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 special 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 to develop appropriate assessment categories for a specific grade and/or subject test. For each grade or course, educators provide advice on an assessment model or structure that aligns with good 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 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 items 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-item formats and include sample items.

6 With educator input, a preliminary test blueprint is developed that sets the length of the test and the number of test items measuring each reporting category.

*7 Professional item writers, many of whom are former or current Texas educators, develop items based on the reporting categories and the item guidelines.

*8 TEA curriculum and assessment specialists review and revise the proposed test items.

*9 Item review committees composed of Texas educators review the revised items to judge the appropriateness of item content and difficulty and to eliminate potential bias.

*10 Items are revised again based on input from Texas educator committee meetings and are field tested with large representative samples of Texas students.

*11 Field-test data are analyzed for reliability, validity, and possible bias.

*12 Data-review committees are trained in statistical analysis of field-test data and review each item and its associated data. The committees determine whether items are appropriate for inclusion in the bank of items from which test forms are built.

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

*14 All field-test items and data are entered into a computerized item bank. Tests are built from the item bank and are designed to be equivalent in difficulty from one administration to the next.

*15 Content validation panels composed of university-level experts in each of the fields of English language arts (ELA), mathematics, science, and social studies review each high school-level test for accuracy because of the advanced level of content being assessed.

*16 Tests are administered to Texas students; results are reported at the student, campus, district, regional, and state levels for state-mandated assessments.

*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 will release 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 new 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

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 contract with Pearson. TEA staff members also conduct quality-control activities for every aspect of the development and administration of the assessment program and monitor the program's security provisions.

Pearson

Pearson is TEA's primary contractor for the provision of support services to the state assessment program. Because of the diverse nature of the services required, Pearson employs subcontractors to perform tasks requiring specialized expertise. During the 2013–2014 school year, Pearson's subcontractors for test development activities were Educational Testing Service (ETS); Tri-Lin Integrated Services, Inc. (Tri-Lin); and Lone Star Assessment and Publishing, L.L.C.

ETS

ETS specializes in the development of test items. As a subcontractor of Pearson, ETS works with Pearson personnel, TEA staff members, and Texas educators to produce mathematics, reading, and social studies items.

Tri-Lin

Tri-Lin Integrated Services, Inc., specializes in translation and transadaptation of test items from English into Spanish. As a subcontractor of Pearson, 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 Pearson personnel, TEA staff members, and Texas educators to develop unique passages and/or items for the reading and writing assessments in Spanish.

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 the State of Texas Assessments of Academic Readiness (STAAR®) and STAAR Modified writing assessments; modifies passages and items for the STAAR Modified writing assessments; and delivers item

development and expert review services for mathematics, reading, science, and social studies items.

Texas Educators

Texas educators, including K–12 classroom teachers, higher education representatives, curriculum specialists, administrators, and ESC staff, play a vital role in the test-development process. 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 closely with the curriculum and good classroom 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.

Additionally, Texas educators participate in meetings to define the grade-specific performance level descriptors (PLDs) and to recommend the performance standards on the assessments.

Item Development and Review

This section describes the item-writing process used during the development of Texas assessment program items. While Pearson assumes the major role for item development, many subcontractors and agency personnel are involved in 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

Pearson and its subcontractors employ item writers who have extensive experience developing items for standardized achievement tests and large-scale criterion-referenced measurements. These individuals are selected for their specific content-area knowledge and their teaching or curriculum development experience in the relevant grades. For each assessment, TEA receives an item-tally sheet that displays the number of test items submitted for each reporting category and TEKS student expectation. Item tallies 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

Pearson and its subcontractors provide extensive training for each item writer prior to item development. During these training seminars, Pearson or its subcontractors 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 Pearson and its subcontractors, as well as 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, which occurs annually, includes a check for content accuracy and fairness of the items for different demographic groups. Pearson instructs reviewers to 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"). Pearson also directs its reviewers to consider the more global issues of passage appropriateness; passage difficulty; interactions among items within passages and between passages; and appropriateness of artwork, graphics, or figures. The items are examined by Pearson editorial staff before they are submitted to TEA for review. Items developed for the STAAR EOC assessments also undergo review by outside content experts.

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; grade appropriateness; clarity of wording; content accuracy; plausibility of the distractors; and identification of any potential economic, regional, cultural, gender, or ethnic bias. Then staff from TEA, Pearson, and, if applicable, the subcontractor meet 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 English language learner 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 other agency divisions. Nomination 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 Pearson, ETS, Tri-Lin, and/or Lone Star staff, train committee members on the proper procedures and the criteria for reviewing newly developed items. Committee members judge each item for 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 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 to choose items for field-testing.

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 field-test items 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 order to minimize burden on test takers, prompts for STAAR grade 4 writing were field-tested through a prompt study in 2014, whereas prompts for STAAR grade 7 writing and English I and English II assessments were embedded in operational tests in 2014. Starting in 2017, prompts will be field-tested through prompt studies for all assessments that include prompts. In these studies, which occur separately from STAAR test administrations once every three years, a representative sample of Texas students responds to newly developed prompts.

Past experience has shown that these 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 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/or ethnicity.

Data-Review Procedures

After field testing, Pearson and TEA curriculum and assessment specialists meet 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

Pearson maintains an electronic item bank for the Texas assessment program. The item bank stores each test item and its accompanying artwork. In addition, TEA and Pearson maintain paper copies of each test item.

The electronic item bank also stores item data, such as the unique item number (UIN), grade, subject, reporting category/TEKS student expectation measured, dates the item was administered, and item statistics. The statistical item bank 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. The 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 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 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 the TEA 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, and

- the level of difficulty of the skills represented in the TEKS,
- 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 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 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.

After completion of test construction, 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 it is administered. 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 reviewed and chosen. 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 state's 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. In addition, TEA investigates every allegation of cheating or breach of confidentiality.

Test Security Supplement

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

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.

Manuals and Test Security

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. These manuals provide guidelines about how to train testing personnel, administer the 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.

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

Online Training

TEA provides training materials that cover test administration best practices and the maintenance of test security on the Texas Assessment website. 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, recommended that districts and charter schools use these modules to help supplement the mandatory training required of all personnel involved in testing.

Security Violations

In accordance with 19 TAC §101.3031(b)(2), the *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.

Light Marks Analysis

Pearson 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 oftentimes, result from an erasure. This information is 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 represents the code for the district and campus of the class group being reported.

- **Grade and Subject.** This represents the grade and subject of the class group being reported.
- **Class Group.** This represents the class group name gridded on the class identification sheet of the class group being reported.
- **# of Students.** This represents the number of students within the class group.
- **All Items.** This represents the average number of total erasures for the students in the class group.
- **Wrong-to-Right.** This represents the average number (and percentage) of erasures from incorrect to correct answers. This value is likely of primary interest.
- **Right-to-Wrong.** This represents the average number of erasures from correct to incorrect answers.
- **Wrong-to-Wrong.** This represents the average number of erasures from one incorrect answer choice to another incorrect answer choice.

In addition, statewide statistics for the tests are reported, including 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.

During the spring and summer of 2014, TEA conducted a series of pilot studies to detect statistical irregularities in STAAR results, which could indicate possible violations of test security. These analyses compared spring 2013 and spring 2014 STAAR results to identify atypical and statistically significant changes in average scale scores and pass rates. The first study examined changes in STAAR performance at the overall campus level. For example, mathematics results from grades 6, 7, and 8 were combined at each middle school campus, which was made possible by the vertical scale in STAAR 3–8 mathematics. The second study replicated the methodology of the first study with a finer grain of analysis. Specifically, separate analyses were conducted for each STAAR assessment and then aggregated to the campus level. Campuses flagged for statistical irregularities on multiple assessments could be prioritized for additional investigation or monitoring. The first approach has

the advantage of larger sample sizes (and greater measurement precision), but the second approach can be applied to assessments that do not have vertical scales and it can potentially detect statistical irregularities at certain grade levels. An operational implementation of these analyses will occur in summer 2015. Results from that analysis will be merged with the annual erasure analysis, which flags campuses having atypical rates of wrong-to-right answer changes. By applying multiple independent methods, TEA will gather stronger 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 statewide testing 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 implemented for every state assessment administered in 2013–2014, including

- STAAR 3–8
- STAAR EOC
- STAAR L
- STAAR Modified
- STAAR Alternate
- TAKS
- TAKS (Accommodated)
- TELPAS

The quality-control process involves internal steps taken by Pearson, as well as implementation of a comprehensive Quality Control System (QCS) that is jointly

supported by TEA and Pearson. Pearson implements 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 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. It is then ready for the QCS process to begin.

Essentially, this QCS process is a complete test run of scoring and reporting. TEA begins the QCS process months in advance of a test date. For each test administration, Pearson and TEA 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 Pearson match. Details of the QCS process can be found in Appendix A.

In addition to checks performed during the QCS process, a small sample of operational answer documents is run through all scoring and reporting processes as part of a step called the first production run process (FPRP). 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 Pearson. These calculations are then additionally verified and accepted by TEA. In some cases, like equating, a third party external to TEA 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 2013–2014 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 2013–2014:

- STAAR grade 4 and 7 writing
- STAAR Spanish grade 4 writing
- STAAR English I and II writing (fall 2013)
- STAAR English I and II reading (fall 2013)
- STAAR English I and II (spring and summer 2014)
- STAAR Modified grade 4 and 7 writing
- STAAR Modified 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 assessments, the types of writing required vary by grade and course and represent the learning progression evident in the TEKS (personal narrative and expository in grade 4, personal narrative [with extension] and expository in grade 7,

expository in English I, and persuasive in English II). For the STAAR Modified assessments, a single purpose for writing is required to represent the learning progression evident in the TEKS (personal narrative in grade 4, expository in grade 7, literary in English I, and expository in English II). For TAKS, the ELA tests at exit level include a single written composition that requires students to write a personal essay.

Written compositions are evaluated using the holistic scoring process, meaning that the essay is considered as a whole. For STAAR, 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 essay is scored on a scale of 1 (a very limited writing performance) to 4 (an accomplished writing performance). Each STAAR Modified essay is scored on a scale of 1 (a very limited writing performance) to 3 (a satisfactory 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, the STAAR Modified Resources page, and the TAKS Resources page.

The STAAR English I and II tests include two short answer responses. The STAAR Modified English assessments do not contain short answer responses. The TAKS ELA tests at exit level include three short answer reading responses. The STAAR Modified English assessments do not contain short answer responses. Each short answer response is scored on a scale of 0 to 3. 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

Pearson conducts an extensive search for the best people to score Texas students' tests. Pearson works with the same employment resources used by school systems across the state and advertises broadly for qualified test readers. All test readers hired by Pearson 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 rigorous training, applicants must 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. The training and monitoring of reader performance is conducted by scoring directors and supervisors, all of whom have extensive experience with the Texas assessment programs and often with 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. They are responsible for 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 scorer questions and reviewing scoring reports. Scoring supervisors are trained on both content and job expectations shortly before scorers are trained. If possible, people with previous scoring experience are hired as supervisors. The project monitor supervises all aspects of performance scoring for the Texas assessment program, writes a plan that specifies the configuration of training materials, and manages the schedule and process for performing the work.

The Austin Performance Scoring Center (PSC) oversees scoring of all essays and short answer reading responses for the Texas assessment program. In addition, the PSC collaborates with TEA on the development of writing prompts and the training of scoring supervisors. The PSC 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 a pool of approximately 54,000 screened applicants.
- More teachers across the state are able to participate in the scoring process.
- Distributed scoring is environmentally responsible through the reduction in commuter carbon emissions. Paper handling and associated costs and risks are reduced.
- Scorers 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 scorers.

All Texas assessments use a blend of distributed scoring and regional scoring, except for the STAAR Modified assessments, which are scored regionally by readers at the Austin PSC.

The ePEN System

Written compositions and short answer responses are scored using the electronic Performance Evaluation Network (ePEN) system. The ePEN system enables readers to view the scanned responses on a computer monitor exactly as they were written and select a score for the response using the applicable rubrics.

Student answer documents are scanned after they are received from districts. During the scanning process, the pages on which students wrote compositions or short answer responses are separated from the multiple-choice section of the answer documents. The sections of the answer document are linked by a unique number printed on each page so that the performance-task scores can be added to the student's record once scoring is complete. The performance-task responses are given a unique ePEN identifying number. The ePEN number is not visible to individual readers. As a result of this process, unless students signed their names, wrote about their hometowns, or in some way provided other identifying information, readers have no knowledge of who students are or where they live. The lack of identifying information on the responses helps ensure unbiased scoring.

The responses are grouped by grade or course and are stored on an ePEN 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 supervisors who work on the STAAR and TAKS performance-task scoring projects receive extensive training, including training through onsite and online modules on materials based on the prompts and/or short answer responses related to each assessment. 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 qualifying sets; any reader who cannot demonstrate satisfactory performance on these sets is dismissed. Only readers who undergo the complete training and qualifying process are allowed to begin scoring operational student responses. After the readers are qualified, they are trained to use the ePEN system.

Written Compositions

Readers first complete several training sets of student compositions. The student compositions in the training sets have already been scored by scoring directors and TEA staff. The first set of training materials is selected to clearly differentiate student performance at the different rubric score points and help readers learn the difference between score points. The second set of training materials is selected to be borderline between two adjacent score points and help readers refine their understanding of differences between adjacent score points. The third set of training materials represents all the rubric score points. A scoring director leads the discussions and answers any questions about the training sets. Once readers complete the training sets, they are administered two qualifying sets of student compositions. As with the

training sets, the student compositions in the qualifying sets have already been scored by scoring directors and TEA staff. All the readers take two qualifying sets and must accurately assign scores to 80 percent of the student responses on at least one of the two sets. Any reader unable to meet the standards established by TEA is dismissed.

Short Answer Responses

Before training, the readers are divided into two groups for STAAR or three groups for TAKS based on the number of short answer responses on the assessment. Each group is trained on and scores one of the short answer responses. This allows each group to focus fully on a particular short answer response without being distracted by the other questions. 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 scoring directors and TEA staff. A scoring director leads the discussions and answers any questions about the training sets. Once readers complete the training sets, they are administered two qualifying sets of short answer responses. All the readers must accurately assign scores to 80 percent of the student responses on at least one of the two sets. Any reader unable to meet the standards established by TEA is dismissed.

Ongoing Training

After initial training, ongoing training is provided routinely to ensure scoring consistency and to ensure high reader agreement. 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 score-agreement models 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 and STAAR Modified written compositions are scored using an adjacent agreement scoring model. Each reader assigns a score from 1 to 4 for STAAR and from 1 to 3 for STAAR Modified. The reported (summed) score for STAAR ranges from

2 to 8 and for STAAR Modified ranges from 2 to 6. Summed score performance information is provided to districts on both the Confidential Student Report (CSR) for individual students and on the Written Performance Summary Report for individual campuses and districts. The STAAR 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 ePEN system creates 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 and scoring directors up-to-date information on the progress of the entire project at all scoring centers.

Nonscorable Responses

Before an essay can be given a nonscorable designation, the response is thoroughly reviewed by a scoring supervisor and the scoring director. If either of these reviewers determines that the response is scorable, it is assigned a score and routed to a second reader. If the scoring director agrees that the response is nonscorable, a second scoring director or the project monitor is brought in to conduct a second 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 supervisor, the scoring director, and the project monitor. Readers can also send difficult-to-score responses to their scoring supervisor, who can respond to the reader or pass the question along to the scoring director or project monitor. 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 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 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 system. The ePEN system allows the 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 at an overall rate of one validity response for every 40 responses scored.

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 and also provide a summary of their overall impressions as to the suitability of each prompt or question for future use on an assessment.

Following the scoring of the field-test responses, Pearson 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 Pearson staff independently score samples of the field-test responses to the prompts and short answer responses to be used on the operational assessments. This scoring is in addition to the scoring already done by field-test readers. TEA and Pearson content and management-level staff, including the respective scoring directors, participate in a series of meetings called “rangefinding sessions” to analyze these responses and to assign “true” scores. The scoring directors select responses from the rangefinding sessions to be included in each scoring guide. The scoring directors 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 Pearson and approved by TEA.

Appeals

If a district has questions about the score a response has been assigned, the district can request that the response be rescored. Through a telephone call to the district contact person, Pearson provides an analysis of the response in question to explain the final outcome of the appeal and whether or not the score was changed.