Chapter 2 Building a High-Quality Assessment System

Test-Development Activities

Texas educators, including K–12 classroom teachers, higher education representatives, curriculum specialists, administrators, and Education Service Center (ESC) staff, play a vital role in all phases 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. These committees represent the state geographically, ethnically, by gender, and by type and size of school district. The procedures described in Figure 2.1 outline the process used to develop a framework for the tests and provide for ongoing development of test items.

Figure 2.1. Test-Development Process

1 Committees of Texas educators review the state-mandated curriculum, the Texas Essential Knowledge and Skills (TEKS), or the English Language Proficiency Standards (ELPS) to develop appropriate assessment categories for a specific grade/subject or course that is assessed. For each grade/subject or course, educators provide advice on an assessment model or structure that aligns with best practices in classroom instruction.

2 Educator committees work with the Texas Education Agency (TEA) both to prepare draft test reporting categories and to determine how these categories would best be assessed. These preliminary recommendations are reviewed by K–12 teachers, higher education representatives, curriculum specialists, and assessment specialists.

3 A draft of the reporting categories and TEKS student expectations or ELPS 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 piloted by 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, many of whom are former or current Texas educators, develop test items based on the reporting categories, the TEKS student expectations or ELPS, 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 committee meetings and are 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 bank of items from which test forms are built.

13 A final blueprint for each test that establishes the number of questions on the test and the number of test questions measuring each reporting category is developed.

***14** All accepted field-test items and data are entered into a computerized item bank. Tests are built from the item bank so that the tests are comparable in difficulty from one administration to the next.

***15** Content validation panels composed of university-level experts in each content area review the end-of-course assessments or high-school level tests for accuracy because of the advanced level of content being assessed.

***16** Tests are administered to Texas students.

***17** Stringent quality control measures are applied to all stages of printing, scanning, scoring, and reporting for both paper and online assessments. Results of the test 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.

*These steps are repeated annually to ensure that tests of the highest quality are developed.

Groups Involved

Several groups are involved in the Texas assessment program. Each of the following groups performs specific functions, and their collaborative efforts significantly contribute to the quality of the assessment program.

Student Assessment Division

TEA's Student Assessment Division is responsible for implementing the provisions of state and federal law for the state assessment program. The Student Assessment Division oversees the planning, scheduling, and implementation of all major assessment activities and supervises the agency's contracts with Educational Testing Service (ETS) and Pearson. TEA staff members in this division conduct quality-control activities for the development and administration of the assessment program, as well as the monitoring of the program's security provisions.

Curriculum Standards and Student Support Division

TEA's Curriculum Standards and Student Support Division is responsible for supporting development and implementation of the TEKS in the foundation curriculum (mathematics, reading language arts, writing, science, and social studies) and the enrichment curriculum (fine arts, health education, languages other than English, physical education, and technology applications). TEA staff members in this division provide content expertise during item development and test build processes for all statewide assessments.

Performance Reporting Division

TEA's Performance Reporting Division is responsible for compiling and analyzing data to develop and report meaningful accountability ratings that help Texas public schools meet the educational needs of all students. As part of administering the state's public school accountability system, the department publishes assessment reporting and accountability information. TEA staff members in this division conduct quality-control activities for the scoring and reporting of the assessment program.

The department also provides guidance and resources to help school administrators, teachers, parents, and the general public understand and benefit from the state's accountability information.

ETS

ETS is the contractor for the provision of support services to the state for the State of Texas Assessments of Academic (STAAR®) program, which includes STAAR Spanish. ETS also serves as the program integration contractor. This role includes working with Pearson to make sure that the state assessment program as a whole is managed per TEA requirements. Due to the diverse nature of the services required, ETS employs subcontractors to perform tasks requiring specialized expertise. During the 2018–2019

school year, ETS's subcontractor to develop STAAR Spanish assessments was Tri-Lin Integrated Services, Inc. (Tri-Lin).

Tri-Lin

Tri-Lin Integrated Services, Inc., specializes in translation and transadaptation of test items from English into Spanish. As a subcontractor of ETS, Tri-Lin researches terminology and cultural and regional differences to generate the proper translations of the grades 3–5 mathematics and science items. In addition to the transadaptations of selected items, Tri-Lin works with ETS personnel, TEA staff members, and Texas educators to develop unique passages and items for the STAAR reading and writing assessments in Spanish.

Pearson

Pearson is TEA's contractor for the provision of support services to the state assessment program for STAAR Alternate 2, the Texas English Language Proficiency Assessment System (TELPAS), and TELPAS Alternate. Due to the diverse nature of the services required, Pearson employs subcontractors to perform tasks requiring specialized expertise. During the 2018–2019 school year, Pearson's subcontractor for test development activities was Lone Star Assessment and Publishing, L.L.C.

Lone Star Assessment and Publishing, L.L.C.

Lone Star Assessment and Publishing, L.L.C., provides management and clerical services to coordinate the hiring and payment of item writers for TELPAS and STAAR Alternate 2. As a subcontractor to Pearson, Lone Star Assessment and Publishing employs item writers who have extensive experience developing items for Texas and/or experience writing items for alternate assessments and English language proficiency tests.

Texas Educators

When a new assessment is developed, committees of Texas educators review the state-required curriculum, help develop appropriate reporting categories for the specific grade/subject or course tested, and provide advice on a model for assessing the content that aligns with the curriculum and instruction.

Draft reporting categories with corresponding TEKS or ELPS student expectations are reviewed by teachers, curriculum specialists, assessment specialists, and administrators. Texas educator committees assist in developing draft guidelines that outline the eligible test content and test item formats. TEA refines and clarifies these draft reporting categories and guidelines based on input from Texas educators.

Following the development of test items by professional item writers, many of whom are current or former Texas teachers, committees of Texas educators review the items to ensure appropriate content and level of difficulty and to eliminate potential bias. Items are revised based on input from these committees, and then the items are field-tested.

Item-Development and Review

This section describes the item-writing process used during the development of Texas assessment program items. While ETS and Tri-Lin, the subcontractor for the Spanish language assessments, assume the major role for STAAR item development, and Pearson assumes the major role for STAAR Alternate 2, TELPAS, and TELPAS Alternate item development, agency personnel are involved throughout the item-development process.

Item Guidelines

Item and performance task specifications provide guidance from TEA on how to translate the TEKS or ELPS into actual test items. Item guidelines are strictly followed by item writers in order to enable 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 and fairness across the development of test items.

Item Writers

ETS and Pearson each employ item writers who have extensive experience developing items for standardized achievement tests, large-scale criterion-referenced measurements, and English language proficiency tests. These individuals are selected for their background in specific content-area knowledge and their teaching or curriculum development experience in the relevant grades or second-language acquisition.

For each STAAR assessment, TEA receives an item inventory that displays the number of test items submitted for each reporting category and TEKS student expectation. Item inventories are examined throughout the review process. If necessary, additional items are written by ETS to provide the requisite number of items per reporting category.

For each STAAR Alternate 2 and TELPAS assessment, TEA receives an item inventory that displays the number of test items submitted for each reporting category and TEKS student expectation or ELPS. Item inventories are examined throughout the review process. If necessary, additional items are written by Pearson or its subcontractors to provide the requisite number of items per reporting category.

For the TELPAS Alternate Assessment, the observable behaviors were developed by Texas educators during a series of TEA led meetings. The educators were guided by Pearson and TEA staff to develop an inventory of items that aligned to the ELPS and covered the alternate Proficiency Level Descriptors (PLDs).

Training

ETS and Pearson each provide extensive training for item writers prior to item development. During these trainings, ETS and Pearson review in detail the content expectations and item guidelines and discuss 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

Experienced staff members from ETS and Pearson, who are content experts in the grades and content areas for which items are developed, participate in the review of each set of newly developed items. This review includes a check for content accuracy and fairness of the items for different demographic groups. ETS and Pearson reviewers consider additional issues, such as 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"). Reviewers also consider the more global issues of universal design; passage appropriateness; passage difficulty; readability measures; interactions among items within passages and between passages; and appropriateness of artwork, graphics, or figures. The items are examined by ETS and Pearson editorial staff before they are submitted to TEA for review.

TEA Review

TEA staff members from the Curriculum Standards and Student Support and Student Assessment Divisions, who are content experts in the grades and content areas for which items are developed, review each item to verify alignment to a particular student expectation in the TEKS/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. Then, provide edits electronically, and meet with ETS and Pearson to discuss the progress of the reviews before each educator item review committee meeting. For STAAR Alternate 2, staff from TEA meet with Pearson to examine, discuss, and edit all newly developed items before each educator item review committee meeting.

Item Review Committee

Each year, TEA's Curriculum Standards and Student Support and Student Assessment Divisions convenes committees composed of Texas classroom teachers (including general education teachers, special education teachers, and bilingual and English as a second language [ESL] teachers), curriculum specialists, administrators, and regional ESC staff 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 Form database where educators can self-nominate to participate in all TEA Educator Committees. Item review committee members are selected based on their established expertise in a content area. Committee members represent the 20 ESC regions of Texas and the major ethnic groups in the state, as well as the various types of districts (e.g., urban, suburban, rural, large, and small districts).

TEA staff, along with ETS, Tri-Lin, and Pearson staff, train committee members on the proper procedures and the 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 the item should be field-tested as written or revised, recoded to a different TEKS or ELPS student expectation, or rejected. All committee members conduct their reviews considering the effect on various student populations and work toward eliminating potential bias against any group. Table 2.1 shows the guidelines item-review committee members follow in their review.

Table 2.1. Item-Review Guidelines

Item-Review Guidelines	
Reporting Category/Student Expectation Item Match	• Does the item measure what it is supposed to assess? • Does the item pose a clearly defined problem or task?
Appropriateness (Interest Level)	• Is the item or passage well written and clear? • Is the point of view relevant to students taking the test? • Is the subject matter of fairly wide interest to students at the grade being tested? • Is artwork clear, correct, and appropriate?
Appropriateness (Format)	• Is the format appropriate for the intended grade? • Is the format sufficiently simple and interesting for the student? • Is the item formatted so it is not unnecessarily difficult?
Appropriateness (Answer Choices)	• Are the answer choices reasonably parallel in structure? • Are the answer choices worded clearly and concisely? • Do any of the choices eliminate each other? • Is there only one correct answer?
Appropriateness (Difficulty of Distractors)	• Is the distractor plausible? • Is there a rationale for each distractor? • Is each distractor relevant to the knowledge and understanding being measured? • Is each distractor at a difficulty level appropriate for both the objective and the intended grade?
Opportunity to Learn	• Is the item a good measure of the curriculum? • Is the item suitable for the grade or course?
Freedom from Bias	• Does the item or passage assume racial, class, or gender values or suggest such stereotypes? • Does the item provide an advantage or disadvantage to any group of students because of their personal characteristics, such as race, gender, socioeconomic status or religion? • Might the item or passage offend any population? • Are minority interests well-represented in the subject matter and artwork?

If the committee finds an item to be inappropriate after review and revision, it is removed from consideration for field testing. TEA field-tests the recommended items to collect student responses from representative samples of students from across the state.

TELPAS Alternate did not have educator review committees like the other assessments. Instead, TELPAS Alternate observable behaviors were written and revised by the educators during the development of the assessment.

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 pilot items of differing types and ranges of difficulty, piloting might occur before the extensive item-development process described on the preceding pages. If the purpose is to pilot test administration logistics, the pilot might occur after major item development but before field testing. No pilot testing was conducted for the Texas assessment program during the 2018-2019 school year.

Field Testing and Data Review

Field testing is conducted prior to a test item being used on an operational test form. However, when there are curriculum changes, newly developed items that have not been field-tested may be used on an operational test form. This is referred to as operational field testing.

Field-Test Procedures

Whenever possible, TEA conducts field-tests of new items by embedding them in multiple forms of operational tests so that the field-test items are randomly distributed to students across the state. This results in a large representative sample of responses gathered on each item. Periodically, TEA conducts standalone field tests of new items (e.g., writing prompts) by administering them to a purposefully selected representative Texas student sample.

For the spring 2019 STAAR grades 3–8 primary administrations, six field-test questions were embedded in each form for mathematics, reading, science, and social studies, and five were embedded in each Spanish grade 4 writing form. For the spring 2019 STAAR end-of-course (EOC) primary administrations, 13 field-test questions were embedded in each English I and English II form, and eight were embedded in each Algebra I, Biology, and U.S. History form.

Past 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 field-test items in a realistic testing situation. Performance on field-test items is not part of the students' scores on the operational tests.

Additionally, through standalone field testing in February 2019, ten writing prompts were evaluated in grade 4 writing, Spanish grade 4 writing, and grade 7 writing assessments; 19 were evaluated in English I; and 16 were evaluated in English II. The standalone field testing followed the same administration procedure as the primary administration. The student participants were carefully sampled to represent the key academic and demographic characteristics. Regular standalone field testing of writing prompts is needed to ensure that sufficient writing prompts are developed to replenish the item banks without requiring students to respond to more than one prompt during their primary administrations. Past experience has shown that the standalone field testing yields sufficient data for precise item evaluation.

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;
- the percentage of students, by gender and by ethnicity, choosing each response;
- point-biserial correlations to determine the relationship between a correct response on a particular test item and the score obtained on the total content-area test;
- 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 (CR) items to identify greater-than-expected differences in group performance on any single item by gender and ethnicity.

Data-Review Procedures

After field testing, TEA curriculum and assessment specialists provide feedback to ETS and Pearson on each test item and its associated data regarding reporting category/student expectation match; appropriateness; level of difficulty; and potential gender, ethnic, or other bias; and 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

ETS and Pearson each maintain an electronic item bank for their respective portion of the assessment program. The item banks store each test item and its accompanying artwork.

Each electronic item bank also stores item data, such as the unique item number (UIN), grade or course, subject, reporting category, TEKS/ELPS student expectation measured, dates the item was administered, and item statistics. Each item bank also warehouses information obtained during data review meetings, which specifies whether a test item is acceptable for use. TEA, ETS, and Pearson uses the item statistics and other information about items during the test construction process to maintain constant test difficulty and adjust the test for content coverage and balance.

Test Construction

Each content-area and grade-level assessment is based on a specific test blueprint that guides how each test is constructed. Test blueprints delineate the number of items from each reporting category that will appear on a given test. Additionally, the STAAR, STAAR Spanish, and STAAR Alternate 2 assessments focus on the TEKS that are most critical to assess by incorporating readiness and supporting standards into the test blueprints. Readiness standards are emphasized annually in the STAAR, STAAR Spanish, and STAAR Alternate 2 assessments. Supporting standards are an important part of instruction and are eligible for assessment, but they may not be tested each year. All decisions about the relative emphasis of each reporting category were based on feedback from Texas educators (from both K–12 and higher education) and are indicated in the Test Blueprints and Assessed Curriculum documents on TEA's website. General characteristics of readiness and supporting standards are shown in Table 2.2.

Table 2.2. 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

TELPAS and TELPAS Alternate are based on all the ELPS, and the assessable ELPS for the Alternate population, respectively. The ELPS do not designate between as readiness or supporting.

Overall, each assessment is designed to reflect

- problem solving and complex thinking skills;
- the range of content (including readiness and supporting standards) 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 proficiency level descriptors 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 each item, in terms of difficulty, to 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 administration to administration. 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 on the overall test and for each reporting category. In addition, blueprints for STAAR, STAAR Spanish, and STAAR Alternate 2 include a specific number of readiness and supporting standards. Items that test each reporting category are included for every administration, but the array of TEKS/ELPS student expectations represented might vary from one administration to the next. Although the STAAR, STAAR Spanish, and STAAR Alternate 2 tests 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.

At the end of test construction for STAAR EOC assessments, panels composed of university-level experts in the fields of mathematics, reading/language arts, science, and social studies review the content of each STAAR EOC assessment before test construction is completed. This review is referred to as content validation and is included as a quality-control step to ensure that each high school 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. After a thorough review of each assessment, committee members note any issues of concern. When necessary, replacement items are chosen and reviewed. There is no content validation review for TELPAS and TELPAS Alternate.

After test construction for STAAR is complete, ETS and TEA work together to develop content and language supports for students who meet eligibility criteria. These embedded accommodations, or designated supports, are available for all online STAAR test forms. Embedded accommodations are not provided on TELPAS or TELPAS Alternate assessments. 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 selections, test questions, and answer choices and are mainly in the form of pop-ups, rollovers, prereading text, and supplementary materials (these supports are not available for Algebra II or English III). All test content, including the embedded supports, is reviewed and approved by TEA. The assessments are then ready to be administered.

The TELPAS Alternate assessment was designed to be a static test that contains the same observable behaviors every year. Thus, there is no annual test construction process.

Security

TEA places a high priority on test security and confidentiality for all aspects of the statewide assessment program. From the development of test items to the construction of tests, and from the distribution and administration of test materials to the delivery of students' score reports, special care is taken to promote test security and confidentiality. TEA investigates every allegation of cheating or breach of confidentiality.

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 Manuals

Test security for the Texas assessment program has been supported by an aligned set of test administration documents that provide clear and specific information to testing personnel. In response to the statutes and administrative rules that are the foundation for policies and documentation pertaining to test security, TEA produces and updates detailed information about appropriate test administration procedures in the *District and Campus Coordinator Manual* and the test administrator manuals.

MANUALS

The annual coordinator manual and test administrator manuals provide guidelines on how to train testing personnel, administer tests, create secure testing environments, and properly store test materials. They also instruct testing personnel on how to report to TEA any confirmed or alleged testing irregularities that might have occurred in a classroom, on a campus, or within a school district. Finally, the manuals provide training and guidelines relative to test security oaths that all personnel with access to secure test materials are required to sign. The manuals give specific details about the possible penalties for violating test procedures.

During the 2018–2019 school year, TEA eliminated the Test Security Supplement, which was a guide designed to help districts implement required testing procedures. Instead, TEA adopted amendments to TAC §101.3031 that included specific language from the supplement detailing the requirements of school districts and charter schools to maintain security and confidently of assessment instruments, including a list of violations and actions that may result from a violation. Other information from the supplement regarding local practices for implementing security policies and procedures was included in the new online resource for district testing coordinators.

Online Training

TEA provides training materials that cover test administration best practices and the maintenance of test security. The online training is divided into three modules: 1) active monitoring, 2) distribution of test materials, and 3) proper handling of secure materials. Although completion of these modules is not a requirement, it is, however, strongly recommended that districts and charter schools use them to help supplement the mandatory training required of all personnel involved in testing. Training modules can be accessed from the training page on the Texas Assessment Management System website.

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 be faced with

- 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.

Any student involved in a violation of test security could have his or her test results invalidated.

Incident Tracking

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

Products and procedures to assist in test administration have been developed to promote test security and include the following:

- 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
- 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

Light Marks Analysis

ETS provides an analysis of light marks for all STAAR test documents in paper format. Scanning capabilities allow for the detection of 16 levels of gray in student responses on scorable documents. During scanning, these procedures collect the darkest response for each item and the location of the next darkest response. These multiple shaded responses often result from an erasure. The changes in the erasures are categorized as wrong-to-right, right-to-wrong, or wrong-to-wrong and are summarized in the erasure analysis report.

Information and descriptive statistics for each group (usually by grade level in each campus) are available in the report. The report includes the following information about each group.

- **County-District-Campus Number** This nine-digit number is the code for the district and campus of the class group being reported.
- **Grade and Subject** This is the grade and subject of the class group being reported.
- **Number of Students** This is the number of students within the class group.
- **All Items** This is the average number of total erasures for the students in the class group.
- **Wrong-to-Right** This is the average number of erasures from incorrect to correct answers.
- **Right-to-Wrong** This is the average number of erasures from correct to incorrect answers.
- **Wrong-to-Wrong** This is the average number of erasures from one incorrect answer choice to another incorrect answer choice.

Statewide statistics for the tests are also reported and include the average erasures of any type, the average and standard deviation of wrong-to-right erasures, and the average number of right-to-wrong and wrong-to-wrong erasures.

It should be stressed that these analyses serve only to identify an extreme number of light marks or erasures. These procedures serve as a screening device and provide no insight into the reason for excessive erasures. Students could, for example, have an extremely high number of erasures if they began marking their answers on the wrong line and had to erase and re-enter answers. Students could also be particularly indecisive and second-guess their answer selections. By themselves, data from light marks analyses cannot provide evidence of inappropriate testing behaviors. Therefore, it is important to consider the results from the light mark analyses within a larger test security process that includes additional evidence such as seating charts, reports of testing irregularities, and records of test security and administration training for districts and campuses.

Statistical Analyses

Each summer, ETS conducts a series of analyses to detect statistical irregularities in STAAR results that could possibly indicate violations of test security. These analyses compare prior-year and current-year STAAR spring results to identify atypical and statistically significant changes in average scale scores and pass rates. Separate analyses are conducted for each STAAR assessment and then aggregated to the campus level (grades 3–5, grades 6–8, or high school). The results from the statistical analyses are compared to the annual erasure analysis report, which flags campuses having atypical rates of wrong-to-right answer changes. Campuses flagged in both areas are prioritized for additional review. By applying multiple independent methods, TEA gathers strong evidential support for inferences about statistical irregularities at the campus level, while minimizing false positives.

Quality-Control Procedures

The Texas assessment program and the data it provides play an important role in decision-making about student performance and in public education accountability. Individual student test scores are used for promotion, graduation, and remediation. In addition, the aggregated student performance results from the student 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 the tests are scored correctly and reported accurately to school districts. TEA verifies the accuracy of the work and the data produced by the testing contractor through a comprehensive verification system. The section that follows describes the quality-control system used to verify the scoring and reporting of test results and the ongoing quality-control procedures in the test-development process.

Data and Report Processing

Prior to reporting test results, an extensive and comprehensive quality-control (QC) process is performed by TEA to verify the quality and accuracy of final reports for Texas assessments. This quality-control process was applied for every state assessment administered in the school year, including:

- STAAR
- STAAR Spanish
- STAAR Alternate 2
- TELPAS
- TELPAS Alternate

The quality-control process involves internal steps taken by ETS and Pearson, as well as implementation of a joint quality-control process supported by TEA and each contractor. ETS and Pearson each implement an internal quality-control system for the reporting of test results. Quality-control testing occurs at two levels: the unit level and the system level. The purpose of the unit test process is to confirm that software modules associated with various business processes, such as online test delivery, scanning, scoring, and reporting, are developed and operating to meet program requirements. The system test confirms that all the modules work together so that outputs from one module match the proper inputs for the next module in the system. The system test is performed by a group that is independent from the software development group. This process allows for independent verification and interpretation of project requirements. Once the independent testing group has completed the test and given its approval, the system is moved into production mode.

The joint TEA/contractor quality-control process is a complete test run of scoring and reporting. TEA begins the quality process months in advance of a test date. For each test administration, TEA and the contractor prepare answer documents and online student response data for thousands of hypothetical students who serve as test cases and who are assigned to a campus in one of three hypothetical districts. Answer documents for each student within this data set are processed like operational data. This processing includes scanning the answer documents, scoring the responses, and generating student- and district-level reports and data files. For online hypothetical student data, this processing includes scoring the responses and generating student- and district-level reports and data files. During every step of the test run, information is independently checked and verified by TEA. Reports are not sent to districts until all discrepancies in the quality-control data set are resolved and the reports generated by TEA and the contractor match. Details of the quality-control process can be found in Appendix A.

In addition to checks performed during the TEA/contractor process, a small sample of operational answer documents is run through all scoring and reporting processes. This serves as an additional quality-control step to test the processing of answer documents. Only after this final quality-control step is completed successfully is the processing of all assessment materials launched.

Technical Processing

In addition to the processing of student answer documents, online data, and generation of reports, psychometric or technical processing of the data also occurs before and

after each test administration. Each type of technical processing includes additional quality-control measures.

Each technical procedure, like scaling and equating, requires calculations or transformations of the data. These calculations are always completed and verified by multiple psychometricians or testing experts at ETS and Pearson. These calculations are then additionally verified and accepted by TEA. In some cases, as with equating, a third party external to TEA, ETS, and Pearson is also included in processing to further enhance the quality-control procedures.

While each year's calculations are verified, they are also compared to historical values to further validate the reasonableness of the results. For example, pass rates from this year were compared to those from previous years. These year-to-year comparisons of the technical procedures and assessment results help to verify the quality of the assessments and to inform TEA of the impact of the program on student achievement.

For more information about the standard technical processes of the Texas assessment program, see chapter 3, "Standard Technical Processes."

Performance Assessments

The STAAR tests include constructed-response items, which required scoring by trained human raters on the following operational assessments in 2018–2019:

- STAAR grade 4 and 7 writing
- STAAR Spanish grade 4 writing
- STAAR English I, English II, and English III
- TELPAS speaking

The Texas assessment program uses written compositions on STAAR and STAAR Spanish, which are a direct measure of the student's ability to synthesize the component skills of writing; that is, the composition task requires the student to express ideas effectively in writing for a specified purpose. To do this, the student must be able to respond in a focused and coherent manner to a specific prompt while organizing ideas clearly, generating and developing thoughts in a way that allows the reader to thoroughly understand what the writer is attempting to communicate, and maintaining a consistent control of the conventions of written language.

For the STAAR and STAAR Spanish assessments, the types of writing required vary by grade and course and represent the learning progression evident in the TEKS.

Written compositions for STAAR are evaluated using a holistic scoring process, meaning that the essay is considered as a whole. It is evaluated according to pre-established criteria: organization/progression, development of ideas, and use of language/conventions. These criteria, explained in detail in the writing scoring rubrics for each grade and type of writing, are used to determine the effectiveness of each

written response. Each essay is scored on a scale of 1 (a very limited writing performance) to 4 (an accomplished writing performance). A rating of 0 is assigned to compositions that are nonscorable. The STAAR writing rubrics can be found on TEA's Student Assessment Division website on the STAAR Resources page.

For the TELPAS Speaking assessment, all student responses are initially scored by an automated scoring engine. To ensure continued validity, reliability, and calibration of the assessment scoring process, 10% of engine-scored responses are reviewed by human raters. Data from these two methods are continuously compared to ensure the process is reliable. Human scoring also takes place for responses identified as "not scorable" by the automated engine. These responses most often have a unique characteristic that makes them ore appropriately scored by a human rater. These unique characteristics may include background noise (school bell rings, static sound in recordings), mumbled or unclear spoken language, and/or the volume of the recorded response is too low and difficult to score. All scorers go through the same extensive training. This process is as standardized as possible and all scorers are trained using the same materials and rubrics. Refer to chapter 6, "Texas English Language Proficiency Assessment System (TELPAS)," for detailed information about the TELPAS Speaking scoring process.

The TELPAS speaking assessment consists of prompts that elicit student speaking responses captured and recorded through the online assessment using a headset with a microphone. Speaking prompts are scored according to a 2- or 4-point rubric depending on the item type. During field-testing, human scorers give ratings on the responses in order to train the automated scoring engine. For operational items, the automated scoring engine scores the responses, while human scorers rate any responses that are considered to be "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 TEA's Student Assessment Division website on the TELPAS Resources page.

Scoring Staff

ETS recruits raters through various mass media and educational organizations. All test raters hired must have at least a four-year college degree and undergo rigorous TEA-approved training before they are allowed to begin work. As part of this training, applicants for rating STAAR compositions must complete a certification process, score practice sets, and pass calibration. Raters are closely monitored on a daily basis, with each student response carefully reviewed by multiple readers to produce scores that are accurate and reliable.

At ETS, the training and monitoring of rater performance is conducted by Scoring Leaders (SLs), content scoring leaders (CSLs), and assessment specialists, all of whom have demonstrated expertise with constructed-response scoring. The SLs guide, support, and monitors Raters during operational scoring sessions. The CSLs guide, support, and monitor the SL during operational scoring. The CSLs will apply will apply all condition codes and reach out to Assessment Development (AD) when they need

guidance. They receive additional training for this role over what SLs and rates receive in order to provide the needed support.

Assessment specialists are responsible for overseeing the scoring of individual assessment items and for building the training materials from field-test responses to represent a full range of scores. During scoring, SLs and CSLs monitor and manage scoring quality by answering rater questions and reviewing scoring reports. Assessment specialists train scoring leadership on both content and job expectations prior to rater training. Program management monitors all aspects of performance scoring for the STAAR assessment program, writes a plan that specifies the configuration of training materials, and manages the schedule and process for performing the work.

For TELPAS Speaking, Pearson advertised through various mass media and educational organizations. All test scorers hired must have at least a four-year college degree and undergo rigorous TEA-approved training before they can score student responses. As part of this training, scorers review the rubric and an anchor set that includes a range of responses to exemplify the score points and delineate the scoring lines. Scorers then take practice sets which reinforce the scoring criteria, and following the completion of the practice sets, scorers must take qualifying sets and qualify by demonstrating a high level of mastery before any student responses are scored. Pearson's content supervisory staff monitor scorer performance daily to ensure accurate and reliable scoring.

Pearson's content supervisory staff consists of scoring supervisors who monitor and work directly with scorers, scoring directors who monitor overall scorer performance and provide direction to the scoring supervisors, and the content specialist who monitors the scoring overall and works directly with the scoring directors to accurate and consistent scoring across all items. All content supervisory staff have demonstrated a high level of expertise and possess years of experience in scoring student assessments. Project management monitors all aspects of performance scoring for TELPAS, develops and executes plans for delivering high quality scoring, and manages the schedule to ensure timely completion of scoring. In 2018–2019, Pearson scored all TELPAS speaking responses on-site at its Austin scoring center and utilized local scorers to complete the work. TEA staff observed the training and scoring of the TELPAS Speaking assessment.

Distributed Scoring

Distributed scoring of STAAR and STAAR Spanish was first used with the Texas assessment program in 2010–2011. Distributed scoring is a system in which raters can participate in the scoring process from any location if they qualify and meet strict requirements. Distributed scoring is a secure, Web-based model that incorporates several innovative components and benefits, including the following:

- The number of raters available locally can be augmented by other highly credentialed raters from across the state and country.
- More teachers across the state are able to participate in the scoring process.
- Paper handling and associated costs and risks are reduced.
- Raters are trained and qualified using comprehensive, self-paced online training modules, which 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 raters.

The Online Network for Evaluation (ONE) System

STAAR written compositions are scored using the ETS ONE system. ONE provides secured access to student handwritten and online delivered constructed responses for raters who have completed training and passed a calibration/qualification test for the applicable prompt. Raters have access to prompt content, TEA-approved rubrics and benchmark papers at any time during training, calibration, and operational scoring. The ONE response viewer renders scanned images and text responses online as they were written/typed by the student. Viewer tools allow raters to adjust contrast, colors, and magnification/zoom levels, which serve to further improve reading clarity, as well as to reduce reading fatigue.

All multiple-choice answers and constructed responses from a particular student and test are linked throughout ETS scoring and reporting processes via a unique identifier. This identifier is associated to each handwritten response during the scanning and image-clipping processes and to online-entered responses after capture. In ONE, student identifiers and other demographic information are not visible to raters in order to protect student anonymity and to reduce bias during scoring.

The responses are grouped by grade or course and are stored on the ONE server. Only qualified scoring leaders, content scoring leaders, and assessment specialists have access to this server. As raters score the responses, more responses are routed into their scoring queues. Each rater independently reads a response and selects a score from a menu on the computer screen. Scoring leaders, content scoring leaders, and assessment specialists can identify which rater reads each response.

Refer to the Performance Assessments section, or to chapter 6, “Texas English Language Proficiency Assessment System (TELPAS),” for information about the TELPAS Speaking scoring process.

Pearson’s Image-based Online Scoring System – Electronic Performance Evaluation Network (ePEN)

Although the automated scoring engine scores the vast majority of TELPAS speaking responses, sometimes responses need a human score, and for these, Pearson scored

them through ePEN, providing secure access to the students' audio files and real-time scoring reports for content supervisory staff. Each scorer independently listens to a response and selects the appropriate score in the scoring grid. When reviewing the scorers' work, content supervisory staff can always identify who scored a particular response. Students' personally identifiable information along with other demographic information are not visible to ePEN users in order to protect student anonymity and to reduce bias during scoring. ePEN provides a wealth of tools and reports to help supervisory staff monitor scoring. Through calibration within ePEN, the rubric and training can be reinforced through calibration sets delivered both regularly and when needed to address a scoring issue.

Rater Training Process

All raters who work on the STAAR and STAAR Spanish performance task scoring projects receive extensive training through the ETS Learning Management System (LMS) online modules. This training covers the materials associated with the prompts for each assessment. In addition, training for STAAR scoring includes orientation within the ONE system. Raters receive training on the scoring guide that provides the rubric and examples of each rubric score point for a particular assessment item. These examples are called "benchmark papers." Additionally, raters score training set responses that have predetermined scores. They also have an opportunity to explanation and discuss the scores. Raters are required to demonstrate a complete understanding of the rubrics and to pass a set of responses called the "calibration set," before being allowed to score operational student responses.

WRITTEN COMPOSITIONS

Raters first complete a rubric overview training before training on item specific responses. The Item Specific Training (IST) modules are a set of student compositions that have already been scored by assessment specialists and TEA staff. The training materials are selected to clearly differentiate student performance at the different rubric score points and to help raters learn the difference between score points. The training materials also contain responses determined to be borderline between two adjacent score points to help raters refine their understanding of differences between adjacent score points. Scoring leaders are available during rater training to assist and answer questions. Once raters complete the training sets, they are administered a calibration set of student compositions on the first day that they are scheduled to score responses. As with the training sets, the student compositions in the calibration set have already been scored by assessment specialists and TEA staff. All the raters must accurately assign scores to student responses in the calibration set. Raters are given two opportunities to qualify, with a different set of responses in each set. Raters are also required to recalibrate at regular intervals throughout the scoring window. Any rater who is unable to meet the standards established by ETS and TEA is dismissed from scoring.

ONGOING TRAINING

After initial training, ongoing training is available to ensure scoring consistency and high rater agreement. Scoring leaders and content scoring leaders monitor scoring and provide mentoring continually during the operational scoring. The ONE scoring system includes a comprehensive set of scoring and monitoring tools such as backreading, validity, and reporting functions, which help identify areas for additional training.

Scoring Process

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

In instances when the scores are discrepant (scores from rater 1 and rater 2 differ by more than 1 point) the student response is routed to a resolution queue and the response is reviewed by a content scoring leader. Only content scoring leaders have access to the resolution queue. The content scoring leader will review the student response and apply a third score. This third score invalidates the two initial scores, by rater 1 and rater 2, and this score is doubled and becomes the reported score.

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

NONSCORABLE RESPONSES

Before an essay can be given a nonscorable designation, the response is thoroughly reviewed by the content scoring leader. If the content scoring leader determines that the response is scorable, it is assigned a score and routed to a second content scoring leader. If the content scoring leader determines that the response is nonscorable, a nonscorable code is applied, and the response is routed to a second content scoring leader for confirmation. Only a content scoring leader can determine if a student response should be scored as nonscorable. While the response is under review, it is held in a review queue that prevents it from being distributed to other raters.

MONITORING OF RATER QUALITY

Raters are closely monitored by their Scoring Leader (SL), who can provide feedback and guidance during scoring. In addition, raters can defer student responses to their SL who can provide feedback on how to score the response or pass the question along to the Content Scoring Leader (CSL) for that prompt. This allows raters to receive regular feedback on their performance. Responses scored by a rater who is identified as having difficulty applying the criteria has his or her scores invalidated and rescored and completes remediation training. Any rater who cannot successfully pass the remediation training set is dismissed from scoring.

The SL guides, supports, and monitors raters during operational scoring sessions. The CSL guides, supports, and monitors the SL during operational scoring. The CSLs will apply all conditional codes and reach out to Assessment Development when guidance is needed. CSLs receive additional training for this role in order to provide the needed support.

Validity responses are student responses that have already been assigned a score during rangefinding and are presented to raters throughout the operational-scoring process to monitor the quality of their scoring. All validity responses are approved by TEA before being introduced into the scoring systems. Validity responses cannot be distinguished from operational responses and are inserted randomly into the scoring queue and scored by raters. Rater accuracy can be evaluated based on the agreement of the rater validity score and the original validity score.

For TELPAS, scoring supervisors closely monitor their scorers, providing feedback and guidance to continually improve scoring accuracy. ePEN allows a supervisor to back-listen to responses scored and send that scorer feedback through the ePEN messaging system. Scorers can also send responses to review, so that a scoring supervisor or scoring director can listen and provide feedback. Along with these, a key tool in monitoring scorer performance are validity responses. All of these responses have had their scores approved by TEA and 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.

RANGEFINDING

TEA, ETS, and Pearson staff independently score samples of the field-test responses to the prompts to be used on the operational assessments. This scoring is in addition to the scoring already done by field-test raters. TEA and ETS content and management staff, including the respective assessment specialists, participate in a series of meetings called rangefinding: to analyze these responses and to assign “true” scores. The assessment specialists select responses from the rangefinding sessions to be included in each scoring guide. The assessment specialists then assign the remaining pre-scored responses from the rangefinding sessions to training sets and qualifying sets for use in future rater training. Prior to scoring, TEA staff review and approve all scoring guides and training sets.

Score Reliability and Validity Information

Throughout the years, TEA has reported on the reliability and validity of the performance-scoring process. Reliability has been expressed in terms of rater agreement (percentage of exact agreement between rater scores) and correlation between first and second ratings. Validity has been assessed by the inclusion of validity responses throughout the operational-scoring process. It is expressed in terms of exact agreement between the score assigned by the rater and the “true” score assigned by ETS and approved by TEA.

Appeals

If a district has questions about the score assigned to a response, a rescore can be requested through submission of the appropriate request form. ETS provides rescore results by posting an updated STAAR Report Card (SRC) to the STAAR Assessment Management System, only if the score has changed. If the score does not change, there is a fee that districts pay. If the score changes, that fee is waived. 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 that explains the final outcome of the appeal, and whether or not the score was changed, will be provided. In 2019, there was no appeal process in place for TELPAS Speaking.