 science
- Algebra I, English I, English II, Biology, and U.S. History

For all STAAR assessments reported on a horizontal scale, a scale score of 4000 represents the Meets Grade Level performance standard. The Approaches Grade Level cut score was set to 3550 for all STAAR assessments except for English I and English II, for which the cut score was set to 3775. The Masters Grade Level cut scores vary across STAAR assessments, but for any given assessment, performance standards remain constant over time.

The STAAR scale scores represent linear transformations of the Rasch proficiency-level estimate (θ). Specifically, the transformation is made by first multiplying θ by a slope constant (*A*) and then adding an intercept constant (*B*). This operation is described by the following equation:

$$SS_\theta = A \times \theta + B,$$

where SS_θ is the scale score for a Rasch proficiency score estimate (θ) and *A* and *B* are referred to as the horizontal scaling constants. These same transformations are applied each year to the Rasch proficiency score estimates (θ) for that year’s set of test items. Values for the horizontal scaling constants are provided in Tables 4.4 and 4.5 for the horizontally scaled STAAR grades 3–8 and EOC assessments, respectively.

Table 4.4. Horizontal Scaling Constants for STAAR Grades 3–8

Assessment			*A*	*B*
Grade	**Language**	**Content Area**		
5	English	Science	555.8300	3661.6663
8	English	Science	630.2521	3873.5084
8	English	Social Studies	571.3560	3726.2633
5	Spanish	Science	555.8300	3661.6663

Table 4.5. Horizontal Scaling Constants for STAAR EOC Assessments

Assessment	*A*	*B*
Algebra I	460.7351	3919.0028
English I	429.3074	3845.4064
English II	444.4006	3852.8590
Biology	435.9620	4042.0267
U.S. History	487.6991	4073.2524

Vertical Reporting Scales

As required by TEC §39.036, TEA developed vertical scales for assessing student performance in grades 3–8 for mathematics and RLA. Vertical scales were developed for the following grades and subjects:

- grades 3–8 mathematics
- grades 3–8 RLA
- Spanish grades 3–5 mathematics
- Spanish grades 3–5 RLA

The vertical scale established for the English versions of grades 3–5 mathematics was also applied to the Spanish versions of grades 3–5 mathematics because the Spanish versions were transadapted from the English. For the STAAR grades 3–8 mathematics scale, a scale score of 1360 represented the Approaches Grade Level performance standard for the grade 3 assessment. The scale's standard deviation was set to 150.

For the STAAR grades 3–8 RLA scale, a scale score of 1345 represented the Approaches Grade Level performance standard for the grade 3 assessment. For the STAAR Spanish grades 3–5 RLA scale, a scale score of 1318 represented the Approaches Grade Level performance standard for the grade 3 assessment. The RLA vertical scales' standard deviations were also set to 150 for the assessment in both languages.

It is important to note that although Approaches Grade Level scale score values are fixed for the lowest grade in the vertical scale, the Approaches Grade Level scale score for the other assessments in the vertical scale will vary across grades. However, these Approaches Grade Level scale score values, as well as the Meets Grade Level and Masters Grade Level scale score values, remain constant over time.

The linear transformation of the underlying Rasch proficiency score estimate (θ) for vertical scale scores is described by the following equation for a vertically scaled test at grade *g*:

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

where SS_θ is the scale score for a Rasch proficiency score estimate (θ), *A* and *B* are the vertical scale score transformation constants, and V_g is the vertical scaling constant for the grade *g* test. The values of *A*, *B*, and V_g for the vertically scaled STAAR assessments are provided in

Table 4.6. Once established, these same transformations are applied each year to the proficiency level estimates for that year's set of test questions.

Table 4.6. Vertical Scale Score Transformation and Scaling Constants for STAAR Grades 3–8 Mathematics and RLA

Assessment			*A*	*B*	V_g
Grade	**Language**	**Content Area**			
3	English/Spanish	Mathematics	130.0052	1454.3188	0
4	English/Spanish	Mathematics	130.0052	1454.3188	0.5911
5	English/Spanish	Mathematics	130.0052	1454.3188	1.0884
6	English	Mathematics	130.0052	1454.3188	1.9965
7	English	Mathematics	130.0052	1454.3188	2.4185
8	English	Mathematics	130.0052	1454.3188	3.0511
3	English	RLA	143.7195	1398.5930	0
4	English	RLA	143.7195	1398.5930	0.6921
5	English	RLA	143.7195	1398.5930	0.6641
6	English	RLA	143.7195	1398.5930	1.4135
7	English	RLA	143.7195	1398.5930	1.2939
8	English	RLA	143.7195	1398.5930	1.9002
3	Spanish	RLA	153.0768	1318.1531	0
4	Spanish	RLA	153.0768	1318.1531	0.4323
5	Spanish	RLA	153.0768	1318.1531	0.6918

Equating

Used in conjunction with the scaling process, equating is the process that considers the differences in difficulty across test forms and administrations and allows scores to be placed onto a common scale. By using statistical methods, TEA equates the results of different test forms so that scale scores across test forms and testing administrations can be compared. TEA uses pre-equating for all STAAR assessments and post-equating for STAAR assessments that include constructed-response items.

To replenish the item bank as new tests are created each year, newly developed items must be field-tested and equated to the item bank scale. During each spring administration, field-test equating is conducted for STAAR, including STAAR Spanish, through an embedded-field-test design for all tests. In some years, stand-alone field tests are conducted for STAAR. Each stand-alone field test also includes some items from the item bank as anchor items, and the field-tested items are equated to the item bank scale through these items.

Refer to Chapter 3, "Standard Technical Processes," for detailed information about equating.

Reliability

Reliability indicates the precision of test scores, referring to the expectation that repeated administrations of the same test should generate consistent results. Reliability for STAAR test

scores is estimated using statistical measures including internal consistency, classical SEM, CSEM, and classification consistency and accuracy. Data for each of these statistical measures from the spring STAAR administration are provided in Appendix B. Refer to Chapter 3, "Standard Technical Processes," for detailed information about reliability.

Validity

Validity refers to the extent to which test scores accurately measure what the test is intended to measure. TEA follows national standards of best practice and annually collects validity evidence to support the interpretations and uses of STAAR, including STAAR Spanish, test scores. TTAC, a panel of national testing experts created specifically for the Texas Assessment Program, provides ongoing input to TEA about STAAR validity evidence. The following sections describe how validity evidence has been collected for STAAR. Refer to Chapter 3, "Standard Technical Processes," for additional information about validity.

Evidence Based on Test Content

Validity evidence based on test content refers to evidence of the relationship between tested content and the construct that the assessment is intended to measure. STAAR, including STAAR Spanish, has been developed to align with content as defined by the TEKS. Content validity evidence is collected at all stages of the test-development process. Nationally established test-development processes for the Texas Assessment Program are followed while developing STAAR. This supports the use of STAAR scores in making inferences about students' knowledge and understanding of the TEKS.

Relationship to the Statewide Curriculum

The TEKS are designed to ensure that Texas students receive a solid education that will enable them to be successful in life, whether they choose to pursue higher education or enter the workforce directly after graduation. The TEKS are specifically aligned to the CCRS. The CCRS specify the knowledge and skills necessary to succeed in entry-level community college and university courses. The CCRS have been incorporated into the secondary TEKS to form a vertically articulated set of curriculum standards. STAAR focuses on fewer skills and addresses those skills in a deeper manner through the identification of readiness and supporting standards in the TEKS and the inclusion of a larger number of items that assess readiness standards in the test blueprint. STAAR, therefore, focuses on the TEKS that are most critical to success in the next grade or course and ultimately on postsecondary readiness.

Educator Input

As part of the development of STAAR, teachers, curriculum specialists, test development specialists, college educators, and TEA staff worked together in advisory committees to identify appropriate assessment reporting categories for STAAR. The input of the advisory committees was reflected in the assessed curricula and test blueprints. In addition, prototype items were developed for the assessments early in the development process. The educator advisory committees reviewed these prototypes to identify how well the items would measure the student expectations to which the items were aligned. These early reviews provided valuable suggestions for item development guidelines and item types. Item development guidelines continued to be refined through the test development process as various STAAR item-review

educator committees shared their feedback about how the student expectations could be effectively assessed.

As part of the annual process of item development, committees of Texas educators meet to review STAAR items and confirm that each item appropriately measures the TEKS to which it is aligned. These committees also review items for content and bias. Two distinct types of educator committee meetings are regularly held to support the validity of test content: item review committees and content validation committees. Item review committees are composed of Texas educators, and these committees revise and edit items, as appropriate, prior to field testing. Item review committees are convened for all STAAR assessments. Content validation committees, by comparison, comprise university faculty who are experts in the relevant subject matter. Though these committees do not edit or revise items prior to field testing, they can recommend that certain items not be placed on STAAR operational assessments. Content validation is conducted for all STAAR EOC assessments before assessments are administered to students.

Test Developer Input

Item writers and reviewers follow test development guidelines that explain how content aligned to given TEKS should be measured. At each stage of development, writers and reviewers verify the alignment of the items with the assessed student expectations.

Evidence Based on Response Processes

Response processes refer to the cognitive behaviors that are required to respond to a test item. TEA collects evidence to show that the way students respond to items on STAAR, including STAAR Spanish, reflects accurate measurement of the construct.

Items

Student response processes on STAAR vary per item type. Across STAAR, 15 types of response interactions are available to measure student learning. For more information about the question types, refer to the STAAR Resources webpage.

TEA gathers theoretical and empirical evidence to confirm that the type of response required for each item does not add construct-irrelevant variance. TEA also gathers evidence from several sources to confirm that response processes do not result in an advantage or disadvantage for any student group. When new item types or changes to the format of existing item types are considered for STAAR, cognitive labs are used to study the way students engage with the various item presentations. After item types are determined to be appropriate for STAAR, evidence about student responses is gathered annually through educator and expert reviews and analyses of individual student responses to these items. During item reviews, educators evaluate whether the content for a given item type is being appropriately assessed and whether students will be able to accurately demonstrate their knowledge of the construct given the items' planned format. When items are field-tested, additional data are gathered about students' responses. Data such as item difficulty, item point-biserial correlations, and DIF are all evaluated regarding the item type. For additional information, refer to the Item Analysis section of Chapter 3, "Standard Technical Processes."

Scoring Process

The process used to score items can provide additional validity evidence based on response processes. This type of validity evidence is predicated on accurate scoring.

For all multiple-choice, multipart, and multiselect items on STAAR, statistical key checks are conducted during the equating process. A statistical key check is a procedure in which the statistical properties of all items on every test form are computed. Items whose statistics do not meet predetermined criteria are flagged for further review by content experts to verify that the items are correctly keyed and scored.

An adjudication process is used to ensure scoring reliability and validity for technology-enhanced items. During adjudication, data files that include all unique responses for each test question are analyzed to identify responses or questions that require more detailed analysis to ensure accurate, consistent scoring. Evaluators who specialize in STAAR content then review student responses to resolve scoring discrepancies or uncertainties.

For short and extended constructed-response questions, rubrics are used by human scorers to evaluate student responses. All rubrics for STAAR are validated by educator committees and content experts. In addition, TEA has implemented a rigorous scoring process for constructed responses that includes training and qualification requirements for scorers, ongoing monitoring during scoring, adjudication and resolution processes for student responses that do not meet the exact or adjacent scoring requirements, and rescoring of responses for which concerns have been raised by districts, campuses, or teachers regarding the assigned score. A more comprehensive description of the scoring process for constructed-response items is available in Chapter 2, “Building a High-Quality Assessment System.”

Score reliability for every STAAR assessment is generated and evaluated in terms of scorer agreement rates and the commonly used kappa with quadratic weights (Fleiss & Cohen, 1973). Constructed responses are scored with the adjacent agreement model. The exact agreement rate, adjacent agreement rate, and total agreement rate (exact and adjacent) between both scores are generated (refer to Table 4.7 and Table 4.8). When both scores are not in exact or adjacent agreement, the student response is adjudicated by a scoring leader.

Table 4.7. Summary of Scorer Agreement (Reliability) for Spring 2023 STAAR Extended Constructed Responses

Item Type	Number of Responses	Agreement Rate (%) after Two Scores			Quadratic Weighted Kappa
		Exact	Adjacent	Exact + Adjacent	
Grade 3 RLA					
Ideas	355,155	77%	21%	98%	0.77
Conventions	355,155	76%	23%	99%	0.71
Grade 4 RLA					
Ideas	363,493	69%	27%	96%	0.74
Conventions	363,493	73%	23%	96%	0.69

Item Type	Number of Responses	Agreement Rate (%) after Two Scores			Quadratic Weighted Kappa
		Exact	Adjacent	Exact + Adjacent	
Grade 5 RLA					
Ideas	371,792	63%	34%	97%	0.75
Conventions	371,792	67%	31%	98%	0.67
Grade 6 RLA					
Ideas	389,737	65%	32%	97%	0.80
Conventions	389,737	68%	29%	97%	0.72
Grade 7 RLA					
Ideas	398,777	65%	33%	98%	0.79
Conventions	398,777	67%	31%	98%	0.71
Grade 8 RLA					
Ideas	407,717	66%	32%	98%	0.83
Conventions	407,717	72%	26%	98%	0.77
English I					
Ideas	509,861	67%	31%	98%	0.83
Conventions	509,861	70%	28%	98%	0.75
English II					
Ideas	463,238	65%	32%	97%	0.81
Conventions	463,238	69%	28%	97%	0.72
Grade 3 Spanish RLA					
Ideas	30,059	82%	17%	99%	0.81
Conventions	30,059	82%	17%	99%	0.71
Grade 4 Spanish RLA					
Ideas	21,259	73%	25%	98%	0.79
Conventions	21,259	75%	24%	99%	0.75
Grade 5 Spanish RLA					
Ideas	15,961	68%	30%	98%	0.70
Conventions	15,961	73%	25%	98%	0.69

Table 4.8. Summary of Scorer Agreement (Reliability) for Spring 2023 STAAR Short Constructed Responses

Item Type	Number of Responses	Agreement Rate (%) after Two Scores			Quadratic Weighted Kappa
		Exact	Adjacent	Exact + Adjacent	
Grade 3 RLA					
Writing	354,892	92%	7%	99%	0.86
Grade 4 RLA					
Writing	363,116	92%	7%	99%	0.85
Grade 5 RLA					
Writing	371,782	89%	7%	100%	0.77
Grade 6 RLA					
Writing	389,880	89%	11%	100%	0.78
Grade 7 RLA					
Writing	399,103	87%	13%	100%	0.75
Grade 8 RLA					
Reading	408,977	77%	22%	99%	0.76
Writing	408,815	90%	10%	100%	0.79
English I					
Reading	512,642	85%	16%	100%	0.88
Writing	512,403	94%	6%	100%	0.89
English II					
Writing	465,604	93%	7%	100%	0.87
Grade 3 Spanish RLA					
Writing	30,298	92%	7%	99%	0.86
Grade 4 Spanish RLA					
Reading	21,663	82%	27%	99%	0.83
Writing	22,004	93%	7%	100%	0.86
Grade 5 Spanish RLA					
Reading	15,975	82%	18%	100%	0.83
Writing	16,181	92%	8%	100%	0.83
Grade 5 Science					
Question 1	378,366	87%	13%	100%	0.91
Grade 8 Science					
Question 1	405,505	84%	15%	99%	0.97
Question 2	406,151	97%	2%	99%	0.84

Item Type	Number of Responses	Agreement Rate (%) after Two Scores			Quadratic Weighted Kappa
		Exact	Adjacent	Exact + Adjacent	
Biology					
Question 1	455,603	92%	8%	100%	0.90
Question 2	459,859	89%	11%	99%	0.88
Grade 5 Spanish Science					
Question 1	9,763	91%	9%	100%	0.83
Social Studies Grade 8					
Question 1	411,961	81%	18%	99%	0.83
Question 2	411,625	83%	16%	99%	0.85
U.S. History					
Question 1	375,688	63%	37%	97%	0.62
Question 2	375,890	64%	36%	98%	0.62

Validity is evaluated through validity papers, which are student responses from the field test and current administrations that are representative of different levels of writing performance based on the scoring rubrics. Validity papers are identified by scoring leaders and are then systematically given to scorers throughout the scoring project. An important feature of validity papers is that they are not identifiable as such; in fact, they are indistinguishable from unscored student responses. Each person's daily scores on validity papers are compared with the approved scores. Validity papers are used throughout the scoring project as a primary quality-control measure, the purpose of which is to ensure that scorers are accurately and reliably scoring on a daily basis and across time. Validity agreement rate in Table 4.9 and Table 4.10 is expressed in terms of exact agreement between the score assigned by a given person and the true score approved by scoring leaders.

Table 4.9. Summary of Validity Results for Spring 2023 STAAR Extended Constructed Responses

STAAR Assessment and Trait	Exact Agreement Rate (%)
Grade 3 RLA	
Ideas	92%
Conventions	91%
Grade 4 RLA	
Ideas	85%
Conventions	88%
Grade 5 RLA	
Ideas	89%
Conventions	89%

STAAR Assessment and Trait	Exact Agreement Rate (%)
Grade 6 RLA	
Ideas	82%
Conventions	88%
Grade 7 RLA	
Ideas	81%
Conventions	78%
Grade 8 RLA	
Ideas	80%
Conventions	84%
English I	
Ideas	84%
Conventions	83%
English II	
Ideas	83%
Conventions	85%
Grade 3 Spanish RLA	
Ideas	94%
Conventions	92%
Grade 4 Spanish RLA	
Ideas	86%
Conventions	82%
Grade 5 Spanish RLA	
Ideas	82%
Conventions	83%

Table 4.10. Summary of Validity Results for Spring 2023 STAAR Short Constructed Responses

STAAR Assessment	Exact Agreement Rate (%)
English I Reading	95%
English I Writing	96%
English II Writing	97%
Biology Question 1	97%
Biology Question 2	98%
U.S. History Question 1	89%
U.S. History Question 2	94%

Evidence Based on Internal Structure

TEA collects evidence that shows the relationship of students' responses between items, within reporting categories of items, and within the full tests to verify that the elements of an assessment conform to the intended test construct and conducts internal consistency studies to gather evidence based on internal structure. The internal consistency of STAAR is evaluated using KR20 for assessments that have only dichotomously scored items. For the STAAR assessments that have a combination of dichotomous and polytomous items, internal consistency is evaluated using stratified coefficient alpha. These internal consistency evaluations are made for all students and for student groups such as female, male, Black or African American, Hispanic or Latino, and White students. Estimates of internal consistency are made for the full test, as well as for each reporting category within a content area, and can be found in Appendix B.

Evidence Based on Relationships to Other Variables

Another method TEA uses to provide validity evidence for STAAR, including STAAR Spanish, is analyzing the relationship between performance on STAAR and performance on other assessments, a process that supports criterion-related validity. Evidence can be collected to show that the empirical relationships are consistent with the expected relationships. Several past and current research studies have been designed to evaluate the relationship between performance on STAAR and performance on other related assessments or criteria, including the following:

- STAAR to TAKS comparison studies, which link performance on STAAR to performance on TAKS (e.g., STAAR grade 7 mathematics to TAKS grade 7 mathematics)
- STAAR linking studies, which link performance on STAAR across grade levels or courses in the same content areas (e.g., grade 4 RLA to grade 5 RLA, English I to English II)
- STAAR intercorrelation estimates, which evaluate the strength of the relationship (or lack thereof) among scores on STAAR across different content areas (e.g., grade 4 mathematics to grade 4 RLA, English I to Biology)
- grade correlation studies, which link performance on the STAAR EOC assessments to course grades
- validity studies, which link performance on STAAR to other measures (e.g., SAT, ACT, Lexiles, Quantiles, STAAR Interim Assessments)
- college students taking STAAR studies, which link performance on STAAR EOC assessments to college course grades

More information on comparisons between STAAR operational assessments and STAAR Interim Assessments is available in the STAAR Interim Assessments Summary Report. In addition, STAAR correlation estimates based on student performance on the spring administration are provided in Appendix B.

Evidence Based on the Consequences of Testing

Another method for providing validity evidence is by documenting the intended and unintended consequences of administering an assessment. The collection of consequential validity evidence typically occurs after a program has been in place for some time and on a regular basis.

Given the important stakes associated with STAAR (including STAAR Spanish), valid test scores are critical in supporting their intended interpretations and uses. The intended interpretations of STAAR results are stated in the policy definitions of the four performance levels. Refer to the Performance Standards section in this chapter for the policy definitions of the STAAR performance levels. Each performance level describes a student's knowledge and skills in a content area and a student's level of preparation for the next grade or course.

Student-Level Performance

The following are the intended uses of STAAR test scores based on the policy definitions for student-level performance:

- Performance on STAAR is one indicator of a student's level of proficiency in a content area or specific course.
- Performance on STAAR is one indicator of a student's readiness for the next grade level or course in the same content area.
- Performance on STAAR is one indicator of a student's possible need for academic intervention.
- Performance on STAAR across years provides one indicator of a student's academic progress within a content area.
- Performance on STAAR may provide information about expected student performance on external assessments, such as the ACT or SAT, that measure similar knowledge and skills.

District- or Campus-Level Performance

The following are the intended uses of STAAR test scores based on the policy definitions for district- or campus-level performance:

- STAAR performance results can be aggregated to provide one indicator of overall student proficiency at a district or campus.
- STAAR performance results can be aggregated to provide one indicator of overall student readiness (for the next grade level or course in the same content area) at a district or campus.
- STAAR performance results can be aggregated across years to provide one indicator of overall student academic progress at a district or campus.

Sampling

Sampling is a procedure that is used to select and examine a small set that is representative of the population from which it was drawn. STAAR uses two types of sampling: stratified random sampling and simple random sampling. Stratified random sampling used in stand-alone field testing ensures that subgroups of a given population are adequately represented within the whole sample. Simple random sampling is used to sample responses for field-test items that are scored by human scorers.

Test Results

Appendix B provides consistency and accuracy data, scale score correlations, CSEMs, mean *p*-values, scale score descriptive statistics, and frequency distributions for the spring STAAR administration. Pass rates for STAAR are available on the Statewide Summary Reports webpage.

TECHNICAL DIGEST 2022–2023

Chapter 5

STAAR Alternate 2

Overview

STAAR Alternate 2 is a standardized alternate academic achievement assessment based on alternate academic achievement standards and designed to measure the extent to which a student has learned and is able to apply the defined knowledge and skills in the TEKS. STAAR Alternate 2 is administered individually to students with the most significant cognitive disabilities who meet the participation requirements. STAAR Alternate 2 fulfills ESSA and IDEA. ESSA requires that all students be assessed in specific grades and subjects throughout their academic careers, whereas IDEA requires that students with disabilities have access to the same standards as their nondisabled peers and that they be included in statewide assessments.

STAAR Alternate 2 is not a traditional paper-pencil or multiple-choice test. Instead, it involves test administrators observing students as they respond to standardized, state-developed assessment questions that align to the grade-level TEKS through prerequisite skills. Teachers evaluate student performance based on standard scoring instructions embedded into each item on STAAR Alternate 2.

STAAR Alternate 2 was implemented in the 2014–2015 school year and includes the following assessments:

- grades 3–8 mathematics,
- grades 3–8 RLA,
- grades 5 and 8 science,
- grade 8 social studies, and
- EOC assessments for:
    - Algebra I,
    - English I,
    - English II,
    - Biology, and
    - U.S. History.

Due to the redesign of STAAR, STAAR Alternate 2 reading assessments were also redesigned to combine reading and writing into an RLA assessment for each grade. The redesigned STAAR Alternate 2 RLA assessments were implemented in spring 2023.

Participation Requirements

Students who receive special education services and have the most significant cognitive disabilities are eligible to participate in STAAR Alternate 2. These students exhibit significant intellectual and adaptive behavior deficits that limit their ability to plan, comprehend, and reason as well as adaptive behavior deficits that limit their ability to apply social and practical skills

(e.g., personal care, social problem-solving skills, dressing, eating, using money) across all life domains. Students with the most significant cognitive disabilities require extensive, direct, individualized instruction and have a need for substantial supports that are neither temporary nor content specific.

STAAR Alternate 2 has specific participation requirements that an ARD committee must carefully review and consider annually. The STAAR Alternate 2 Participation Requirements, available in English and Spanish on the STAAR Alternate 2 Resources webpage, detail the ARD committee's responsibility for ensuring that a student is eligible for STAAR Alternate 2. Prior to reviewing the eligibility criteria for STAAR Alternate 2, the ARD committee must understand all assessment options, including the characteristics of each assessment and the potential implications of each assessment choice. If STAAR Alternate 2 is being considered, the ARD committee must review the participation requirements against the supporting documentation within the student's individualized education program (IEP), such as in the present levels of academic achievement and functional performance, to determine eligibility.

Students in grades 3–8 who meet the participation requirements will take all applicable STAAR Alternate 2 subject assessments at their enrolled grade level.

Students in grades 9–12 who meet the participation requirements will take STAAR Alternate 2 EOC assessments—Algebra I, English I, English II, Biology, and U.S. History—as they are completing the corresponding course. The ARD committee makes educational decisions for a student with a disability, including decisions related to graduation requirements as described in TAC §89.1070.

In rare circumstances a student's ARD committee may determine prior to the administration of the assessment that the student will not participate in STAAR Alternate 2 because the student meets the eligibility criteria for a medical exception or no authentic academic response (NAAR). For both exceptions, the ARD committee reviews educational records and eligibility requirements. For more information, refer to the eligibility criteria on the STAAR Alternate 2 Resources webpage.

Test Development

Maintaining a high-quality student assessment program involves a complex and detailed test-development process, and TEA relies on input from educators to ensure that all measures of learning for Texas public school students are equitable and accurate. Test items for STAAR Alternate 2 are developed annually, reviewed by educator committees, field-tested, reviewed with their data, and, if approved, added to the STAAR Alternate 2 item bank. Newly developed items are embedded in STAAR Alternate 2 operational assessments each spring. For more information regarding each step of the STAAR Alternate 2 test-development process, refer to Chapter 2, which outlines the processes used to develop each STAAR Alternate 2 assessment's framework and explains ongoing test development.

For the initial development of STAAR Alternate 2, TEA sought input from educator committees and a statewide steering committee that included state assessment experts, parents, advocacy group representatives, related service providers, administrators, and ESC professionals. Consistent with the idea of universal design, particular attention was given to:

- students' response modes to allow students to show what they know and can do,
- differentiated supports and materials to allow students to access the content of the assessment, and
- multiple means of engagement to allow students more time to complete each task.

To ensure STAAR Alternate 2 is linked to grade-level TEKS, TEA worked with experts in test development, special education, and specific subject areas to develop vertical alignments for each content area and curriculum framework tools. The vertical alignments link content standards across grades, and the curriculum frameworks list the grade-level TEKS and the associated prerequisite skills for each grade and subject area. Essence statements, also known as strand statements in RLA, act as a bridge between grade-level content standards and STAAR Alternate 2 prerequisite skills. Specific essence statements are selected each year and provided to educators in the fall, giving them time to plan instruction and develop standards-based IEPs for that school year.

Accommodations

The goal of accommodations for STAAR Alternate 2 is to ensure that each student can interact appropriately with the content, presentation, and response modes of the state assessments. STAAR Alternate 2 is a standardized assessment intended to be appropriate for eligible students in its original, intact form. However, it is critical that students with disabilities be provided access to the assessment through careful use of accommodations wherever appropriate. Therefore, allowable accommodations may be provided to enable students with disabilities to participate meaningfully in the assessment.

Test administrators may use accommodations only if they are routinely provided in classroom instruction and listed in the student's IEP. Some accommodations provided during classroom instruction may not be allowed during testing, as certain accommodations used in the classroom would invalidate the content being assessed or compromise the security and integrity of the test. A list of allowable accommodations can be found in the *STAAR Alternate 2 Test Administrator Manual*, which is available on the STAAR Alternate 2 Resources webpage.

Training

TEA develops instructional materials, including manuals, presentations, online modules, and videos, to support the training of all testing personnel on test security and administration procedures. Preparation for test administration begins every year with a training-of-trainers session for testing coordinators from each of the 20 Texas regional ESCs as well as district testing coordinators from the state's 25 largest districts. Using materials and information provided in the TEA training session, ESC regional testing coordinators train the district coordinators in their respective regions. District coordinators then train their campus testing coordinators, who are responsible for training test administrators.

Test security and administration procedures provided in the *Coordinator Resources* and the *STAAR Alternate 2 Test Administrator Manual* must be followed so that all students have an equal opportunity to demonstrate their academic knowledge and skills. The *Coordinator*

Resources guide district and campus coordinators through their responsibilities as they oversee the administration of the Texas Assessment Program. This online resource contains preparation and administration procedures for each state-required assessment and is available prior to the annual ESC training.

In addition, TEA produces the *STAAR Alternate 2 Educator Guide*, available on the STAAR Alternate 2 Resources webpage, to familiarize educators with the assessment. The guide includes information on test design, alignment with state curriculum, training, and test results.

Test Administration

All STAAR Alternative 2 assessments—grades 3–8 mathematics, grades 3–8 RLA, grades 5 and 8 science, grade 8 social studies, Algebra I, English I, English II, Biology, and U.S. History—are administered on paper. The STAAR Alternate 2 testing window occurs over a five-week period during the spring, and retest opportunities are not offered. The number of students tested for each STAAR Alternate 2 assessment is shown in Table 5.1.

Table 5.1. STAAR Alternate 2 Assessments Administered in 2022–2023

STAAR Assessment	Assessments Administered
Grade 3 mathematics	7,615
Grade 3 RLA	7,617
Grade 4 mathematics	7,544
Grade 4 RLA	7,546
Grade 5 mathematics	7,084
Grade 5 RLA	7,083
Grade 5 science	7,081
Grade 6 mathematics	6,722
Grade 6 RLA	6,723
Grade 7 mathematics	6,557
Grade 7 RLA	6,561
Grade 8 mathematics	6,445
Grade 8 RLA	6,448
Grade 8 science	6,445
Grade 8 social studies	6,446
Algebra I	6,310
English I	6,325
English II	6,077
Biology	6,325
U.S. History	5,502

Each STAAR Alternate 2 test question measures a targeted prerequisite skill. A cluster of four test questions tests a common skill or concept at varying levels of difficulty. Five clusters make up a test form of 20 base test questions. Test forms also include one field-test cluster.

The assessment is designed with scripted presentation instructions that mirror instructional techniques for a student with the most significant cognitive disabilities. Student responses during a STAAR Alternate 2 test administration may be verbal, physical, or visual as appropriate for the student at the time of testing. Each question has a unique set of scoring instructions that describe what the student must do for his or her response to be marked correct. The test administrator must refer to the scoring instructions for each question to determine how to score the student's response.

STAAR Alternate 2 is scored polytomously using a standard scoring rubric with item score ranges from 0 to 2. Each item is scored according to the level of independence with which a student responds. The scoring rubric is as follows:

- If a student responds correctly to the first presentation of an item, he or she receives a score point of 2. If the student does not respond or responds incorrectly, the item is presented again with allowable teacher assists.
- If the student responds correctly to the second presentation of the item, he or she receives a score point of 1.
- If the student does not respond or responds incorrectly to the second presentation, he or she receives a score point of 0.

Performance Standards

Performance standards directly relate levels of test performance to what students are expected to learn, as defined in the statewide curriculum. Standard setting is the process of establishing cut scores that define the performance levels on an assessment.

Performance Levels and Policy Definitions

For STAAR Alternate 2, the performance levels and policy definitions are as follows:

Level I: Developing Academic Performance

Performance in this category indicates that students require additional instructional supports for accessing the curriculum through prerequisite skills. Students are able to acknowledge some concepts, but they demonstrate a minimal or inconsistent understanding of the knowledge and skills that are linked to content measured in this grade or course. Even with continued support, students in this category need significant intervention to show progress in the next grade or course.

Level II: Satisfactory Academic Performance

Performance in this category indicates that students are sufficiently prepared for the next grade or course with instructional supports for accessing the curriculum through prerequisite skills. Students demonstrate sufficient understanding of the knowledge and skills that are linked to

content measured at this grade or course. Students exhibit the ability to determine relationships, integrate multiple pieces of information, extend details, identify concepts, and match concepts that are similar. With continued support, students in this category have a reasonable likelihood of showing progress in the next grade or course.

Level III: Accomplished Academic Performance

Performance in this category indicates that students are well prepared for the next grade or course with instructional supports for accessing the curriculum through prerequisite skills. Students demonstrate a strong understanding of the knowledge and skills that are linked to content measured at this grade or course. Students exhibit the ability to use higher-level thinking and more complex skills, which includes making inferences and comparisons and solving multi-step problems. With support, students in this category have a high likelihood of showing progress in the next grade or course.

Standard Setting

Standards for all assessments were originally set for STAAR Alternate 2 in 2015. Standard setting for STAAR Alternate 2 involved 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. Standard-setting committees comprised diverse groups of stakeholders who carefully considered the interaction of these elements for each STAAR Alternate 2 assessment. The task of each standard-setting committee was to recommend cut scores that would define the performance levels for each STAAR Alternate 2 assessment. In 2023, performance standards were reset for STAAR Alternate 2 RLA based on the redesign of STAAR.

The current performance standards for STAAR Alternate 2 are provided in Table 5.2.

Table 5.2. Performance Levels for STAAR Alternate 2

Subject Area	Grade/Course	Level II: Satisfactory	Level III: Accomplished
Reading Language Arts	Grade 3	300	388
	Grade 4	300	380
	Grade 5	300	374
	Grade 6	300	370
	Grade 7	300	378
	Grade 8	300	371
	English I	300	365
	English II	300	370

Subject Area	Grade/Course	Level II: Satisfactory	Level III: Accomplished
Mathematics	Grade 3	300	375
	Grade 4	300	387
	Grade 5	300	379
	Grade 6	300	373
	Grade 7	300	375
	Grade 8	300	365
	Algebra I	300	361
Science	Grade 5	300	387
	Grade 8	300	382
	Biology	300	383
Social Studies	Grade 8	300	372
	U.S. History	300	368

Refer to the standard setting technical reports, which are available on the Assessment Reports and Studies webpage, for more information.

Scores and Reports

TEA publishes resources on both the TEA and Texas Assessment websites to assist school personnel in understanding and interpreting student performance data and to help parents understand their child's STAAR Alternate 2 results. School personnel can access STAAR Alternate 2 results through CRS, parents can access their child's STAAR Alternate 2 results in the Family Portal, and the public can access STAAR Alternate 2 statewide, region, district, and campus data using the Research Portal.

TEC §39.030 and TAC §101.3014 specify the requirements for maintaining the confidentiality of individual student results and for reporting district-level and campus-level results. The results of individual student performance on state assessments are confidential and may be released only in accordance with FERPA. Districts must provide each student's state assessment results to the student, to his or her parent or guardian, and to his or her teacher for the applicable subject area. In addition, all state assessment results must be included in each student's academic achievement record.

Description of Scores

Scores for STAAR Alternate 2 include raw scores, scale scores, and the resulting performance level associated with the student's score. The number of points that a student earns on a STAAR Alternate 2 assessment is the student's raw score. A scale score is a conversion of the raw score onto a scale that is common to all test forms for that assessment. The scale score is used to determine whether a student achieved the Level I: Developing Academic Performance, Level II: Satisfactory Academic Performance, or Level III: Accomplished Academic Performance standard. Refer to Chapter 3, "Standard Technical Processes," for more information about raw scores and scale scores.

Assessment Reports

TEA provides reports of student performance on STAAR Alternate 2 to all Texas public school districts and open-enrollment charter schools. For STAAR Alternate 2, TEA provides student report cards, student labels, campus rosters, summary reports, and reporting data files. In addition, TEA periodically releases STAAR Alternate 2 assessments, which can be found on the STAAR Alternate 2 Released Test Questions webpage.

For more information about scoring and reporting for STAAR Alternate 2, refer to the Interpreting Results page of the *Coordinator Resources*.

Use of Test Results

Test results can be used to evaluate the performance of a group over time. Average scale scores and the percentage of students meeting the Level I, Level II, and Level III performance standards can be analyzed by grade and content area across administrations to provide insight into whether student performance is improving across years. For example, the average scale score for students who took the STAAR Alternate 2 grade 4 mathematics test can be compared over time.

Test results can be used when evaluating instruction or programs that require average-score or year-to-year comparisons. The tests are designed to measure content areas within the required state curriculum, so the consideration of test results by content area and reporting category might be helpful when evaluating curriculum and instructional programs. All test scores can be compared with statewide and regional performance within the same content area for any administration.

Test scores can also be used to identify where an individual student needs additional instruction or support in each subject. This identification can help educators plan the most effective instructional intervention. Finally, individual student test scores are also used in conjunction with other performance indicators to assist in making placement decisions. While scores can contribute to decisions regarding placement, educational planning for a student should take into account as much student information as possible.

Generalizations from test results can be made from the specific content area being measured on the test. However, because each test measures a finite set of skills with a limited set of items, any generalizations about student achievement derived solely from a particular test should be made with great caution and with full reference to the fact that the conclusions are based only on that test. Instruction and program evaluations should take into account as much information as possible, rather than relying on test scores alone, to provide a more complete picture of student performance.

Standard Setting

Standards for all assessments were originally set for STAAR Alternate 2 in 2015. Standard setting for STAAR Alternate 2 involved 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. Standard-setting committees comprised diverse groups of stakeholders who carefully considered the interaction of these elements for each STAAR Alternate 2 assessment. The task of each standard-setting committee was to recommend cut scores that would define the performance levels for each STAAR Alternate 2 assessment. In 2023, performance standards were reset for STAAR Alternate 2 RLA based on the redesign of STAAR.

The current performance standards for STAAR Alternate 2 are provided in Table 5.2.

Table 5.2. Performance Levels for STAAR Alternate 2

Subject Area	Grade/Course	Level II: Satisfactory	Level III: Accomplished
Reading Language Arts	Grade 3	300	388
	Grade 4	300	380
	Grade 5	300	374
	Grade 6	300	370
	Grade 7	300	378
	Grade 8	300	371
	English I	300	365
	English II	300	370
Mathematics	Grade 3	300	375
	Grade 4	300	387
	Grade 5	300	379
	Grade 6	300	373
	Grade 7	300	375
	Grade 8	300	365
	Algebra I	300	361
Science	Grade 5	300	387
	Grade 8	300	382
	Biology	300	383
Social Studies	Grade 8	300	372
	U.S. History	300	368

Refer to the standard setting technical reports, which are available on the Assessment Reports and Studies webpage, for more information.

Measures of Student Progress

Student progress measures provide information beyond performance level by considering performance over time. Whereas performance-level information describes students' current achievement, progress measures describe students' achievement in adjacent years.

STAAR Alternate 2 Progress Measure

The STAAR Alternate 2 Progress Measure was reported for the first time in the 2015–2016 school year. For STAAR Alternate 2, progress is measured based on a student's stage change from the prior year to the current year. Stage change is determined by: 1) classifying the

student's scores from the previous school year and the current school year in terms of the stage of performance achieved and then 2) comparing the stages from year to year. Student progress is then categorized as Did Not Meet, Met, or Exceeded. These progress targets define the expectation of annual progress for each grade and content area. The progress targets are grounded in the STAAR Alternate 2 performance standards.

Steps for calculating a student's stage change and progress indicator for the STAAR Alternate 2 progress measure can be found in *STAAR Alternate 2 Progress Measure Questions and Answers* on the Progress Measures webpage.

Progress measures were not calculated or reported for the 2022–2023 school year.

STAAR Alternate 2 On-Track Measure

The STAAR Alternate 2 on-track measure examines a student's progress and projects where that student will be in a future target year if that student continues to make progress at the same rate over future years. The student is then classified as On Track or Not On Track to achieve Level II: Satisfactory in the target year.

If a student has scores for STAAR Alternate 2 in two consecutive grades and subject or courses in two consecutive years, the on-track measure can be calculated for the student. If any of the required information for STAAR Alternate 2 on-track measure calculation is lacking, the on-track measure is not available. This includes students who have received exceptions through the medical exception or NAAR policies in the previous or current grade.

Steps for calculating a student's STAAR Alternate 2 on-track measure can be found in *STAAR Alternate 2 On-Track Measure Questions and Answers* on the Progress Measures webpage.

Scaling

Scaling is a statistical procedure that places raw scores on a common scoring metric to make test scores comparable across test administrations. Scaling associates numbers with characteristics of interest to provide information about measurable quantities for those characteristics. STAAR Alternate 2 uses the RPCM to place test items on the same Rasch scale across administrations for a given assessment. Once performance standards have been set for an assessment, the Rasch scale is then transformed to a more user-friendly metric to facilitate interpretation of the test scores. Details of the RPCM scaling method are provided in Chapter 3, "Standard Technical Processes."

Reporting Scales

STAAR Alternate 2 scale scores are reported on a horizontal scale. Horizontal scale scores allow for direct comparisons of student performance between specific sets of test items from different test administrations for a specific grade and subject or course.

For all STAAR Alternate 2 assessments, a scale score of 300 represents the Level II: Satisfactory performance standard. The desired standard deviation for each grade and subject and course is 50. The Level III scale score values vary across STAAR Alternate 2 assessments, but for any given assessment, performance standards remain constant over time.

STAAR Alternate 2 scale scores represent linear transformations of Rasch proficiency-level estimates (θ). Specifically, the transformation is made by first multiplying θ by a slope constant (*A*) and then adding an intercept constant (*B*). This operation is described by the following equation:

$$SS_\theta = A \times \theta + B,$$

where SS_θ is the scale score for a Rasch proficiency-level estimate (θ) and *A* and *B* are the horizontal scaling constants. These same transformations will be applied each year to the Rasch proficiency-level estimates (θ) for that year's set of test items. Values for the horizontal scaling constants for STAAR Alternate 2 are provided in Table 5.3.

Table 5.3. Horizontal Scaling Constants for STAAR Alternate 2

Subject Area	Grade/Course	*A*	*B*
Mathematics	Grade 3	43.9599	297.2305
	Grade 4	42.3406	297.9677
	Grade 5	42.9221	293.4758
	Grade 6	47.3082	293.8972
	Grade 7	45.0653	292.6994
	Grade 8	45.9897	283.5357
	Algebra I	46.1042	287.8285
Reading Language Arts	Grade 3	51.7409	300.6002
	Grade 4	52.5281	289.7045
	Grade 5	52.1646	285.3261
	Grade 6	52.7711	284.9813
	Grade 7	53.4243	290.3676
	Grade 8	50.2019	283.8651
	English I	49.3225	294.3526
	English II	49.5385	294.8480
Science	Grade 5	43.8943	291.6601
	Grade 8	38.5892	298.4950
	Biology	38.2614	293.1129
Social Studies	Grade 8	41.4662	282.7501
	U.S. History	41.3565	283.7055

Equating

Used in conjunction with the scaling process, equating is the process that considers the differences in difficulty across test forms and administrations and allows scores to be placed onto a common scale. By using statistical methods, TEA equates the results of different test forms so that scale scores across test forms and test administrations can be compared. TEA uses pre-equating and post-equating for all STAAR Alternate 2 assessments.

To replenish the item bank as new tests are created each year, newly developed items must be field-tested and equated to the item bank scale. During each spring administration, field-test equating is conducted for STAAR Alternate 2 through an embedded-field-test design for all tests.

Refer to Chapter 3, "Standard Technical Processes," for detailed information about equating.

Reliability

Reliability indicates the precision of test scores, referring to the expectation that repeated administrations of the same test should generate consistent results. Reliability for STAAR Alternate 2 test scores is estimated using statistical measures including internal consistency, classical SEM, CSEM, and classification consistency and accuracy. Data for these statistical measures from the spring STAAR Alternate 2 administration are provided in Appendix C. Refer to Chapter 3, “Standard Technical Processes,” for detailed information about reliability.

Validity

Validity refers to the extent to which test scores accurately measure what the test is intended to measure. TEA follows national standards of best practice and annually collects validity evidence to support the interpretations and uses of STAAR Alternate 2 test scores. TTAC, a panel of national testing experts created specifically for the Texas Assessment Program, provides ongoing input to TEA about STAAR Alternate 2 validity evidence. The following sections describe how validity evidence has been collected for STAAR Alternate 2. Refer to Chapter 3, “Standard Technical Processes,” for additional information about validity.

Evidence Based on Test Content

Validity evidence based on test content refers to evidence of the relationship between tested content and the construct that the assessment is intended to measure. STAAR Alternate 2 has been developed to align with content as defined by the TEKS through prerequisite skills. Content validity evidence is collected at all stages of the test-development process. Nationally established test-development processes for the Texas Assessment Program are followed while developing STAAR Alternate 2. This supports the use of STAAR Alternate 2 scores in making inferences about students’ knowledge and understanding of the TEKS.

Relationship to the Statewide Curriculum

The TEKS are designed to ensure that Texas students receive a solid education that will enable them to be successful in life, whether they choose to pursue higher education or enter the workforce directly after graduation. STAAR Alternate 2 assesses the TEKS through prerequisite skills.

In 2015–2016, an independent third-party analysis of the alignment between items on the 2016 STAAR Alternate 2 tests and the TEKS was conducted to inform TEA about the degree of alignment between the test items and curriculum standards. The study concluded that the 2016 STAAR Alternate 2 items demonstrated strong linkages across all grades and subjects and courses. All items were found to have an academic foundation and to have content connections to the grade-level student expectations.

Educator Input

As part of the initial development of STAAR Alternate 2, teachers, curriculum specialists, special education experts, test development specialists, and TEA staff worked together in advisory

committees to identify the best way to assess students with the most significant cognitive disabilities. The input of the advisory committees was reflected in the vertical alignment documents, prerequisite skills, essence statements, and test blueprints. In addition, prototype items were developed for the assessments early in the development process. The educator advisory committees reviewed these prototypes to identify how well the items would measure the TEKS through the prerequisite skills to which the items were aligned. These early reviews provided valuable suggestions for item development guidelines and item types. Item development guidelines continued to be refined through the test development process as various STAAR Alternate 2 item review committees of educators shared their feedback about how the TEKS could be effectively assessed.

As part of the annual process of item development, committees of Texas educators meet to review STAAR Alternate 2 items and confirm that each item appropriately measures the TEKS through prerequisite skills. These committees also review items for content and bias. Item review committees are composed of Texas educators, including special education teachers, and these committees revise and edit items, as appropriate, prior to field testing. Item review committees are convened for all STAAR Alternate 2 assessments.

Test Developer Input

Item writers and reviewers, including content experts and special education experts, follow test development guidelines and item specifications that explain how the content of the assessed TEKS should be measured. At each stage of development, writers and reviewers verify the alignment of the test items with the assessed essence statements.

Evidence Based on Response Processes

Response processes refers to the cognitive behaviors that are required to respond to a test item. TEA collects evidence to show that the way students respond to items on STAAR Alternate 2 reflects accurate measurement of the construct.

Items

TEA gathers theoretical and empirical evidence supporting the expectation that the way students respond to test items does not add construct-irrelevant variance. During yearly item reviews, educators evaluate whether the content for a given item is being appropriately assessed and whether students will be able to accurately demonstrate their knowledge of the construct given the items’ planned format. When items are field-tested, additional data are gathered about students' responses. Data such as item difficulty, item-total correlations, and item fit are all evaluated. For additional information, see the Item Analysis section of Chapter 3, “Standard Technical Processes.”

Scoring Process

The process used to score items can provide additional validity evidence based on response processes. This type of validity evidence is predicated on accurate scoring. The test administrator booklet provides test administrators with exact scoring rules and scripted instructions on how to present every item to a student. Test administrators are provided with resources to prepare for a STAAR Alternate 2 test administration, including a scheduled period

directly before the testing window in which they can preview the test booklet to prepare for a valid test administration.

Evidence Based on Internal Structure

TEA collects evidence that shows the relationship between items and reporting categories to verify that the elements of an assessment conform to the intended test construct and conducts internal consistency studies to gather evidence based on internal structure. The internal consistency of STAAR Alternate 2 is evaluated using coefficient alpha. These internal consistency evaluations are made for all students and for student groups such as female, male, Black or African American, Hispanic or Latino, and White.

Evidence Based on Relationships to Other Variables

Another method TEA uses to provide validity evidence for STAAR Alternate 2 is analyzing the relationship between performance on STAAR Alternate 2 and performance on other assessments, a process that supports criterion-related validity. Evidence can be collected to show that the empirical relationships are consistent with the expected relationships. STAAR Alternate 2 correlation estimates are calculated to evaluate the strength of the relationship (or lack thereof) among scores on STAAR Alternate 2 assessments across different content areas (e.g., grade 4 mathematics to grade 4 RLA, English I to Biology).

Evidence Based on Consequences of Testing

Another method for providing validity evidence is by documenting the intended and unintended consequences of administering an assessment. The collection of consequential validity evidence typically occurs on a regular basis after a program has been in place for some time. Some of the intended consequences of STAAR Alternate 2 are as follows:

- Students with the most significant cognitive disabilities can receive challenging instruction that is linked to state content standards.
- Students with the most significant cognitive disabilities can be included in state assessment programs.
- STAAR Alternate 2 can assess the achievement of students with the most significant cognitive disabilities.

Given the important stakes associated with STAAR Alternate 2, valid test scores are critical in supporting their intended interpretations and uses. The intended interpretations of STAAR Alternate 2 results are stated in the policy definitions of the three performance levels. Refer to the Performance Standards section in this chapter for the policy definitions of the STAAR Alternate 2 performance levels. Each performance level describes a student's knowledge and skills in a content area and a student's level of preparation for the next grade or course.

Student-Level Performance

The following are the intended uses of STAAR Alternate 2 test scores based on the policy definitions for student-level performance:

- Performance on STAAR Alternate 2 is one indicator of a student's level of proficiency in a content area or specific course.
- Performance on STAAR Alternate 2 is one indicator of a student's readiness for the next grade or course in the same content area.
- Performance on STAAR Alternate 2 is one indicator of a student's possible need for academic intervention.
- Performance on STAAR Alternate 2 across years provides one indicator of a student's academic progress within a content area.

District- or Campus-Level Performance

The following are the intended uses of STAAR Alternate 2 test scores based on the policy definitions for district- or campus-level performance:

- STAAR Alternate 2 performance results can be aggregated to provide one indicator of overall student proficiency at a district or campus.
- STAAR Alternate 2 performance results can be aggregated to provide one indicator of overall student readiness (for the next grade or course in the same content area) at a district or campus.
- STAAR Alternate 2 performance results can be aggregated across years to provide one indicator of overall student academic progress at a district or campus.

Sampling

Sampling is a procedure that is used to select and examine a small set that is representative of the population from which it was drawn. For the STAAR Alternate 2 administration, campus assignment of forms uses an annual sampling process wherein a single form is assigned to each campus. A sample of students who represent the state demographic makeup respond to each form. This approach ensures that each campus administers the same form to all students and that teachers need only administer a single form.

Test Results

Appendix C provides consistency and accuracy data, scale score correlations, CSEMs, mean *p*-values, scale score descriptive statistics, and frequency distributions for the spring STAAR Alternate 2 administration. Pass rates for STAAR Alternate 2 are available on the Statewide Summary Reports webpage.

TECHNICAL DIGEST 2022–2023

Chapter 6

Texas English Language Proficiency Assessment System

Overview

TELPAS is an English language proficiency assessment that measures the progress that EB students make in acquiring the English language. It fulfills the requirements of ESSA, which requires that all EB students be assessed annually until they are determined to be proficient in the English language.

TELPAS assesses EB students in kindergarten through grade 12 in the language domains of listening, speaking, reading, and writing. For grades 2–12, TELPAS consists of online assessments for listening and speaking and for reading and writing. For kindergarten and grade 1, holistically rated assessments based on ongoing classroom observations are used for all four language domains.

Participation Requirements

All EB students in kindergarten through grade 12 are required to participate in TELPAS unless the student meets the participation requirements for TELPAS Alternate. EB students are assessed annually in English language proficiency until they are determined to be proficient by meeting the EB reclassification criteria. This includes students classified as emergent bilingual (EB)/English learner (EL) in the Public Education Information Management System (PEIMS) whose parents have declined bilingual or ESL program services.

In rare cases, it might be necessary for the ARD committee, in conjunction with the LPAC, to determine that an EB student receiving special education services should not be assessed in listening, speaking, reading, or writing for reasons associated with the student's disability. Participation must be considered on a domain-by-domain basis. The reason for not assessing the student must be related to the student's disability and be well supported and documented in the student's IEP by the ARD committee and in the student's permanent record file by the LPAC.

Test Development

Maintaining a high-quality student assessment program involves a complex and detailed test-development process, and TEA relies on input from educators to ensure that all measures of learning for Texas public school students are equitable and accurate. For more information regarding each step of the TELPAS test-development process, refer to Chapter 2, "Building a High-Quality Assessment System," which outlines the processes used to develop each TELPAS assessment's framework and explains ongoing test development.

Test items for TELPAS online assessments are developed annually, reviewed by educator committees, embedded in operational assessments each spring for field testing, reviewed with their data, and, if approved, added to the TELPAS item bank. TELPAS grades 2–12 online assessments were developed as combined listening and speaking assessments for multiple grade bands and combined reading and writing assessments for specific grades and grade bands, as shown in Table 6.1.

Table 6.1. TELPAS Grades 2–12 Online Assessments

Listening and Speaking	Reading and Writing
Grades 2–3	Grade 2
Grades 4–5	Grade 3
Grades 6–8	Grades 4–5
Grades 9–12	Grades 6–7
	Grades 8–9
	Grades 10–12

TEA developed the TELPAS holistically rated components in collaboration with test development experts, bilingual and ESL consultants, and focus group members including teachers, bilingual and ESL directors, assessment directors, campus administrators, and university professors. Like the TELPAS grades 2–12 assessments, these assessments align with the ELPS, assessing the English communication skills that EB students need to engage meaningfully and effectively in learning the academic knowledge and skills required by the TEKS. The holistically rated assessments draw on second language acquisition research, research-based standards, the experience of Texas educators, and observational assessment practices.

More information about the development of TELPAS is available in the *TELPAS Educator Guide* available on the TELPAS Resources webpage. Provided to familiarize educators with TELPAS, the guide shows the integral relationship between TELPAS and the ELPS. It explains the TELPAS language domains of listening, speaking, reading, and writing and provides examples of classroom instruction and annotated test item descriptions.

Training

TEA develops instructional materials, including manuals, guides, presentations, online modules, and videos, to support the training of all testing personnel on test security and administration procedures. Preparation for test administration begins every year with a TEA-provided training-of-trainers session for testing coordinators from each of the 20 Texas regional ESCs as well as district testing coordinators from the state's 25 largest districts. Using materials and information provided in the TEA training session, ESC regional testing coordinators train the district coordinators in their respective regions. District coordinators then train their campus testing coordinators, who are responsible for training test administrators.

Test security and administration procedures provided in the *Coordinator Resources*, the *TELPAS Test Administrator Manual*, and the *TELPAS Rater Manual* must be followed so that all students have an equal opportunity to demonstrate their knowledge and skills. The *Coordinator Resources* guide district and campus coordinators through their responsibilities as they oversee the administration of the Texas Assessment Program. This online resource contains preparation and administration procedures for each state-required assessment and is available prior to the annual ESC training.

In addition, TEA produces the *TELPAS Educator Guide* to familiarize educators with the assessment. The guide includes information on test design, alignment with the ELPS, training, and test results.

TELPAS raters must have trained and calibrated successfully before rating students. The training that TELPAS raters receive supports the administration of TELPAS and provides teachers with ongoing professional development to support effective implementation of the ELPS.

The Online Basic Training course teaches new raters the essentials of second language acquisition theory and how to use the ELPS PLDs to accurately identify the English language proficiency levels of EB students based on how well the students understand and use English during daily academic instruction and classroom interaction. The trainings are specific to grade clusters, and raters should complete the holistic rating training in the grade cluster that corresponds to the grade levels of the students they will rate. Online courses for kindergarten through grade 1 contain numerous practice rating activities composed of student writing samples and video segments in which EB students demonstrate their listening, speaking, reading, and writing skills in authentic Texas classroom settings. The courses give raters practice applying the scoring rubrics (i.e., PLDs) and provide detailed feedback about their rating accuracy.

New raters are required to successfully complete the online holistic rating trainings and separate practice activities for the grade cluster they are assigned before they access rater calibration activities and must then complete the online calibration activities to demonstrate their ability to apply the PLD rubrics consistently and accurately before they rate students for the operational assessment. There are two sets of calibration activities, and all applicable language domains are represented. Raters finish the calibration activities when they demonstrate sufficient accuracy. If sufficient accuracy is not obtained on the first set, the rater attempts a second and final online calibration set. Individuals not successful on the final set are either not used as raters or are provided rater support in accordance with test administration procedures. More information about TELPAS rater training can be found on the TELPAS Resources webpage.

Test Administration

TELPAS is administered once a year, in the spring, during a six-week testing window. The number of TELPAS assessments that were administered to eligible students in 2022–2023 is indicated in Table 6.2.

Table 6.2. TELPAS Assessments Administered in 2022–2023

Grade	Listening	Speaking	Reading	Writing
Kindergarten	96,717	96,519	96,425	96,449
Grade 1	103,723	103,510	103,341	103,338
Grade 2	100,204	100,196	100,259	100,251
Grade 3	100,597	100,591	100,651	100,650

Grade	Listening	Speaking	Reading	Writing
Grade 4	101,659	101,649	101,693	101,692
Grade 5	103,145	103,142	103,176	103,176
Grade 6	100,388	100,383	100,455	100,455
Grade 7	98,437	98,432	98,513	98,514
Grade 8	95,401	95,395	95,483	95,480
Grade 9	96,472	96,467	96,465	96,466
Grade 10	77,503	77,502	77,622	77,616
Grade 11	55,395	55,394	55,459	55,460
Grade 12	45,346	45,346	45,476	45,476
Total	1,174,987	1,174,526	1,175,018	1,175,023

Holistic Assessments

A holistically rated assessment process is used for kindergarten and grade 1 for all four language domains. To conduct these assessments, raters are specially trained to use the ELPS PLDs as holistic rating rubrics and determine the English language proficiency of EB students based on classroom observations and daily interactions with students. EB students in grades 2–12 may qualify for special holistic administrations of TELPAS listening, speaking, or writing, which follow this same process.

Online Assessments

EB students in grades 2–12 take two online TELPAS assessments—one combined assessment for listening and speaking and one combined assessment for reading and writing. In addition to a special holistic administration of listening, speaking, and writing, an EB student may qualify for a special paper administration of TELPAS reading.

The Test Delivery System

TELPAS online assessments are administered using TDS. TDS includes the Test Administrator Interface, which is used for scheduling test sessions; the Student Interface, which enables students to participate in testing; and the Secure Browser application, which provides a secure online environment for testing. TDS allows for the secure transfer and storage of test data while remaining scalable to support the student testing population. The TDS architecture has demonstrated stability and efficiency by seamlessly handling over 1.2 million concurrent users.

Make-up Testing

Make-up testing opportunities for students who are absent on the scheduled day of testing are available during the TELPAS testing window for all grades and domains.

Proficiency Standards

For TELPAS holistically rated assessments, proficiency standards are established through descriptions of student performance in the scoring rubrics and student exemplars used in scorer training. The scoring rubrics are the ELPS PLDs, and the student exemplars are the student writing collections and student videos used in rater training.

For TELPAS online assessments, proficiency standards are established by determining the score students need to obtain to be classified into specified performance categories. The proficiency categories are the proficiency levels described in the ELPS.

Proficiency Levels and Policy Definitions

As an English language proficiency assessment, TELPAS provides an indicator of where EB students are on a continuum of English language development. This continuum is divided into four proficiency levels: Beginning, Intermediate, Advanced, and Advanced High.

Beginning

Beginning students have little or no ability to understand and use English. They may know a little English but not enough to function meaningfully in social or academic settings.

Intermediate

Intermediate students have some ability to understand and use English. They can function in social and academic settings as long as the tasks require them to understand and use simple language structures and high-frequency vocabulary in routine contexts.

Advanced

Advanced students are able to engage in grade-appropriate academic instruction in English, although ongoing second language acquisition support is needed to help them understand and use grade-appropriate language. These students function beyond the level of simple, routinely used English.

Advanced High

Advanced high students have attained the command of English that enables them, with minimal second language acquisition support, to engage in regular all-English academic instruction at their grade level.

Standard Setting

Initial proficiency standards for TELPAS reading were established in 2008. Proficiency standards for TELPAS listening and speaking were established in 2018 and were reset for TELPAS reading with the shift to online assessments for listening and speaking. In 2023, proficiency standards for TELPAS writing were established with the shift to online writing assessments. The current proficiency standard ranges for TELPAS are provided in Table 6.3.

Table 6.3. TELPAS Proficiency Standards

Domain	Grade or Grade Band	Beginning	Intermediate	Advanced	Advanced High
Listening	Grades 2–3	1000 to 1441	1442 to 1524	1525 to 1599	1600 to 2000
	Grades 4–5	1000 to 1455	1456 to 1524	1525 to 1599	1600 to 2000
	Grades 6–8	1000 to 1430	1431 to 1524	1525 to 1599	1600 to 2000
	Grades 9–12	1000 to 1447	1448 to 1524	1525 to 1599	1600 to 2000
Speaking	Grades 2–3	1000 to 1410	1411 to 1524	1525 to 1599	1600 to 2000
	Grades 4–5	1000 to 1466	1467 to 1524	1525 to 1599	1600 to 2000
	Grades 6–8	1000 to 1459	1460 to 1524	1525 to 1599	1600 to 2000
	Grades 9–12	1000 to 1484	1485 to 1524	1525 to 1599	1600 to 2000
Reading	Grade 2	1000 to 1439	1440 to 1524	1525 to 1599	1600 to 2000
	Grade 3	1000 to 1434	1435 to 1524	1525 to 1599	1600 to 2000
	Grades 4–5	1000 to 1430	1431 to 1524	1525 to 1599	1600 to 2000
	Grades 6–7	1000 to 1446	1447 to 1524	1525 to 1599	1600 to 2000
	Grades 8–9	1000 to 1437	1438 to 1524	1525 to 1599	1600 to 2000
	Grades 10–12	1000 to 1426	1427 to 1524	1525 to 1599	1600 to 2000
Writing	Grade 2	1000 to 1431	1432 to 1524	1525 to 1599	1600 to 2000
	Grade 3	1000 to 1400	1401 to 1524	1525 to 1599	1600 to 2000
	Grades 4–5	1000 to 1408	1409 to 1524	1525 to 1599	1600 to 2000
	Grades 6–7	1000 to 1428	1429 to 1524	1525 to 1599	1600 to 2000
	Grades 8–9	1000 to 1412	1413 to 1524	1525 to 1599	1600 to 2000
	Grades 10–12	1000 to 1445	1446 to 1524	1525 to 1599	1600 to 2000

Refer to the TELPAS standard-setting technical reports, which are available on the Assessment Reports and Studies webpage, for more information.

Scores and Reports

TEA publishes resources on both the TEA and Texas Assessment websites to assist school personnel in understanding and interpreting student performance data and to help parents understand their child's TELPAS results. School personnel can access TELPAS test results through CRS, parents can access their child's TELPAS results in the Family Portal, and the public can access TELPAS statewide, region, district, and campus data using the Research Portal.

TEC §39.030 and TAC §101.3014 specify the requirements for maintaining the confidentiality of individual student results and for reporting district-level and campus-level results. The results of individual student performance on state assessments are confidential and may be released only in accordance with FERPA. Districts must provide each student's state assessment results to the student, to his or her parent or guardian, and to his or her teacher for the applicable subject

area. In addition, all state assessment results must be included in each student's academic achievement record.

Description of Scores

Results for TELPAS include proficiency-level ratings for each domain, composite scores, composite ratings, and yearly progress indicators.

For TELPAS online assessments, scores include raw scores and scale scores. The number of points that a student earns on a TELPAS assessment is the student's raw score. A scale score is a conversion of the raw score onto a scale that is common to all test forms for that assessment. The scale score is used to determine whether a student achieved the Beginning, Intermediate, Advanced, or Advanced High proficiency level. Refer to Chapter 3, "Standard Technical Processes," for more information about raw scores and scale scores.

Yearly Progress Indicator

The student's yearly progress indicator provides information about the yearly proficiency-level progress that an EB student makes in acquiring the English language. This measure is based on a comparison of a student's composite rating in the previous year with his or her composite rating in the current year. The yearly statewide summary reports provide the number and percentage of students who progressed one, two, or three proficiency levels. The yearly statewide summary reports also provide the number and percentage of students who progressed at least one proficiency level. The yearly progress indicator is set as follows:

- If a student received a composite rating one level higher than the previous year, the student's yearly progress indicator is 1. Additionally, if a student received an Advanced High composite rating in the current year and in the previous year, the student's yearly progress indicator is also 1.
- If a student received a composite rating two levels higher than the previous year, the student's yearly progress indicator is 2.
- If a student received a composite rating three levels higher than the previous year, the student's yearly progress indicator is 3.
- If a student received a current year composite rating that is the same as the previous year's composite rating (excluding an Advanced High composite rating) or lower than the previous year's rating, the yearly progress indicator is 0.

The yearly progress indicator was not calculated in 2023 due to the shift to online writing assessments.

Composite Score and Rating

In addition to receiving a proficiency-level rating for each domain, students also receive a composite score and composite rating. TELPAS composite scores and ratings indicate a student's overall level of English language proficiency and are determined from the student's

listening, speaking, reading, and writing proficiency ratings. To calculate the composite score, the proficiency rating for each of the domains is converted to a domain score from 1 (Beginning) to 4 (Advanced High). The domain scores are equally weighted, as shown in Table 6.4, and added for one composite score.

Table 6.4. Language Domain Weights for TELPAS Composite Scores

Listening	Speaking	Reading	Writing
25%	25%	25%	25%

After a composite score is calculated, a composite rating is determined according to the rules below. All criteria listed for a particular rating must be met for a student to receive that rating.

Beginning:

- a composite score that fails to meet the Intermediate requirements

Intermediate:

- a composite score of 1.5 or higher
- a minimum proficiency level of Intermediate in at least half of the domains in which the student was assessed

Advanced:

- a composite score of 2.5 or higher
- a minimum proficiency level of Intermediate in all domains
- a minimum proficiency level of Advanced in at least half of the domains in which the student was assessed

Advanced High:

- a composite score of 3.5 or higher
- a minimum proficiency level of Advanced in all domains

Figure 6.1 provides a student example to show how composite results are generated. Each domain rating is converted to a domain score from 1 (Beginning) to 4 (Advanced High).

Figure 6.1. Sample Calculation of Composite Results All Domains Assessed

Domain	Proficiency Level	Domain Score
Listening	Advanced	3
Speaking	Intermediate	2
Reading	Advanced	3
Writing	Intermediate	2

The domain scores are multiplied by the appropriate weight in Table 6.3 and then added together to obtain the composite score, as shown:

Composite Score = (Listening × 0.25) + (Speaking × 0.25) + (Reading × 0.25) + (Writing × 0.25)

Using the sample scores from the chart above, the composite score is calculated as follows:

Composite Score = (3 × 0.25) + (2 × 0.25) + (3 × 0.25) + (2 × 0.25) = 2.5

TELPAS composite scores are converted to TELPAS composite ratings. In this example, the composite score of 2.5 results in a composite rating of Advanced due to the ratings profile having:

- a TELPAS composite score of 2.5 or higher,
- a minimum proficiency level of Intermediate in all domains, and
- a minimum proficiency level of Advanced in at least half of the domains in which the student was assessed.

A small subset of EB students with disabilities who cannot be assessed in all four domains will receive a composite score if they have results for at least two domains. This is applicable only to students who have a decision from the ARD committee, in conjunction with the LPAC, to not be evaluated in one or two domains. In such instances when not all four domains are assessed, the composite score will be calculated based on the number of domains assessed.

Figure 6.2 provides a student example to show how composite results are generated when one domain is not assessed.

Figure 6.2. Sample Calculation of Composite Results
One Domain Not Assessed

Domain	Proficiency Level	Domain Score
Listening	Intermediate	2
Speaking	Intermediate	2
Reading	Beginning	1
Writing	Not Assessed	Not Assessed

The domain scores are multiplied by the appropriate weight and then added together to obtain the composite score, as shown:

Composite Score = (Listening × $\frac{1}{3}$) + (Speaking × $\frac{1}{3}$) + (Reading × $\frac{1}{3}$)

Using the sample scores from the chart above, the composite score is calculated as follows:

Composite Score = (2 × $\frac{1}{3}$) + (2 × $\frac{1}{3}$) + (1 × $\frac{1}{3}$) = 1.7

TELPAS composite scores are converted to TELPAS composite ratings. In this example, the composite score of 1.7 results in a composite rating of Intermediate due to the ratings profile having:

- a TELPAS composite score of 1.5 or higher, and
- a minimum proficiency level of Intermediate in at least half of the domains in which the student was assessed.

Assessment Reports

TEA provides reports of student performance on TELPAS to all Texas public school districts and open-enrollment charter schools. For TELPAS, TEA provides student report cards, student labels, campus rosters, summary reports, and reporting data files. In addition, TEA periodically releases TELPAS online assessments to the public through the Practice Test Site.

For more information about scoring and reporting for TELPAS, refer to the Interpreting Results page of the *Coordinator Resources*.

Use of Test Results

Test results can be used to evaluate the performance of a group over time. Average scale scores and the percentage of students achieving each proficiency level can be analyzed by grade and domain across administrations to provide insight into whether student performance is improving across years. Test results can be used when evaluating instruction or programs that require average-score or year-to-year comparisons. The tests are designed to measure English language proficiency based on the ELPS, and so the consideration of test results by domain and reporting category might be helpful when evaluating curriculum and instructional programs.

All test scores can be compared with statewide and regional performance within the same domain for any administration.

TELPAS student performance reports are used to:

- help families monitor the progress their child is making in acquiring English;
- inform instructional planning for individual students;
- report results to local school boards, school professionals, and the community;
- evaluate programs, resources, and staffing patterns; and
- evaluate district effectiveness in accountability measures.

Audits

Since the 2004–2005 school year, TEA has conducted periodic audits of the TELPAS assessment processes as a means of collecting reliability and validity evidence for the assessment program. Audits allow for the collection of information from school districts that can be used to evaluate the training, administration, and scoring of the holistically rated assessments. Information collected during TELPAS audits has been useful in the refinement of TELPAS holistic rating training and administration procedures.

Last conducted in spring 2011, an audit process for the listening and speaking domains was used in which documentation was collected from teachers at selected sites to evaluate the accuracy of holistic ratings. Due to the replacement of holistically scored assessments with an online assessment, no further audits are needed for TELPAS listening and speaking.

In the TELPAS writing audit conducted in 2019, expert raters provided second ratings of student writing samples and testing personnel at the sampled sites completed questionnaires that allowed for a conformity evaluation of training and administration procedures. Due to the replacement of the holistically rated writing assessment with an online assessment in 2023, audits are no longer necessary for TELPAS writing.

Scaling

Scaling is a statistical procedure that places raw scores on a common scoring metric to make test scores comparable across test administrations. Scaling associates numbers with characteristics of interest to provide information about measurable quantities for those characteristics. TELPAS uses the RPCM to place test items on the same Rasch scale across administrations for a given assessment. Once performance standards have been set for an assessment, the Rasch scale is then transformed to a more user-friendly metric to facilitate interpretation of the test scores. Details of the RPCM scaling method are provided in Chapter 3, “Standard Technical Processes.”

Reporting Scales

TELPAS scale scores are reported on a horizontal scale. Horizontal scale scores allow for direct comparisons of student performance between specific sets of test items from different test administrations. Refer to Chapter 3, "Standard Technical Processes," for detailed information about the scaling process.

Scale for Online Assessments

The reporting scales for each domain (listening, speaking, reading, and writing) are independent horizontal scales with lowest obtainable scale scores of 1000 and highest obtainable scale scores of 2000. The cut scores on the reporting scale for the Advanced and Advanced High proficiency levels are 1525 and 1600, respectively, to create common points of reference across the assessments for each grade and domain. It is important to note that although the Advanced and Advanced High scale score values are fixed across horizontally scaled assessments, the Intermediate scale score values vary across assessments. For any given assessment, the proficiency standards remain constant over time.

TELPAS scale scores represent linear transformations of Rasch proficiency-level estimates (θ). Specifically, the transformation is made by first multiplying θ by a slope constant (*A*) and then adding an intercept constant (*B*). This operation is described by the following equation:

$$SS_\theta = A \times \theta + B,$$

where SS_θ is the scale score for a Rasch proficiency score estimate (θ) and *A* and *B* are referred to as the horizontal scaling constants. These same transformations are applied each year to the Rasch proficiency level estimates (θ) for that year's set of test items. Values for the horizontal scaling constants are provided in Table 6.5.

Table 6.5. Horizontal Scaling Constants for TELPAS

Domain	Grade or Grade Band	*A*	*B*
Listening	Grades 2–3	67.4946	1497.4015
	Grades 4–5	64.5661	1482.9804
	Grades 6–8	67.6285	1486.0798
	Grades 9–12	53.7172	1497.3517
Speaking	Grades 2–3	35.0533	1511.4519
	Grades 4–5	24.6208	1522.0652
	Grades 6–8	19.5008	1530.4446
	Grades 9–12	21.0574	1545.1456

Domain	Grade or Grade Band	*A*	*B*
Reading	Grade 2	66.7438	1423.0422
	Grade 3	88.0488	1396.6160
	Grades 4–5	86.5951	1391.3838
	Grades 6–7	79.5756	1380.2599
	Grades 8–9	68.8452	1408.3486
	Grades 10–12	64.4607	1389.4972
Writing	Grade 2	37.5921	1452.6615
	Grade 3	41.4342	1450.0496
	Grades 4–5	57.2738	1484.0778
	Grades 6–7	58.9855	1504.2252
	Grades 8–9	58.3794	1480.8360
	Grades 10–12	68.8389	1508.4649

Scale for Holistically Rated Assessments

The scale for TELPAS holistically rated assessments (all domains for kindergarten and grade 1) ranges from 1 to 4 and is defined by the four proficiency levels: Beginning, Intermediate, Advanced, and Advanced High.

Scale for Composite Ratings

TELPAS composite ratings use a scale from 1 to 4.

Equating

Used in conjunction with the scaling process, equating is the statistical process that considers the differences in difficulty across test forms and administrations and allows scores to be placed onto a common scale. By using statistical methods, TEA equates the results of different test forms so that scale scores across test forms and testing administrations can be compared. TEA uses pre-equating, post-equating, and field-test equating for all online TELPAS assessments. During each administration, field-test equating is conducted for online TELPAS assessments through an embedded-field-test design.

Equating is not necessary for TELPAS holistically rated assessments. The difficulty level of these assessments is maintained using consistent rating rubrics developed to define the proficiency levels. Prior to test administration, raters complete training activities that provide consistency in the way the rubrics are applied. By calibrating the raters to the assessment rubric, the training maintains the difficulty of the assessment across administrations.

Refer to Chapter 3, “Standard Technical Processes,” for detailed information about equating.

Reliability

Reliability indicates the precision of test scores, referring to the expectation that repeated administrations of the same test should generate consistent results. Reliability for TELPAS test scores is estimated using statistical measures including internal consistency, classical SEM, CSEM, and classification consistency and accuracy. Data for each of these statistical measures from the spring TELPAS administration is provided in Appendix D.

In addition to the statistical measures mentioned above, TEA also collects inter-rater reliability evidence, median response times for speaking, and composite score reliability estimates.

Inter-Rater Reliability

Evidence that the holistically rated components of TELPAS result in reliable observation and rating of student performance is collected through periodic inter-rater reliability studies. Evidence of inter-rater reliability is collected through the audit process by having a second rater provide independent ratings for a sample of students.

For the TELPAS grades 2–12 writing audit conducted in 2019, districts were required to submit writing collections for approximately 2,200 EB students selected for the pure random sample, which was spread across grade levels and stratified across proficiency levels. The writing collections included writing from classroom instruction in a variety of core content areas and were rescored by Pearson scorers after the original scores were collected from the TELPAS raters. Audit results were documented in the TELPAS Writing Audit Report, available on the Assessment Reports and Studies webpage, and add to the body of validity and reliability evidence collected to support the assessment system. This process enables the evaluation of classroom activities on which the assessments are based and the way raters statewide interpret the PLD rubrics. The same information collected during TELPAS audits has been useful in the refinement of TELPAS rater training and administration procedures.

For TELPAS speaking, field-test items are examined for human-to-human and human-to-machine agreement. Evidence of inter-rater reliability is gathered by examining the perfect agreement rates and the Pearson correlations. An additional validity check is performed on the automated scoring of the responses to check inter-rater reliability between automated and human scoring. A random sample of 15 percent of students per grade band is selected for human scoring. The grade band correlations between the total raw scores on the human-scored and automated scored samples are presented in Table 6.6.

Table 6.6. TELPAS Speaking Inter-Rater Correlations

Grade	*N*	Inter-Rater Correlation
Grades 2–3	29,925	0.85
Grades 4–5	30,973	0.82
Grades 6–8	44,052	0.86
Grades 9–12	41,491	0.90

Median Response Time

When students are ready to respond to a speaking prompt, they use a speech capture tool in the online testing interface to record their responses. They have 45 seconds to respond to simple prompts and 90 seconds to respond to more complex prompts. Students are allowed two recording attempts per item; they may listen to their first recorded response and, if desired, delete it and record a second response.

Analysis was conducted on student response time to speaking items and the relationship to the overall student proficiency level on the speaking domain. This information is useful to educators and students to help demonstrate how the time spent responding impacts student performance.

Table 6.7 shows the median response time per item (for both 45-second and 90-second responses), by proficiency level, for a random sample of 5,000 students per grade band from this year's administration.

Table 6.7. TELPAS Speaking
Median Response Time per Item

Proficiency Level	Time per Item in Seconds			
	Grades 2–3	Grades 4–5	Grades 6–8	Grades 9–12
Beginning	0.0	0.0	0.0	0.0
Intermediate	13.4	11.4	8.2	8.5
Advanced	29.6	23.2	23.9	18.2
Advanced High	44.2	38.6	42.2	33.3

Composite Score Reliability Estimates

TELPAS composite score reliability estimates are analyzed annually to evaluate the impact of the reliability of the listening, speaking, reading, and writing domains on the TELPAS composite reliability estimates. The composite score reliability estimates are calculated using a stratified alpha approach. This approach is described by the equation below:

$$\alpha_{Strat} = 1 - \frac{\sum_{i=1}^{k} w_i^2 \sigma_{X_i}^2 \left(1 - \rho_{X_i X_i'}\right)}{\sigma_Z^2} ,$$

where k is the number of the components or domains, w_i is the weight of each domain, X_i represents the domain score of each domain, $\rho_{X_i X_i'}$ is the internal consistency of each domain, and z is the composite score. The internal consistency values of listening, speaking, reading, and writing on the categorical scale were estimated based on their internal consistency values on the continuous scale. The results of these analyses, presented in Table 6.8, show that the weighted TELPAS composite scores have reliability estimates of at least 0.913.

Table 6.8. TELPAS Composite Score Reliability Estimates

Grade	Domain	Mean	Standard Deviation	Internal Consistency	Composite Reliability
Grade 2 (*n*=97926)	Listening	2.834	1.001	0.786	0.913
	Speaking	1.962	0.746	0.814	
	Writing	1.882	0.822	0.783	
	Reading	1.894	0.860	0.756	
Grade 3 (*n*=99466)	Listening	3.363	0.897	0.789	0.929
	Speaking	2.248	0.812	0.821	
	Writing	2.080	0.879	0.814	
	Reading	2.450	1.120	0.837	
Grade 4 (*n*=100728)	Listening	2.685	1.012	0.811	0.929
	Speaking	2.442	0.891	0.825	
	Writing	2.180	0.834	0.775	
	Reading	2.728	1.054	0.822	
Grade 5 (*n*=102472)	Listening	2.987	0.995	0.809	0.929
	Speaking	2.514	0.914	0.831	
	Writing	2.467	0.853	0.767	
	Reading	3.072	1.015	0.823	
Grade 6 (*n*=99576)	Listening	2.928	0.891	0.735	0.913
	Speaking	2.389	0.788	0.820	
	Writing	2.389	0.838	0.759	
	Reading	2.662	1.017	0.793	
Grade 7 (*n*=97447)	Listening	3.013	0.908	0.743	0.917
	Speaking	2.319	0.817	0.831	
	Writing	2.519	0.858	0.762	
	Reading	2.799	1.026	0.803	
Grade 8 (*n*=94274)	Listening	3.128	0.908	0.762	0.922
	Speaking	2.322	0.842	0.851	
	Writing	2.347	0.833	0.757	
	Reading	2.880	0.937	0.815	
Grade 9 (n=92661)	Listening	2.746	0.901	0.778	0.931
	Speaking	2.178	0.974	0.879	
	Writing	2.265	0.860	0.767	
	Reading	2.801	0.962	0.820	

Grade	Domain	Mean	Standard Deviation	Internal Consistency	Composite Reliability
Grade 10 (n=74724)	Listening	2.842	0.884	0.756	0.927
	Speaking	2.268	0.984	0.880	
	Writing	2.408	0.914	0.763	
	Reading	2.642	0.935	0.814	
Grade 11 (n=53384)	Listening	2.917	0.865	0.762	0.924
	Speaking	2.329	1.000	0.880	
	Writing	2.500	0.899	0.750	
	Reading	2.729	0.929	0.808	
Grade 12 (n=44100)	Listening	2.879	0.846	0.749	0.918
	Speaking	2.260	0.995	0.883	
	Writing	2.466	0.886	0.741	
	Reading	2.700	0.901	0.796	

Refer to Chapter 3, “Standard Technical Processes,” for detailed information about reliability.

Validity

Validity refers to the extent to which test scores accurately measure what the test is intended to measure. TEA follows national standards of best practice and annually collects validity evidence to support the interpretations and uses of TELPAS results. TTAC, a panel of national testing experts created specifically for the Texas Assessment Program, provides ongoing input to TEA about TELPAS validity evidence. The following sections describe how validity evidence has been collected for TELPAS. Refer to Chapter 3, “Standard Technical Processes,” for additional information about validity.

Evidence Based on Test Content

Validity evidence based on test content refers to evidence of the relationship between tested content and the construct that the assessment is intended to measure. Content validity evidence is collected at all stages of the test-development process. Nationally established test-development processes for the Texas Assessment Program are followed while developing TELPAS. This supports the use of TELPAS results in making inferences about students’ English language proficiency. TELPAS measures student performance in direct alignment with the English language acquisition skills and PLDs defined by the Texas ELPS that are part of the TEKS curriculum. The ELPS outline the instruction that EB students must receive to support their ability to develop academic English language proficiency.

Online Assessments

TELPAS online assessments are designed to assess English language proficiency in a manner that provides information about how well EB students understand and produce the English they need for academic success in Texas schools, as well as the types of language supports they require to comprehend written or spoken English independently.

As part of the development of TELPAS online assessments, teachers, curriculum specialists, test development specialists, and TEA staff worked together in advisory committees to identify appropriate assessment reporting categories. The input of the advisory committees was reflected in the assessed curricula and test blueprints. In addition, prototype items were developed for the assessments early in the development process. The educator advisory committees reviewed these prototypes to identify how well the items would measure the student expectations to which the items were aligned. These early reviews provided valuable suggestions for item development guidelines and item types. Item development guidelines continued to be refined through the test development process as various TELPAS item-review educator committees shared their feedback about how the student expectations in the ELPS could be effectively assessed.

As part of the annual process of item development, committees of Texas educators meet to review TELPAS items and confirm that each item appropriately measures the ELPS to which it is aligned. These committees also review items for bias. Item review committees are composed of Texas educators, and these committees revise and edit items, as appropriate, prior to field testing. Item review committees are convened for all TELPAS online assessments.

Item writers and reviewers follow test development guidelines that explain how content aligned to given ELPS should be measured. At each stage of development, writers and reviewers verify the alignment of the items with the assessed student expectations.

TELPAS online assessments are built using four levels of built-in linguistic support addressing the gradually reduced degree of linguistic accommodation that EB students need as they progress from knowing little or no English to becoming fluent in English. The levels of linguistic support are integrally related to the four proficiency levels assessed, as each proficiency level described in the ELPS is characterized by the degree of linguistic accommodation that students at that level need to understand and speak English. The staged linguistic accommodation test design is shown in Table 6.9.

Table 6.9. Staged Linguistic Accommodation Test Design

Proficiency Level	Degree of Linguistic Accommodation Applied to Stimulus and Item Development	
Beginning	Extensive	• maximum picture support • short stimuli that require comprehension of words, phrases, and short sentences that use the type of high-frequency, concrete vocabulary first acquired by learners of a second language
Intermediate	Substantial	• frequent picture support • short stimuli written primarily on familiar topics • commonly used everyday English and routine academic English
Advanced	Moderate	• occasional picture support • contextual aids and organizational features support comprehension of longer stimuli on both familiar and unfamiliar social and content-area topics
Advanced High	Minimal	• minimal linguistic accommodation • stimuli highly comparable to those intended for native English speakers

This test design supports the validity of TELPAS online assessments in that it provides built-in, staged linguistic accommodations validated by second language acquisition theory and empirical data as it measures skills in the ELPS that students need for academic success in all content areas.

Holistic Assessments

TELPAS holistically rated assessments are aligned with the ELPS and are designed to assess the English communication skills that EB students need to engage meaningfully and successfully in learning the TEKS. They draw on second language acquisition research, research-based standards, the experience of Texas educators, and observational assessment practices.

The TELPAS holistically rated components are based on ongoing observations of the ability of EB students to understand and use English during the grade-level content-area instruction required by the state-mandated curriculum and assessed by STAAR. TELPAS holistically rated assessments measure the ELPS student expectations from the cross-curricular second language acquisition knowledge and skills and use the ELPS PLDs as assessment rubrics. Rater training and administration procedures require these ratings to be based on the ability of students to use English in a variety of content areas.

Evidence Based on Response Processes

Examining students' response processes provides an additional source of validity evidence.

Online Assessments

Student response processes on TELPAS online assessments vary per item type. Across TELPAS, a variety of question types (e.g., multiple-choice, fill in the blank, drag and drop, hot spots) and response interactions are available to measure second language acquisition.

TEA gathers theoretical and empirical evidence to confirm that the type of response required for each item does not add construct-irrelevant variance. TEA also gathers evidence from several sources to confirm that response processes do not result in an advantage or disadvantage for any student group. When new item types or changes to the format of existing item types are considered for TELPAS, cognitive labs are used to study the way students engage with the various item presentations. After item types are determined to be appropriate for TELPAS, evidence about student responses is gathered annually through educator and expert reviews and analyses of individual student responses to these items. During item reviews, educators evaluate whether the content for a given item type is being appropriately assessed and whether students will be able to accurately demonstrate their knowledge of the construct given the items' planned format. When items are field-tested, additional data are gathered about students' responses. Data such as item difficulty, item point-biserial correlations, and DIF are all evaluated regarding the item type. For additional information, refer to the Item Analysis section of Chapter 3, "Standard Technical Processes."

The process used to score items can provide additional validity evidence based on response processes. This type of validity evidence is predicated on accurate scoring. For all multiple-choice and multiselect items on TELPAS, statistical key checks are conducted during the equating process. A statistical key check is a procedure in which the statistical properties of all items on every test form are computed. Items whose statistics do not meet predetermined criteria are flagged for further review by content experts to verify that the items are correctly keyed and scored. An adjudication process is used to ensure scoring reliability and validity for technology enhanced items. During adjudication, data files that include all unique responses for each test question are analyzed to identify responses or questions that require more detailed analysis to ensure accurate, consistent scoring. Evaluators who specialize in English language proficiency then review student responses to resolve scoring discrepancies or uncertainties.

For constructed-response questions, rubrics are used to evaluate student responses. All rubrics for TELPAS are validated by educator committees and content experts. In addition, TEA has implemented a rigorous scoring process for constructed responses that includes training and qualification requirements for scorers, ongoing monitoring during scoring, adjudication and resolution processes for student responses that do not meet the exact or adjacent scoring requirements, and rescoring of responses for which concerns have been raised by districts, campuses, or teachers regarding the assigned score. A more comprehensive description of the scoring process for constructed-response items is available in Chapter 2, "Building a High-Quality Assessment System."

Validity is evaluated through validity papers, which are student responses from the field test and current administrations that are representative of different levels of writing performance based on the scoring rubrics. Validity papers are identified by scoring leaders and are then

systematically given to scorers throughout the scoring project. An important feature of validity papers is that they are not identifiable as such; in fact, they are indistinguishable from unscored student responses. Each person's daily scores on validity papers are compared with the approved scores. Validity papers are used throughout the scoring project as a primary quality-control measure, the purpose of which is to ensure that scorers are accurately and reliably scoring on a daily basis and across time.

Holistic Assessments

TELPAS holistically rated assessments are based on ongoing classroom observations and daily interaction with students. As is typical of holistically scored assessments, students are evaluated on their overall performance in a global and direct way. TELPAS holistically rated assessments meet the goal of English language proficiency assessments to effectively assess the extent to which EB students are making progress in attaining academic language proficiency by serving as direct measures of the ability of students to understand and use English while engaging in state-required academic instruction. As such, the assessments provide strong validity evidence related to the response process.

Evidence Based on Internal Structure

TEA collects evidence that reflects the relationship between item performance and proficiency levels to verify that patterns of item performance are consistent with the constructs the assessment is intended to measure.

Online Assessments

Internal consistency reliability estimates provide a measure of the consistency with which students respond to the items in an assessment and show the relationship of students' responses between items, within reporting categories of items, and within domains to verify that the elements of an assessment conform to the intended test construct. The internal consistency of TELPAS online assessments is evaluated using KR20 for assessments that have only dichotomously scored items. For TELPAS online assessments that have a combination of dichotomous and polytomous items, internal consistency is evaluated using coefficient alpha and stratified alpha. These internal consistency evaluations are made for all students and for female and male student groups. Estimates of internal consistency can be found in Appendix D.

Holistic Assessments

Evidence of the validity of TELPAS holistic assessments is supported by comprehensive training and administration procedures that prepare raters to perform their duties and district administrators to follow procedures to maintain the integrity of the test administration. In addition to holistic rating training, raters must perform calibration activities to demonstrate high accuracy in rating student activities across all TELPAS holistically rated domains they will assess. Additional support is provided to raters who cannot calibrate on their first two attempts in order to help them assess assigned students in a manner consistent with the PLDs.

TELPAS holistic rating audits provide both validity and reliability evidence based on the internal structure by examining the extent to which raters follow the defined protocol for rating these components. As part of the audit, reports of rater adherence to the assessment protocol are made and used to provide evidence that the internal structure of the assessment is intact and that educators are administering the assessment and applying the scoring rubrics appropriately.

In addition to directly supporting the state's goal of having a valid and authentic assessment, TELPAS holistically rated assessments also serve an ongoing critical role as a professional development tool that supports effective instruction, enabling teachers to better understand and meet the educational needs of EB students.

Evidence Based on Consequences of Testing

Another method for providing validity evidence is by documenting the intended and unintended consequences of administering an assessment. The collection of consequential validity evidence typically occurs after a program has been in place for some time and on a regular basis.

Given the important stakes associated with TELPAS, valid test scores are critical in supporting their intended interpretations and uses. The intended interpretations of TELPAS results are stated in the policy definitions of the four proficiency levels. Refer to the Proficiency Standards section for the policy definitions of the TELPAS proficiency levels. The ELPS PLDs describe a student's English language acquisition skills in each domain based on the student's proficiency level.

Student-Level Performance

The following are the intended uses of TELPAS results based on the policy definitions for student-level performance:

- Proficiency on TELPAS is one indicator of a student's level of proficiency in learning English.
- Proficiency on TELPAS is one indicator of a student's possible need for academic intervention.
- Proficiency on TELPAS across years provides one indicator of a student's English language acquisition within a domain.

District- or Campus-Level Performance

The following are the intended uses of TELPAS test results based on the policy definitions for district- or campus-level performance:

- TELPAS results provide an indicator of overall student English language proficiency at a district or campus.
- TELPAS results can be aggregated across years to provide an indicator of overall student progress in English language acquisition at a district or campus.

Evidence based on the consequences of testing can also be found by comparing performance from past administrations, which is represented in Appendix D. The proficiency-level classifications of students for the listening, speaking, reading, and writing domains of TELPAS have been continually collected since the first administration. In general, long-term trends show a gradual increase in student performance after the introduction of TELPAS, and such improvement may have resulted, in part, from the use of test data to inform instruction.

While TELPAS has continued to assess the same ELPS, changes to the assessment design over time make comparisons to earlier results difficult to interpret. Comparisons in performance are only appropriate across certain years. For example, TELPAS writing results for all grades can be compared from 2005 until 2022. For grades 2–12 listening and speaking, results can be compared within the periods of 2005–2017 and 2018–present. For reading, results are comparable within the periods of 2005–2013, 2014–2017, and 2018–present. However, direct comparisons across these distinct periods are not appropriate. If historical trends hold, however, over time the percentages of students across proficiency levels are expected to remain relatively stable, with the possibility of a gradual increase in performance.

In addition to district and campus consequences, based on what educators learn during rater training and from the observation process, the administration of TELPAS holistically rated assessments leads to improvements in students' language acquisition for both formative and summative purposes. For example, educators learn how developing academic language proficiency in English relates to and supports academic achievement in English.

Sampling

Sampling is a procedure that is used to select and examine a small set that is representative of the population from which it was drawn. For TELPAS, sampling occurs when observed *n*-counts for handscored field-test items exceed 3,000.

Test Results

Appendix D provides consistency and accuracy data, scale score correlations, CSEMs, mean *p*-values, scale score descriptive statistics, frequency distributions, and proficiency level distributions for the spring TELPAS administration. The percentages of students in each proficiency level for all four domains as well as for the composite rating are available on the Statewide Summary Reports webpage.

TECHNICAL DIGEST 2022–2023

Chapter 8

Resources

Resources

Information about the Texas Assessment Program can be found on the TEA Student Assessment webpages and the Texas Assessment website. A summary of some available resources is provided in Table 8.1.

Table 8.1 Texas Assessment Program Online Resources

Topic	URL
Texas Assessment Systems	http://texasassessment.gov/testing-personnel.html
District and Campus Coordinator Resources	https://txassessmentdocs.atlassian.net/wiki/spaces/ODCCM/overview
STAAR Resources	https://tea.texas.gov/student-assessment/testing/staar/staar-resources
STAAR Alternate 2 Resources	https://tea.texas.gov/student-assessment/testing/staar-alternate/staar-alternate-2-resources
TELPAS Resources	https://tea.texas.gov/student-assessment/testing/telpas/telpas-resources
TELPAS Alternate Resources	https://tea.texas.gov/student-assessment/testing/telpas/telpas-alternate-resources
Assessments for Special Populations	https://tea.texas.gov/student-assessment/testing/student-assessment-overview/assessments-for-special-populations
Accommodation Resources	https://tea.texas.gov/student-assessment/testing/student-assessment-overview/accommodation-resources
Test Administration Resources	https://tea.texas.gov/student-assessment/testing/student-assessment-overview/test-administration-resources
Student Assessment Results	https://tea.texas.gov/student-assessment/testing/student-assessment-results
Assessment Reports and Studies	https://tea.texas.gov/student-assessment/testing/student-assessment-overview/assessment-reports-and-studies
Assessment Resources for Educators	https://www.texasassessment.gov/educators.html
Assessment Resources for Students and Families	https://www.texasassessment.gov/families.html