Chapter 2 Building a High-Quality Assessment System

Test Development Activities

Texas educators—K–12 classroom teachers, higher education representatives, curriculum specialists, administrators, and Education Service Center (ESC) staff—play a vital role in the test-development process. The involvement of these education professionals enables the development of high-quality assessments that accurately measure what Texas students have learned in the classroom.

Thousands of Texas educators have served on one or more of the educator committees that are 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. They include educators with knowledge of the needs of a variety of student populations, including students with disabilities and English language learners (ELLs).

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), to develop appropriate assessment categories for a specific grade/subject or course. 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, assessment specialists, and administrators.

3 A draft of the reporting categories and TEKS 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 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 questions.

6 With educator input, a preliminary test blueprint is developed that sets the length of 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 questions based on the reporting categories, the TEKS student expectations, and the item guidelines.

***8** TEA content specialists from the curriculum and assessment divisions review and revise the proposed test questions.

***9** Item review committees composed of Texas educators review the revised test questions 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-review committees are trained in statistical analysis of field-test data and review each question and its associated data. The committees determine whether questions are appropriate for inclusion in the bank of questions from which test forms are built.

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

***14** All field-test questions and data are entered into a computerized item bank. TEA staff build tests from the item bank so that the tests are equivalent 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 for accuracy because of the advanced level of content being assessed.

***16** Tests are administered to Texas students. Results of these tests are reported at the student, campus, district, regional, and state levels.

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

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 annually to provide verified technical information about the tests to schools and the public.

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

Groups Involved

A number of 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

The Texas Education Agency's (TEA) 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 ETS and Pearson. TEA staff members conduct quality-control activities for every aspect of the development, administration, scoring, and reporting of the assessment program and monitor the program's security provisions.

ETS

Educational Testing Service (ETS) is the contractor for the provision of support services to the state for the STAAR program. ETS also serves as the program integration contractor. This role includes working with Pearson and various subcontractors to make sure that the Texas 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 2015–2016 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 Texas Assessment of Knowledge and Skills (TAKS). Due to the diverse nature of the services required, Pearson employs subcontractors to perform tasks requiring specialized expertise. During the 2015–2016 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., specializes in the creation of writing passages and test items. As a subcontractor of Pearson, Lone Star Assessment and Publishing works with Pearson personnel, TEA staff members, and Texas educators to develop complex stimuli and test items for TELPAS reading.

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 particular content that aligns with the curriculum and instruction.

Draft reporting categories with corresponding Texas Essential Knowledge and Skills (TEKS) 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 and Lone Star assume the major role for STAAR Alternate 2 and TELPAS item development, agency personnel are involved throughout the item development process. All items developed for these tests are the property of TEA.

Item Guidelines

Item and performance task specifications provide guidance from TEA on how to translate the TEKS into actual test items. Item guidelines are strictly followed by item writers in order to enable the accurate measurement of the TEKS student expectations. In addition, guidelines for 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, English language proficiency tests (for TELPAS item development), and large-scale criterion-referenced measurements. These individuals are selected for their background in second-language acquisition (for TELPAS) or specific content-area knowledge and their teaching or curriculum development experience in the relevant grades.

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 English Language Proficiency Standard (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.

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 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 and Student Assessment Divisions, who are content experts in the grades and content areas for which items are developed, scrutinize 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 staff from TEA meet with ETS (for STAAR) and Pearson (for STAAR Alternate 2 and TELPAS) to examine, discuss, and edit all newly developed items before each educator item review committee meeting.

Item Review Committee

Each year TEA's Student Assessment Division convenes committees composed of Texas classroom teachers (including general education teachers, special education teachers, and Bilingual and 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, content-area specialists in TEA's Curriculum Division, and staff from other agency divisions. Recommendation forms are provided to districts and ESCs on the Assessment Resources for Teachers and Administrators page on TEA's Student Assessment Division website. Item review committee members are selected based on their established expertise in a particular 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's Student Assessment Division 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, revised, recoded to a different eligible TEKS student expectation or ELP standard, 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? • 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.

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.

Field Testing and Data Review

Field testing is conducted prior to a test item being used on an operational test form. However, when there is curriculum change, 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.

In 2016, STAAR grades 3–8 assessments did not include embedded field-test items to fulfill the requirement in House Bill 743 for shorter tests. Embedded field testing will resume in 2017. There was no field testing of written compositions for English I, II, or III assessments in 2016. Starting in 2018, written compositions will be field-tested through prompt studies for all assessments that include written compositions. In these studies, which occur separately from STAAR test administrations once every three years, a representative sample of Texas students respond to newly developed written compositions.

Past experience has shown that both field-test procedures yield sufficient data for precise item evaluation and allow 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.

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 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 meet with ETS assessment specialists (for STAAR and STAAR Spanish) and Pearson assessment specialists (for STAAR Alternate 2 and TELPAS) to examine each test item and its associated data with regard to 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 test.

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. In addition, TEA, ETS, and Pearson maintain paper copies of each test item.

Each electronic item bank also stores item data, such as the unique item number (UIN), grade, 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 uses the item statistics and other information about items during the test construction process to regulate test difficulty and adjust the test for content coverage and balance. Each electronic item bank can generate files of item information for review or printing.

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 and the identification of readiness and supporting standards 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

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;
- the level of difficulty of the skills represented in the TEKS; and
- the application of content and skills in different contexts, both familiar and unfamiliar.

TEA constructs tests 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 scaling 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 year to year. 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, which includes 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 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, English, 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 that are of concern. When necessary, substitute items are chosen and reviewed. After content validation is complete, the assessments are ready to be administered.

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 equivalent 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.

The *Test Security Supplement* and 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 *Test Security Supplement*, *District and Campus Coordinator Manual*, and the test administrator manuals.

TEST SECURITY SUPPLEMENT

Beginning in 2012, the commissioner of education adopted the *Test Security Supplement* into the Texas Administrative Code, 19 TAC §101.3031(b)(2). Updated annually, this guide is designed to help districts implement required testing procedures and foster best practices for maintaining a secure testing program.

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.

Online Training

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

14-Point Test Security Plan

To bolster ongoing efforts to improve security measures in the state's assessment program, TEA introduced a comprehensive 14-point plan in June 2007 designed to assure parents, students, and the public that test results are meaningful and valid. The document, Recommendations for Implementation of the 14-point Test Security Plan, is available on TEA's Student Assessment Division website.

Security Violations

In accordance with 19 TAC §101.3031(b)(2), *Test Security Supplement*, 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 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 Light Marks Analysis Report.

The Light Marks Analysis Report lists every class group whose average number of wrong-to-right erasures is greater than three standard deviations above the statewide average for each subject within each grade tested. Districts determine the composition of these class groups by how they complete the “Class Identification Sheet” and how they assemble answer documents beneath each Class Identification Sheet.

Information and descriptive statistics for each flagged class group are available in the report. The report includes the following information about flagged class groups.

- **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.
- **Class Group.** This is the class group name gridded on the class identification sheet 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 (and percentage) 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 statistical 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

During the summer of 2016, TEA conducted a series of analyses to detect statistical irregularities in STAAR results that could possibly indicate violations of test security. These analyses compared spring 2015 and spring 2016 STAAR results to identify atypical and statistically significant changes in average scale scores and pass rates. Separate analyses were 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 were compared to the annual Light Marks Report, which flags campuses having atypical rates of wrong-to-right answer changes. Campuses flagged in both areas were prioritized for additional review. By applying multiple independent methods, TEA gathered 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 process is enacted to verify the accuracy of final reports for Texas assessments. This quality control process was applied for every state assessment administered in 2015–2016, including

- STAAR
- STAAR Spanish
- STAAR L
- STAAR A
- STAAR braille
- STAAR Alternate 2
- TAKS
- TELPAS

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, like 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 considered in comparison to historical values to further validate the reasonableness of the results. For example, pass rates from 2015–2016 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

STAAR and TAKS included constructed-response items, which required scoring by trained human readers on the following operational assessments in 2015–2016:

- STAAR grade 4 and 7 writing
- STAAR Spanish grade 4 writing
- STAAR A grade 4 and 7 writing
- STAAR English I, II, and III
- STAAR A English I and II
- TAKS exit level English language arts (ELA)

The Texas assessment program includes two different types of constructed-response items—written compositions and short answer reading responses. Written compositions 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. Short answer reading responses are designed to test students' ability to understand and analyze published pieces of writing. Students must be able to generate clear, reasonable, and thoughtful ideas or analyses about some aspect of the published literary and informational selections. In addition, students must be able to support these ideas or analyses with relevant, strongly connected textual evidence.

For the STAAR, STAAR Spanish, and STAAR A assessments, the types of writing required vary by grade and course and represent the learning progression evident in the TEKS. For TAKS, the ELA tests at exit level include a single written composition that requires students to write a personal essay.

Written compositions for both STAAR and TAKS are evaluated using the holistic scoring process, meaning that the essay is considered as a whole. For STAAR, STAAR Spanish, and STAAR A, 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 STAAR, STAAR Spanish, and STAAR A 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 writing rubrics can be found on TEA's Student Assessment Division website on the STAAR Resources page and the TAKS Resources page.

The STAAR English I, English II, and English III and STAAR A English I and English II tests include two short answer responses. The TAKS exit level ELA tests include three short answer reading responses. The criteria are explained in the scoring rubrics for short answer responses for both single selection and connecting selections. The short answer rubrics can be found on the STAAR Resources page and the TAKS Resources page.

Scoring Staff

ETS and Pearson each recruit readers through various mass media. In addition, ETS also uses various educational organizations to recruit readers. All test readers 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 STAAR must complete a certification process, take practice sets, and pass calibration. Applicants for TAKS complete practice sets and pass qualifying sets before being eligible to work. Readers 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 reader performance is conducted by scoring leaders, chief scoring leaders, and assessment specialists, all of whom have demonstrated expertise with constructed-response scoring. 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, scoring leaders and chief scoring leaders monitor and manage scoring quality by answering reader questions and reviewing scoring reports. Assessment specialists train scoring leadership on both content and job expectations prior to reader 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.

At Pearson, the training and monitoring of reader performance is conducted by scoring directors and supervisors, all of whom also have extensive experience with the Texas assessment program and numerous other large-scale writing assessments. TEA approves all management-level staff at the scoring centers, including the scoring directors for the various projects. Scoring directors are responsible for overseeing the scoring of individual assessment items and building the training materials from field-test responses to represent a full range of scores. During scoring, supervisors help scoring directors monitor and manage scoring quality by answering reader questions and reviewing scoring reports. Scoring supervisors are trained on both content and job expectations prior to reader training. If possible, people with previous scoring experience are hired as supervisors. The project monitor supervises all aspects of performance scoring for the TAKS assessment program, writes a plan that specifies the configuration of training materials, and manages the schedule and process for performing the work.

Pearson’s Austin-based Scoring Services oversees scoring of all essays and short answer reading responses for the TAKS assessment program. In addition, Pearson’s Scoring Service collaborates with TEA on the training of scoring supervisors. Scoring Services recruits and hires scoring personnel, coordinates the handling of student papers, maintains security, and transmits scoring data to the scoring system.

Distributed Scoring

Distributed scoring was first used with the Texas assessment program in 2010–2011. Distributed scoring is a system in which readers 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 readers available locally can be augmented by other highly credentialed readers 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.
- Readers 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 readers.

The Online Network for Evaluation (ONE) System

STAAR written compositions and short answer responses are scored using the ETS Online Network for Evaluation (ONE) system. ONE provides secured access to student handwritten and online delivered constructed responses for readers who have completed training and passed a calibration/qualification test for the applicable prompt. Raters have access to prompt content, TEA approved rubrics and anchor papers (aka “benchmarks”) 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 readers to adjust contrast, colors, and magnification/zoom levels, which serves 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 readers, 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 directors, readers, and project monitors have access to this server. As readers score the responses, more responses are routed into their scoring queues. Each reader independently reads a response and selects a score from a menu on the computer screen. Scoring supervisors, scoring directors, and project monitors can identify which reader reads each response.

Reader Training Process

All readers and scoring leaders/supervisors who work on the STAAR, STAAR Spanish, and TAKS performance task scoring projects receive extensive training, including training through online modules or onsite training. This training covers the materials associated with the prompts and/or short answer responses for each assessment. In addition, training for STAAR scoring includes orientation within the ONE system. Readers 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 “anchor papers.” Additionally, readers score training set responses that have predetermined scores and have the opportunity for explanation and discussion of those scores. Readers are required to demonstrate a complete understanding of the rubrics before operational scoring begins. Readers are required to perform satisfactorily on sets of responses called “calibration sets.” Any reader who cannot demonstrate satisfactory performance on these sets is dismissed. Only readers who successfully undergo the complete training and qualifying process are allowed to begin scoring operational student responses.

Written Compositions

Readers first complete a training set of student compositions. The student compositions in the training set have already been scored by assessment specialists (for STAAR and STAAR Spanish) or scoring directors (for TAKS) and TEA staff. The training materials are selected to clearly differentiate student performance at the different rubric score points and help readers learn the difference between score points. The training materials also contain responses selected to be borderline between two adjacent score points and help readers refine their understanding of differences between adjacent score points. An assessment specialist/scoring director leads the discussions and answers any questions about the training set. Once readers complete the training sets, they are administered a calibration set of student compositions. As with the training sets, the student compositions in the calibration set have already been scored by assessment specialists/scoring directors and TEA staff. All the readers must accurately assign scores to student responses on the calibration set. Readers are given two opportunities to qualify, with a different set of responses in each set. Readers are also required to recalibrate at regular intervals throughout a scoring project. Any reader who is unable to meet the standards established by TEA is dismissed.

Short Answer Responses

Before training, readers are assigned to a group corresponding to each short answer response. This allows each group to focus fully on a particular short answer response without being distracted by the other question(s). Any questions about the material are answered. Readers work through the training sets, which contain examples of short answer responses that have already been scored by assessment specialists, scoring directors, and TEA staff. An assessment specialist or scoring director leads the discussions and answers any questions about the training sets. Once readers complete the training sets, they are administered a calibration/qualifying set of student responses. As with the training sets, the student responses in the calibration set have already been scored by assessment specialists, scoring directors, and TEA staff. All readers for STAAR must accurately assign scores to students’ responses in a calibration set. All readers for TAKS must accurately assign scores to 80 percent of the student responses in a qualifying set. Readers are given two opportunities to qualify, with a different set of responses in each set. Readers are also required to recalibrate at regular intervals throughout a scoring project. Any reader unable to meet the standards established by TEA is dismissed.

Ongoing Training

After initial training, ongoing training is available to ensure scoring consistency and high reader agreement. Scoring leaders, directors, and chief scoring leaders monitor the scoring and provide mentoring continually during the operational scoring. In addition, for TAKS, scoring directors plan for at least three ongoing training sessions a week. Every week the scoring directors review the rubrics with readers and have them reread their anchor papers, emphasizing any area that appears to be giving readers problems. The scoring system includes a comprehensive set of scoring and monitoring tools, such as backreading, calibration, and reporting functions, which helps identify areas for additional training.

Scoring Process

Two different types of reader agreement metrics are used with the Texas assessment program—adjacent and exact agreement. In an adjacent agreement model, each student response is independently scored by two readers. If the student response receives exact or adjacent scores (scores that differ by one point), the scores are summed to create the reported score. Student responses that receive non-adjacent scores receive additional review and scoring by the scoring director (refer to the Resolution Procedures section of this chapter). Similarly, in an exact agreement model, each student response is independently scored by two readers. However, if the response receives exactly the same score from the two readers, this value is used as the reported score. Student responses for which scores do not agree exactly receive additional review and scoring by highly trained staff (refer to the Resolution Procedures section of this chapter).

The STAAR, STAAR Spanish, and STAAR A written compositions are scored using an adjacent agreement scoring model. Each reader assigns a score from 1 to 4 for STAAR, STAAR Spanish, and STAAR A. The reported (summed) score for STAAR, STAAR Spanish, and STAAR A ranges from 2 to 8. Summed score performance information is provided to districts on both the Confidential Student Report (CSR) for individual students and on the Constructed Responses Summary Report for individual campuses and districts. The STAAR, STAAR Spanish, and STAAR A short answer responses are scored using an exact agreement scoring model. Reported score information ranges from 0 to 3.

TAKS written compositions are scored using an exact agreement model, with reported score information ranging from 1 to 4. The TAKS short answer responses are scored using an exact agreement scoring model, with reported scores ranging from 0 to 3.

Resolution Procedures

After a reader has completed a first reading of a student response, the response is routed into a second reader’s queue for an independent reading. Following completion of both the first and second readings, responses that do not meet the score agreement criteria are routed into a resolution queue. Only readers identified as above average in the accuracy of their scoring are allowed to be resolution (or third) readers. Occasionally, a fourth reading of a student response is necessary if the initial two readers and the resolution reader all differ in their scores more than the agreement criteria allow. For example, if a short answer response is given a 1 and a 2 by the initial readers and a 3 by the resolution reader, a fourth reading would be required. When this occurs, the fourth readings are placed in a separate queue and scored only by scoring directors or project monitors. Throughout the scoring project, TEA staff members are consulted on decision papers, which are responses that are highly unusual or require a policy decision from TEA.

After the scores for the first and second readings of a response have been processed, the scoring systems create the resolution readings (third readings and fourth readings), if needed. Project status reports based on data collected for first, second, third, and fourth readings give senior staff up-to-date information on the progress of the entire project.

NONSCORABLE RESPONSES

Before an essay can be given a nonscorable designation, the response is thoroughly reviewed by a scoring leader/supervisor or the chief scoring leader/scoring director. If the review determines that the response is scorable, it is assigned a score and routed to a second reader. If the leader/directors agree that the response is nonscorable, a content scoring expert (at ETS for STAAR) or a project monitor (at Pearson for TAKS) is brought in to conduct an independent reading of the response. While the response is under review, it is held in a review queue that prevents it from being distributed to other readers.

MONITORING OF READER QUALITY

Readers are closely monitored by their scoring leader or supervisor, the chief scoring leader, and content experts or project monitors. Readers can also send difficult-to-score responses to their scoring leader or supervisor, who can respond to the reader or pass the question along to the chief scoring leader or scoring director or the content expert or project monitor for that prompt. This allows readers to receive regular feedback on their performance. Responses scored by a reader who is identified as having difficulty applying the criteria are retrieved and rescored by his or her scoring leader or supervisor or by a reader with above average scoring accuracy. Any reader who cannot be successfully retrained on the criteria is dismissed.

Validity responses are student responses that have already been assigned a score by a scoring leader/director but that are presented to readers 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. The systems allow project staff to include validity responses so that readers cannot distinguish them from operational responses. The validity responses are inserted randomly into the scoring queue and scored by raters. Reader accuracy can be evaluated based on the agreement of the reader validity score and the original validity score.

FIELD-TEST SCORING

After all operational scoring is completed, small groups of experienced readers are selected to score the responses generated by representative samples of students during field testing. Student performance on field-test prompts and short answer responses provides information that helps determine which prompts and questions will be selected for future operational administrations. In addition, field-test responses form the basis for the reader training materials once a prompt or a short answer question is placed on an operational test. Field-test readers score the responses as they would during an operational administration.

Following the scoring of the field-test responses for STAAR and STAAR Spanish, ETS staff compile a summary of the performance of each prompt and short answer question, focusing on such factors as the variety of content seen in the responses, the variety of approaches used, the clarity of the wording of the prompt/short answer question, and an overall impression of the suitability of the prompt/short answer question for possible administration on an operational state assessment. These summaries, along with the statistical data from the scoring process, are presented to TEA for discussion and comment during data review.

RANGEFINDING

TEA and ETS staff independently score samples of the field-test responses to the prompts and short answer questions to be used on the operational assessments. This scoring is in addition to the scoring already done by field-test readers. TEA and ETS content and management staff, including the respective assessment specialists, participate in a series of meetings called “rangefinding sessions” 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 prescored responses from the rangefinding sessions to training sets and qualifying sets for use in future reader training. Prior to the scoring project, 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 reader agreement (percentage of exact agreement between reader scores) and correlation between first and second readings. 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 a given reader and the “true” score assigned by ETS and Pearson 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. Both ETS and Pearson provide rescore results by posting an updated Confidential Student Report (CSR) to the Texas Assessment System. If the score did not change, there is a fee that districts pay. If the score did change, 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.