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 assessment instruments that accurately reflect what Texas students have learned in the classroom.

Thousands of Texas educators have served on one or more of the educator committees involved in the development of the Texas assessment program. These committees represent the state geographically, ethnically, by gender, and by type and size of school district. They routinely include educators with knowledge of the needs of all students, including students with disabilities and English language learners (ELLs).

The procedures described in Figure 1 outline the process used to develop a framework for the tests and provide for ongoing development of test items.

Figure 1. Test Development Process

1 Committees of Texas educators review the state-mandated curriculum to develop appropriate assessment objectives for a specific grade and/or subject test. For each subject area, 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 objectives and to determine how these objectives 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 objectives and the 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 objective and, when necessary, are piloted by Texas students from volunteer classrooms.

5 Educator committees assist in developing guidelines for assessing each objective. 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 objective.

*7 Professional item writers, many of whom are former or current Texas educators, develop items based on the objectives 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 composed of Texas educators 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 objective 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 content 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 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 serves a specific function, and their collaborative efforts contribute significantly to the quality of the assessment program.

Student Assessment Division

TEA's Student Assessment Division is responsible for implementing the provisions of state and federal law for the statewide 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 statewide assessment program. Because of the diverse nature of the services required, Pearson employs subcontractors to perform tasks requiring specialized expertise. During the 2010–2011 school year, Pearson's subcontractors for test development activities were Educational Testing Service (ETS) and Tri-Lin Integrated Services, Inc. (Tri-Lin).

ETS

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

Tri-Lin

Tri-Lin Integrated Services, Inc., specializes in translation and transadaptation of assessment items from English into Spanish. As a subcontractor of Pearson, Tri-Lin researches terminology as well as cultural and regional differences to ensure the proper translations of the grades 3–5 mathematics and science items and reading passages for grades 3–5. 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 in Spanish.

Texas Educators

Texas educators, including K–12 classroom teachers, higher education representatives, curriculum specialists, administrators, and ESC staff, play a vital role in all phases of the test development process. When a new assessment is developed, committees of Texas educators review the state-required curriculum, help develop appropriate objectives for the specific grades and/or subject areas tested, and provide advice on a model for assessing the particular subject that aligns closely with the curriculum and good classroom instruction.

Draft objectives 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 objectives 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 that the content and level of difficulty are appropriate 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 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 owned by TEA.

Item Guidelines

Item guidelines are strictly followed by item writers to ensure the accurate measurement of the TEKS student expectations.

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 subject-area knowledge and their teaching or curriculum development experience in the relevant grades. For each subject area and grade, TEA receives an item-tally sheet that displays the number of test items submitted for each objective and TEKS student expectation. Item tallies are examined throughout the review process. If necessary, additional items are written by Pearson or its subcontractors to complete the requisite number of items per objective.

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 objectives and item guidelines as well as discuss the scope of the testing program, security issues, adherence to the measurement specifications, and avoidance of possible economic, regional, cultural, gender, and ethnic bias.

Contractor Review

Experienced staff members from Pearson and its subcontractors, as well as content experts in the grades and subject areas for which the items were 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, as they may impact various demographic groups. Pearson instructs reviewers to consider additional issues, such as the alignment between the items and the objectives, range of difficulty, clarity, accuracy of correct answers, and plausibility of distractors. Pearson also directs its reviewers to consider the more global issues of passage appropriateness, passage difficulty, interactions between items within passages and between passages, and appropriateness of artwork, graphs, or figures. The items are examined by Pearson editorial staff before they are submitted to TEA for review. Items developed for the end-of-course (EOC) subjects are also subjected to expert content review by recognized experts in the subject areas under review.

TEA Review

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. The task during these internal sessions is to scrutinize each item to ensure alignment to a particular portion of the Texas Essential Knowledge and Skills (TEKS), grade-level appropriateness, clarity of wording, content accuracy, plausibility of the distractors, and any potential economic, regional, cultural, gender, and ethnic bias.

Educator Committee Review

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, subject-area specialists in TEA's Curriculum Division, and other agency divisions. Nomination forms are provided to districts and education service centers through TEA's Student Assessment Division website. Educator review committee members are selected based on their established expertise in a particular subject 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 (such as urban, suburban, rural, large, and small districts).

Item-Review Committees

TEA's Student Assessment Division staff, along with Pearson, ETS, and/or Tri-Lin 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 bias against any group.

Table 1 shows the guidelines educator committee members follow to choose items for assessments.

Table 1. Item Review Guidelines

Item Review Guidelines	
Objective/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 level? • 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 level?
Opportunity to Learn	• Is the item a good measure of the curriculum? • Is the item suitable to the grade level?
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 to prepare a field test for a new assessment area 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 may occur before the extensive item-development process described on the preceding pages. If the purpose is to pilot-test administration logistics, the pilot may occur after major item development but before field testing.

Field Testing and Data Review

Before a test item can be used on an operational test form, it must be field-tested.

Sampling Procedures

TEA conducts field tests of all new items either by embedding items in operational tests or by administering separate field-test forms. Whenever possible, field-test items are embedded in multiple forms of operational tests so the field-test items are randomly distributed to students across the state. This ensures that a large representative sample of responses is gathered on each item. Past experience has shown that these procedures yield sufficient data for precise item evaluation and allow 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. The percentage of students responding to each item is included in the item-analysis data presented to the data-review committees.

When separate Spanish-version field tests occur, a sample of students is not sufficient to provide valid data given the smaller population of students involved. Therefore, all students who take the operational administration of the tests are required to participate in the separate field testing.

To examine each item for potential ethnic bias, the sample selection program is designed in such a way 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

- number of students by ethnicity and gender in each sample;
- percentage of students choosing each response;
- 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 subject-area test; and
- various Rasch and Mantel-Haenszel statistical indices to determine the relative difficulty of each test item and to identify greater than expected differences in group performance on any one item by gender and/or ethnicity.

Data-Review Committees

After field testing, TEA and Pearson curriculum and assessment specialists and psychometricians meet to examine each test item with regard to objective/student expectation match, appropriateness, level of difficulty, and bias (economic, regional, cultural, gender, and ethnic) 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 identified and eliminated from use on any test.

Statistical Analyses

Various statistical analyses, including classical measurement theory and item response theory (Rasch model measurement), are used to analyze the field-test data. Analysis includes an examination of the psychometric properties of the tests, the performance of individual test items, and the distributions of test scores at the student, campus, district, and state levels.

For the purpose of reviewing the quality of new test items, reviewers are provided with various data to assist them in decision-making. Three types of differential item functioning (for example, item bias) data are presented during data review: separately calibrated Rasch difficulty comparisons, Mantel-Haenszel Alpha and associated chi-square significance, and response distributions for each analysis group.

The differential Rasch comparisons provide item-difficulty estimates for each analysis group. Under the assumptions of the Rasch model, the item-difficulty value obtained for one group can be different from that of another group only because of variations in some group characteristics and not because of variations in achievement. When the Rasch item-difficulty estimate shows a statistically significant difference between groups, the item is flagged to indicate that further examination of the particular item is needed.

The Mantel-Haenszel Alpha is a log/odds probability indicating when it is more likely for one of the demographic groups to answer a particular item correctly. When this probability is significantly different across the various groups, the item is flagged for further examination.

Response distributions for each analysis group indicate whether members of a group were drawn to one or more of the answer choices for the item. If a large percentage of a particular group selected an answer choice not chosen by other groups, the item is inspected carefully.

However, statistical analyses merely serve to identify test items that have unusual characteristics. They do not specifically identify items that are "biased;" such decisions are made by item reviewers who are knowledgeable about the state's content standards, instructional methodology, and student testing behavior.

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 a paper copy of each test item.

The electronic item bank also stores item data, such as the unique item number (UIN), grade level, subject, objective/TEKS student expectation measured, dates the item was administered, and item statistics. The statistical item bank warehouses information obtained during the data-review committee meetings specifying whether a test item is acceptable for use. TEA uses the item statistics during the test construction process to calculate and adjust for differential test difficulty and to adjust the test for content coverage and balance if needed. The files are also used to review or print individual item statistics.

Test Construction

Each subject-area and grade-level test is based on a specific test blueprint that guides how each test is constructed. Test blueprints delineate the relative emphasis for each objective, as recommended by educator review committees and TEA's curriculum and assessment staff. The tests are designed to

- reflect the range of content and level of difficulty of the skills represented in the TEKS;
- include only those items judged to be free of possible gender, ethnic, and/or cultural bias and deemed acceptable by the educator review committees; and
- reflect problem-solving and complex thinking skills.

TEA constructs tests from the bank of items deemed acceptable after data review. Field-test data are used to place the item difficulty parameters on a common Rasch (one-parameter) logistic 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 within a content objective not only to meet sound content and test construction practices but also to provide objectives of comparable difficulty from year to year.

Tests are constructed to meet a blueprint for the required number of test items for each objective. Items testing each objective are included for every administration, but the array of TEKS student expectations represented may vary from one administration to the next. The tests are constructed to measure a variety of TEKS student expectations and represent the range of content eligible for each objective being assessed.

Panels composed of university-level experts in the fields of English language arts (ELA), mathematics, science, and social studies meet each year to review the content of each of the EOC assessments to be administered. This critical review is referred to as a content validation review and is one of the final activities in a series of quality-control steps to ensure that each high school test is of the highest quality. A content-validation review is considered necessary for high school tests because of the advanced level of content being assessed.

Security

TEA has always placed a high priority on test security and confidentiality in all aspects of the state's assessment program. From the development of test items to the construction of tests, from the distribution and administration of test materials to the delivery of students' score reports, special care is taken to help ensure test security and confidentiality. In addition, TEA investigates every allegation of cheating or breach of confidentiality.

Test Security Supplement

TEA has implemented numerous measures to strengthen test security. It has developed and instituted various administrative procedures to train and support personnel on ensuring test security and confidentiality. The Student Assessment Division developed the *Test Security Supplement* to help guide districts in implementing these requirements and to foster best practices for maintaining a secure testing program.

14-Point Plan

In June 2007, TEA introduced a comprehensive 14-point plan designed to assure parents, students, and the public that test results are meaningful and valid. Maintaining the security and confidentiality of the Texas state assessment program is crucial for ensuring valid test scores and providing standard and equal testing opportunities for all students. The 14-point security plan is available at TEA's Student Assessment website.

Manuals

Test security for the Texas assessment program has been supported by an organized set of test administration documents that provide clear and specific information to testing personnel. In addition to the statutes and administrative rules that are the foundation for test security-related policies and documentation, TEA produces and continually updates the district and campus testing coordinator manuals and test administrator manuals containing detailed information about appropriate test administration procedures. The manuals provide guidelines about how to administer the tests, ensure secure testing environments, and properly store test materials. They also instruct testing personnel about how to report to TEA any confirmed or alleged testing irregularities that may have occurred in the classroom, on campus, or within the school district. Finally, all education personnel with access to secure test materials are required to sign a test security oath for each role they fulfilled during testing. 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.

In addition, administrative products and procedures have been developed to ensure test security on the statewide assessments including the following:

- an internal database that allows TEA to track and report testing irregularities and security violations submitted by districts
- a resolution process that tracks missing secure test materials after each administration, and provides suggested best practices that districts can implement to ensure the proper handling and return of secure materials
- training materials specific to test security and test administration best practices for posting to TEA's Student Assessment Division website

Security Violations

In accordance with the Texas Administrative Code (TAC), 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 may be penalized under 19 TAC, §101.65(e). An educator involved with a testing irregularity may be faced with the following:

- restrictions on the issuance, renewal, or holding of a Texas educator certificate, either indefinitely or for a set term;
- issuance of an inscribed or noninscribed reprimand;
- suspension of a 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 students involved in a violation of test security may be faced with the invalidation of his or her test results.

Light Marks Analysis

Pearson provides an analysis of light marks of all test documents administered in the paper format. Scanning capabilities allows 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, but not always, result from an erasure. Under the assumption that such marks potentially result from an erasure, this information is summarized in the Light Marks Analysis Report.

The Light Marks Analysis Report displays any header group whose average wrong-to-right erasures is greater than three standard deviations above the statewide average for each of the subjects within each grade tested. Each header group represents a testing unit. Districts determine the composition of these header groups by how they complete the "Return Batch Header." Assuming the distribution of the mean wrong-to-right erasures for header groups is normally distributed, fewer than 1 percent of the header groups will be flagged.

Information and descriptive statistics for each flagged header group is available in the report. The information types and what they represent include the following:

- County-District—This six-digit number represents the code for the county and the district number.
- State Summary—This line provides the average number (and standard deviation) of wrong-to-right erasures made on this test statewide.
- Campus—This line provides the three-digit campus number and name of the campus.
- Header Group—This line provides the name of the header group.
- # of Students—This line provides the number of students within the header group.
- All Items—This line provides the average number of total erasures for the students in the group.
- Wrong-to-Right—This line provides the average number (and percentage) of erasures from incorrect to correct answers. This number may be the primary area of interest in the report.
- Right-to-Wrong—This line provides the average number of erasures from correct to incorrect answers.
- Wrong-to-Wrong—This line provides 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 right-to-wrong and wrong-to-wrong erasures.

The Light Marks Analysis Report has two parts. The first part of the report presents the results of header groups ranked by average number of wrong-to-right erasures. The second part of the report, known as the district summary report, presents the same results grouped by county/district code.

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.

A sample Light Marks Analysis Report for a TAKS grade 3 mathematics test is provided in Appendix A. All identifying information has been removed to preserve confidentiality.

Quality-Control Procedures

The Texas assessment program and its data play an important role in decision-making about student performance and in public education accountability. TEA verifies the accuracy of the work completed and the data produced by the testing contractor, Pearson, 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.

Reporting of Test Results

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 the state and federal accountability systems that are 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. Pearson is responsible for scoring the tests, aggregating the results, and printing and shipping the reports to school districts. TEA created and implemented a comprehensive quality-control system (QCS) to verify the accuracy of the data and reports produced by Pearson. The QCS was implemented for every TAKS assessment (both paper and online), including TAKS (Accommodated), TAKS–M, TAKS–Alt, LAT administrations, TELPAS assessments, and EOC assessments.

In addition to the comprehensive QCS developed by TEA, Pearson implemented an internal quality-control system for the reporting of test results that uses a business process known as the Capability Maturity Model (CMM). CMM is a description of the stages through which organizations evolve as they define, implement, measure, control, and improve their software processes. This model provides a guide for selecting process improvement strategies by facilitating the determination of current process capabilities and the identification of the issues most critical to software quality and process improvement. Through CMM, documents are created that assist in the requirement definition, development, testing, and implementation of the software required to support each testing program. Examples of these documents include a customer requirements allocation document, a project schedule, functional specifications, a software development project plan, unit test plans, and verification and validation plans.

Once software requirements have been identified, project schedules, project plans, functional specifications, and design documents are created. From these, unit test plans and system test plans can be determined. A unit test plan is a list of code-unit test cases that is executed and recorded by the software developer. The purpose of the code-unit test process is to ensure that software is developed, maintained, documented, and verified to meet the project requirements for coding and unit testing. As such, the process provides the mechanisms necessary to implement the software requirements and design and provides code-units quality assurance prior to execution of a system test.

After all modules (units) have been tested within a system, the CMM process requires a system test. The system test helps ensure that all the units work together and that outputs from one module match the proper inputs for the next module in the system. The CMM process also uses expected results to verify that all requirements have been met. The system test is performed by a group that is independent of the software development group. This process allows for independent verification and interpretation of the requirements. Once the independent testing group has completed the test and given its approval, the system is moved into production mode. It is ready to process the QCS documents and files supplied by TEA, as described in the following paragraphs.

TEA begins the QCS process months in advance of a test date. For each test administration Pearson and TEA prepare answer documents for thousands of fictitious students who are assigned to a campus in one of three fictitious districts. Pearson grids these students' answer documents (marks the answer choices and student identification information) using detailed instructions provided by TEA. The answer documents represent real-world scenarios of the numerous correct and incorrect ways answer documents are completed by students and districts.

Pearson processes, scores, and prepares reports for these fictitious students using answer keys, editing rules, and formats previously approved by TEA. TEA simultaneously processes the same student-level information and produces its own reports. When TEA receives Pearson's reports for the fictitious students and districts, it compares Pearson's reports with its own reports.

In addition to scores, calculations, and other numerical data printed on the reports, all text, formats, and customized messages are verified. The goal of this part of the quality-control process is to verify that edits are made properly when the document scanner encounters missing or invalid data. Reports are not sent to districts until all discrepancies in the comparative data for the fictitious districts are resolved and the reports generated by TEA and Pearson match. In addition, the verification system allows TEA to monitor the distribution of all test materials, reports, and information letters.

As part of the QCS process for report verification, TEA and Pearson complete the following tasks:

1. Prepare a test design for each test administration. This is a set of specific instructions to Pearson for preparing answer documents for fictitious students.

    - Check the proposed answer document for the upcoming administration for any design changes that might affect the QCS process (for example, new or revised data fields).
    - Determine whether any new policies have been established since the last administration of the test that would affect how answer documents are edited or how scores are reported. Decide how these policies affect the QCS process and whether these new edits should be tested with additional fictitious students.
    - Create a new database of fictitious students. A new test administration will have most of the same students as the previous administration of the same test but with additions or changes necessary to reflect new reporting policies and/or new conditions that should be tested.
    - Prepare a written test design consisting of coding and gridding instructions to Pearson.
    - Send the test design and text file to Pearson according to an approved schedule of processing deadlines created for the particular test administration.

2. Receive scales from Pearson.

    - Pearson sends a table to TEA that shows the scale score corresponding to each achievable raw score point. If a test administration uses pre-equated scales, these true scales will be used to convert the raw scores to derived (or scale) scores and assign a passing status (for TAKS and EOC) or proficiency rating (for TELPAS) for each fictitious student. These tables are verified, approved, and incorporated into computer programs that produce the student and district/campus files and reports.
    - If a test is post-equated, an artificial scale is used initially for processing the fictitious students' answer documents. Because the QCS is a lengthy process, waiting for the true scale to be created before verifying the accuracy of the reporting system would compromise the delivery schedule of reports to districts. For most of the spring tests, there are only 1–3 days between scale approval and sign-off for QCS. Therefore, there is an initial thorough comparison of files and reports (see below) using artificial scales and an additional comparison of reports with scores generated with the post-equated, or true, scales.

3. Create a student-level data file. This file contains the data from the simulation of the processing of answer documents from the fictitious students.

- Verify that "resolved" fields are correct in the database. The resolved fields simulate the changes that would be made in the Pearson editing process if coding or bubbling errors are made on the answer document.
- Export the data from the database as a text file and create a SAS dataset.

4. Receive student-level file from Pearson.
   - Pearson sends a student-level file (text file) to TEA with student names, demographic data, and scores. This file is the result of a procedure that using the instructions in the test design provided by TEA, simulates the completing of answer documents by districts, followed by processing, editing, and scoring of answer documents by Pearson. These data are in the format of the Electronic Individual Student Record File, an optional report available to districts.
   - Create a Statistical Analysis System (SAS) dataset from Pearson's student-level text file.
5. Compare Pearson and TEA files.
   - For each student record, compare each variable in the Pearson student-level data set with the corresponding variable in the TEA student-level data set.
   - Investigate each mismatch, if any, and determine the source and reason for the discrepancy.
   - Make corrections, if necessary, in accordance with established policy and edit rules.
   - Repeat the process by regenerating student-level files, comparing, and resolving discrepancies until the files are identical.
6. Print reports.
   - Pearson prints reports for the three fictitious districts (all standard and optional individual and aggregated reports) and sends them to TEA.
   - A corresponding TEA report is produced for each Pearson report.
   - Reports by TEA and Pearson are generated with independently produced computer programs.
7. Verify reports.
   - Reports are compared to verify that they contain identical information.
   - Any discrepancies found are investigated and corrected.

8. Approve reports.

   - When all the reports for the fictitious districts are verified to be free of error, TEA notifies Pearson that reports can be shipped to school districts when Pearson's quality assurance process is complete.

When all the standard and optional reports for the fictitious districts have been verified by TEA, the system then is available to process the school districts' documents. Before the bulk of districts' documents are allowed to be processed, the first production run process (FPRP) is performed. A small representative sample of documents is readied for processing and is processed through the entire system. The FPRP documents record a sequence of defined activities used in the first run of live production data through Pearson's Operations Department functions that are specified for a program/project. The FPRP serves as the transition point from the planning, development, and testing phases to the delivery phase and from software development and testing to the production environment. The formal process allows all participating functional groups to

- formally accept the readiness of a system for full production,
- confirm receipt and comprehension of processing specifications, and
- confirm the receipt of required production materials for the project.

Once the FPRP is complete and Pearson's Operations and Quality Assurance departments have approved the functions, Pearson completes the scoring and reporting process for the districts. Pearson also ships final reports to the addresses of the fictitious districts on the same schedule as shipments to the districts, and TEA monitors these shipments for timeliness and completeness.

Ongoing Quality Control

Content Validation

Content validation review is an established step in item quality-control procedures. Panels composed of university-level experts in the fields of English language arts, mathematics, science, and social studies are assembled to review the content of the EOC assessments before their administration. This review is necessary for EOC assessments because of the advanced level of content covered. Experts independently review the test forms for each EOC subject area. After a thorough review of each test, committee members discuss all items, noting any issues that were of concern. When necessary, substitute items are reviewed and chosen.

Verification of Equating Activities

As another quality-control step, Pearson and TEA verify all equating activities. Each live test is calibrated and post-equated by a number of people, including two Pearson psychometricians and at least one external psychometrician. An additional Pearson psychometrician acts as a Quality-Control Coordinator and reviews the equating results. Any discrepancies across the results are resolved prior to the generation of the final scoring tables. In addition, the Quality-Control Coordinator compares the current year's post-equated results to those from previous years and conducts additional data quality and reasonableness checks. The TELPAS reading test calibration and post-equating procedures are conducted independently and verified by two Pearson psychometricians with results reviewed by an additional Pearson psychometrician acting as a Quality-Control Coordinator. Field-test items for all testing programs, whether embedded or separately field-tested, also are independently calibrated and equated by two Pearson psychometricians.

Scoring Table Verification Process for Pre-equated Tests

The scoring table verification process for pre-equated tests supports the accuracy of scoring tables prior to any student tests being scored. In this process, scoring tables are pulled from the Pearson scoring system and compared to scoring tables generated through Pearson's test tracking and construction software. If no discrepancies are found, these tables are forwarded to TEA for verification. If these two reviews concur, the tables are approved by TEA and Pearson personnel and used to score student tests. This process differs slightly for English language arts tests, in that the scoring tables are generated by a Pearson psychometrician, verified, and loaded to the scoring system, rather than being pulled from the scoring system and then verified. Equating is discussed in detail in chapter 3.

Performance Assessments

The 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. 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.

Written compositions are evaluated through the holistic scoring process, meaning that the writing sample is considered as a whole. It is evaluated according to pre-established criteria: focus and coherence, organization, depth of development, voice, and control of conventions. These criteria, explained in detail in the written composition scoring rubric, are used to determine the effectiveness of each written response. Each regular assessment response is scored on a scale of one (low) to four (high). Each modified assessment written composition is scored on a scale of one to three. A rating of zero is assigned to compositions that are nonscorable. In addition, all responses that receive a rating of zero or a score of one are evaluated analytically to determine why they are unsuccessful. This information is provided to districts in two forms: analytic designation(s) on the Confidential Student Report (CSR) for individual students and aggregations of analytic designations in the “Analytic Information Summary” section of the Written Performance Summary Report for individual campuses and districts.

The short-answer component is designed to test students’ ability to understand and analyze published pieces of writing. Students must be able to generate clear, reasonable, thoughtful ideas or analyses about some aspect of the published literary and expository selections. In addition, students must be able to support these ideas or analyses with relevant, strongly connected textual evidence. The criteria are clearly explained in the scoring rubrics for short-answer responses.

Scoring Facility

The Pearson Austin Performance Scoring Center (PSC) oversees scoring of all open-ended items for the Texas assessment programs. 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 Pearson scoring site in Iowa City.

The PSC introduced distributed scoring to the Texas assessment program in 2010–2011. Distributed scoring is a system in which holistic scorers can participate in the scoring process from any location, if they qualify and meet strict requirements. Distributed scoring is a secure, Web-based scoring model that incorporates several innovative components and benefits, including the following:

- The group of regional scorers can be augmented by other highly-credentialed readers for a pool of 42,000 screened applicants.
- More teachers are able to participate in the scoring process.
- Distributed scoring is environmentally responsible.
- Paper handling and associated costs and risks are reduced.
- It introduces state-of-the-art, innovative approaches to the scoring program.

Scoring Staff

The Pearson contract with TEA stipulates that TEA must approve all management-level staff at the scoring centers, including the scoring directors for the various projects. All management-level staff have extensive experience with Texas assessment programs and with numerous other large-scale writing assessments. Throughout the scoring process, senior Pearson staff serve as on-site monitors at each of the four scoring centers.

All performance assessments are scored by readers hired by Pearson. Readers are organized into teams. Each team is coordinated by a scoring supervisor under the leadership of a scoring director. Scoring supervisors are chosen from experienced readers and past scoring supervisors. Each project also employs an assistant scoring director, whose duties include assisting the scoring director with various administrative and quality-control activities.

Prompt Development, Field Testing, and the Rangefinding Process

Numerous writing prompts and short-answer items are field tested in early spring, and responses are generated by representative samples of Texas students. The field-test responses are scored each summer. The scoring process is the same as that used for the live responses. Following the scoring of the field-test responses, Pearson staff compile a summary of the performance of each prompt and short-answer item, focusing on such factors as the variety of content seen in the responses, the variety of approaches used, the clarity of the prompt/short-answer item wording, and an overall impression of the suitability of the prompt/short-answer item for possible administration on a live statewide assessment. These summaries, along with the statistical data from the scoring process, are presented to educator review committees for discussion and comment. The field-test responses serve as the basis for assembling training materials once TEA has selected the live prompts and short-answer items for the following school year's assessments.

TEA and Pearson staff independently score samples of the field-test responses to the prompts and short-answer items to be used on the live assessments. This scoring is in addition to the scoring already done by the field-test readers. TEA and Pearson management-level staff, including the respective scoring directors, then met in a series of meetings called rangefinding sessions to analyze these responses and to assign "true" scores. Compositions are assigned both holistic and analytic scores. The scoring directors select responses from the rangefinding sessions to be included in each scoring guide. After TEA approval of these selections, the scoring directors write annotations for the guide responses; all annotations are reviewed and approved by TEA staff. The scoring directors then assign the remaining prescored responses from the rangefinding sessions to practice sets and qualifying sets for use in reader training.

Reader Training Process

All readers and scoring supervisors receive extensive training, including training through online modules, on materials based on the prompts and/or short-answer items related to each assessment. Readers attend focused sessions during which they receive training on the scoring guide for a particular project, score practice 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 live scoring begins.

Holistic 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 from the project. Only readers who undergo the complete training and qualifying process are allowed to begin scoring live student responses.

All responses, both compositions and short-answer responses, are scored based on the "perfect agreement" model, meaning that two readers must assign the same score for the response to be considered resolved. During holistic scoring, responses receive scores from two independent readers. If the two readers assign the same score, the response is given that score and considered resolved. If the two scores do not agree, then a third reader, who is especially experienced, independently scores the response. In almost all cases, the third reader agrees with either the first or the second reader, allowing a final score to be assigned. Occasionally a fourth reading of a response is necessary. When this occurs, the response is given to the scoring director or project monitor for final resolution.

For projects including written compositions, two sets of readers are employed: holistic and analytic. Holistic readers score the compositions, and analytics readers review all of the unsuccessful papers to assign categories that specify each composition's weaknesses. In addition, for the high-stakes exit level written compositions, a special team of readers, the specialists, are trained to provide a score-verification procedure to further evaluate all responses that received a score of 1 (unsuccessful) during the holistic scoring process. This step is taken before the responses are sent to the analytics group. If the score verification specialists determine that a particular response may be higher than a 1, the specialist coordinator also evaluates the test response. If the specialist coordinator agrees, the response then is read by the scoring director. At that point, the score may be changed, or the response may be referred to the project monitor for a final scoring decision. TEA staff may be consulted as well.

TEA representatives are on site at each scoring center during the training of readers. In addition, TEA representatives select validity responses and work with Pearson staff and analytics coordinators in preparation for analytics scoring and specialist scoring. Throughout the scoring project, TEA staff are consulted on "decision papers," which are responses that are highly unusual or require a policy decision from TEA.

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 read the scanned response on a computer monitor and select a score for the response from a menu on the screen. Like the readers who read responses on paper answer documents, the readers who work on an ePEN project read the responses exactly as the students wrote them and make scoring judgments using the applicable rubrics. The readers receive the same training in the application of the relevant scoring criteria whether they work on a paper document project or on an ePEN project. Supervisors and readers were trained using a combined traditional presentation and online approach.

Training Procedures

Scoring Supervisor and Reader Training: Written Compositions

The scoring directors conduct the scoring supervisor and reader training for holistic scoring. To ensure that scoring supervisors are prepared to answer reader questions during and after the training and to ensure that scoring supervisors are highly qualified to perform their roles during the scoring process, scoring supervisor candidates are trained before the readers.

The guidelines for scoring supervisor and reader training are essentially the same. After completing all the training sets, the scoring supervisors take the qualifying sets. Regardless of whether a scoring supervisor scores well enough on Set 1 to qualify, the supervisor also takes Set 2 and Set 3. Taking all the sets is important because scoring supervisors are responsible for working directly with readers and must understand all the qualifying sets. All the readers take qualifying Sets 1 and 2. A reader who does not qualify on one of these sets has the opportunity to take Set 3. Any reader unable to meet the standards established by TEA is dismissed.

Training of the analytics readers for all grades and of the verification specialists at exit level follows a similar pattern, except that the training is performed by the respective coordinators. Although no qualifying sets are used in analytics or specialist training, readers can begin "live" scoring only when they are able to demonstrate accuracy.

Scoring Supervisor and Reader Training: Short-Answer Responses

Before training, the readers are divided into three groups. Each group is trained on and scores one of the short-answer items. This allows each group to focus fully on a particular question without being distracted by the other short-answer items.

As with written composition training, the scoring supervisors are trained before the readers arrive, and the process is essentially the same. The reading selections that appear on the test are read by the trainees, and any questions about the material are answered. The scoring director presents the guide responses. Trainees work through the practice sets, and the scoring director leads the discussions and answers any questions. After the readers are qualified, they are trained to use the ePEN system.

Ongoing Roomwide Training: All Projects

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.

Monitoring of Individual Readers

In addition to the ongoing training, readers are closely monitored by their scoring supervisor, the scoring director, and the project monitor. Readers can also send responses that are difficult to score 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 constant 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 at or above the room average. Any reader who cannot be successfully retrained on the criteria is dismissed.

Validity System

The ePEN system allows the project staff to insert validity responses within the scoring cycle without the readers being aware that what they are scoring is a validity response. The scoring directors and TEA staff must agree on the scores of all validity responses. Proposed validity responses are transferred to a Validity Folder on the ePEN system. TEA staff members have access to these files and approve or reject the proposed responses. Once the responses are approved, they are placed in a validity queue. The validity responses are inserted into the scoring queue at a rate of one validity response for every 40 responses scored.

Nonscorable Responses

During holistic scoring, if a reader believes a response may be nonscorable, the response is sent to a review queue in the ePEN system for the scoring supervisor to review. If the scoring supervisor determines that the response is scorable, he or she scores it and responds to the reader. If the scoring supervisor believes the response to be nonscorable, he or she alerts the scoring director and leaves the response in the review queue. If the scoring director finds the response to be nonscorable, the second reading is performed independently by the other scoring director or by the project monitor. Nonscorable responses then are sent to the analytics queue for evaluation by the analytics readers.

Resolution Procedures

When first received from districts, student answer documents are scanned. During the scanning process, the pages on which students wrote short-answer or composition responses are separated from the multiple-choice section of the answer document. The sections of the answer document are linked by a unique number printed on each page so 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 the students are or where they live. The lack of identifying information on the responses helps ensure unbiased scoring.

The responses are grouped by grade and 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 shunted 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 which response. After a reader has completed a first reading of the response, the response is shunted into a second reader's queue for an independent reading.

Following completion of both the first and second readings, responses needing an additional reading are identified and shunted into a resolution queue. Only readers identified as above the room 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. When this occurs, the fourth readings are placed in a separate queue and scored only by scoring directors or project monitors. Compositions receiving a holistic score of 1 or a score of 0 (nonscorable) are shunted to the analytics queue for evaluation by the analytics group.

Short-answer responses do not go through the analytics process.

Data Entry Procedures and Resulting Reports

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.

Score Reliability and Validity Information

Throughout the years, TEA has reported on the reliability and validity of the performance task scoring process. Reliability has been expressed in terms of reader agreement and correlation between first and second readings. Validity has been assessed via validity packets composed of responses selected and examined by TEA staff.

Reader agreement rate is expressed in terms of absolute agreement (the first reader's score equals the second reader's score). Validity is expressed in terms of perfect agreement between the score assigned by a given reader and the "true" score assigned by TEA.

Student response scores are based on the score that has been agreed upon independently by at least two of three readers. Only a fourth reader, limited to senior scoring staff, can determine the final score when a response has been given discrepant scores by three independent readers.

Field-Test Response Scoring

After all live scoring was completed, small groups of experienced readers were selected to score the responses generated by representative samples of students during field testing. As explained earlier, student performance on field-test prompts and short-answer items provides information that helps determine which prompts and items will be selected for future operational administrations. In addition, field-test responses are the basis for the reader-training materials once a prompt or a short-answer item is used on a live 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 item for possible future use on an assessment.

Appeals

Pearson rescores any response about which questions have been raised regarding the assigned score. If Pearson scoring leadership determines that a score may need to be changed, TEA is consulted before a final decision is made. 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 the score was changed or not.