Chapter 6: Annual Test Development Activities

Overview

Maintaining a student assessment system designed to be of the highest quality requires completing a set of tasks that must be executed at specified times throughout the year. This chapter provides a description of those tasks.

Texas educators—classroom teachers, 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 are taught 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 below outline the process used to develop a framework for the tests and provide for ongoing development of test items. Steps marked with an asterisk are repeated annually to ensure the development of tests of the highest quality.

1. Committees of Texas educators review the state-mandated curriculum to develop appropriate assessment objectives for a specific grade and/or subject-area 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 teachers, curriculum specialists, assessment specialists, and administrators.

3. A draft of the objectives 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 objective and, when necessary, are piloted by Texas students from volunteer classrooms. (See “Pilot Testing” later in this chapter.)

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 teachers, 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 accuracy because of the advanced level of content being assessed.

*16. Tests are administered to Texas students, and 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 State Board of Education (SBOE) uses impact data and statewide opportunity-to-learn information, along with

recommendations from standard-setting panels, to set a passing standard for a new state assessment.

*20. A technical digest is developed annually to provide verified technical information about the tests to schools and the public.

Item Development and Review

This section describes the item writing process used during the development of Texas Assessment of Knowledge and Skills (TAKS), Texas English Language Proficiency Assessment System (TELPAS), and end-of-course (EOC) test 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 developed for TAKS, TELPAS, and EOC assessments are strictly followed by item writers to ensure the accurate measurement of the Texas Essential Knowledge and Skills (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 test 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 high school grade levels are also subjected to expert content review. The individuals conducting these reviews are 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 committee item-review meeting. The task during these internal sessions is to scrutinize each item for match to the objective and underlying student expectation, item appropriateness for the grade level being assessed, clarity of wording, content accuracy, plausibility of the distractors, and any potential economic, regional, cultural, gender, and ethnic bias.

Educator Committee Review

During the 2007–2008 school year, as has been done since statewide assessment began in Texas in 1980, TEA's Student Assessment Division convened committees composed of teachers, curriculum directors, principals, other district professionals, and administrators from regional ESCs to work with TEA staff in reviewing 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 by TEA's Student Assessment Division and can be found on the TEA website. In partnership with TEA, Pearson builds the educator review committees and selects committee members 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).

Texas educator committees were convened to review all newly developed test items and all new field-test data for the TAKS, TAKS–Modified (TAKS–M), TAKS–Alternate (TAKS–Alt), TELPAS, and EOC assessments. Item review and data review meetings were held in Austin between August 1, 2007, and July 31, 2008. The composition of these committees is shown in Table 2 on the following page.

Table 2. Texas Educator Review Committees' Demographic Data*

		Number	Percentage
Gender	Female	1,427	80
	Male	363	20
	Total	**1,790**	**100**
Ethnicity	African American	149	8
	Hispanic	626	35
	White	973	54
	Other	42	3
	Total	**1,790**	**100**

*The demographic data presented in the table include information about attendees at the 2007 item review committee meetings and attendees at the 2008 data review committee meetings.

Item-Review Committees

TEA's Student Assessment Division staff, along with Pearson, Educational Testing Service (ETS), and/or Questar Assessment, Inc. 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. If the committee finds an item to be inappropriate after review and revision, the item is removed from consideration for field testing.

After the educator committee meetings, Pearson provides TEA with a summary for each subject and grade reviewed that includes a tally of the number of items recommended for either retention or rejection by committee members.

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. In 2007–2008, pilot testing was not conducted.

Field Testing and Data Review

Before a test item can be used on an operational test form, it must be field-tested. Field testing was conducted during the 2007–2008 school year for the TAKS, TAKS–M, TELPAS, and EOC assessments.

Sampling Procedures

The statewide assessment program conducts field tests of all new items by either 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 that 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 actual tests. The percentage of students responding to each item is included in the item-analysis data presented to the data review committees.

By contrast, TAKS field tests for grades 4 and 7 writing, grade 9 reading, and grade 10 and exit level English language arts must be separately administered because embedding test items in a live test form is not possible due to the structure of the tests and the performance tasks (compositions and/or open-ended reading responses). Instead, these field tests are conducted with a sample of students from across the state wherein each student is administered a separate field-test booklet. Spanish-version TAKS field tests are also separately administered, but given the small population of students involved, a sample of students is not sufficient to provide valid data; therefore, all students who take the live administrations of these 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. Districts are notified at the beginning of the school year about which campuses and classes are chosen for the administration of each test form so any issues related to sampling or to the distribution of materials can be resolved before the test materials arrive. TEA field-tests only items that are deemed acceptable after committee review. 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 an item by gender and ethnicity.

Data Review Committees

After field testing, TEA convenes data review committees composed of Texas teachers, TEA curriculum and assessment specialists, and principals. Much effort is made to ensure that these committees of Texas educators represent the state demographically, with regard to ethnicity, gender, type and size of district, and geographical region. The committees receive training on how to interpret the psychometric data that TEA compiles for each field-test item. Pearson and its subcontractors supply psychometricians, content experts (usually former teachers, content specialists, and item writers), and group facilitators for the data review committee meetings.

A comprehensive training video that explains the review process and serves as an introduction to the statistical analysis is presented to each data review committee. Specific directions regarding the use of the statistical information and review booklets are also provided. Committee members examine each test item with regard to objective/student expectation match, appropriateness, level of difficulty, and bias (economic, regional, cultural, gender, and ethnic) and 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 precluded 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. Pearson provides an array of statistical analyses useful in understanding the psychometric properties of the tests, the performance of individual test items, and the distributions of test scores at the student, school, district, and state levels.

For the purpose of reviewing the quality of new test items, data review committees are provided with data to assist them in decision-making. Appendix 12 in the *2005–2006 Technical Digest* contains a sample of the overview given to committee members about the types of field-test data they review to determine the quality of each item. This digest can be found online at http://www.tea.state.tx.us/student.assessment/resources/techdigest/. Three types of differential item functioning (for example, item bias) data are presented during committee 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 characteristic 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 warranted.

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 than another group. 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 should be inspected carefully.

It is important to understand that 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. This system allows test items to be readily available to TEA for test construction and reference and to Pearson personnel for test booklet production and printing.

The electronic item bank also stores item data, such as the unique item number, 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 and Pearson use 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 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 pool of items deemed acceptable based on recommendations made by educators attending the data review meetings. 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 pool. Hence, 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 the blueprint for the required number of test items for each test 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 ELA, mathematics, science, and social studies meet each year in Austin to review the content of each of the high school level TAKS and EOC assessments to be administered that year. 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 at the tested high school grades (9, 10, and 11) because of the advanced level of content being assessed.