

CHAPTER 5: ANNUAL TEST DEVELOPMENT ACTIVITIES

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

Overview of the Test Development Process

Texas educators—classroom teachers, curriculum specialists, administrators, and education service center 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 state assessments. These committees represent the state geographically, ethnically, by gender, and by type and size of school district and routinely include educators with knowledge of the needs of all students, including students with disabilities and students with limited English proficiency. The procedures described below outline the steps used to develop a framework for the tests and provide for ongoing development of test items.

1. Committees of Texas educators review the state-mandated curriculum to develop appropriate assessment objectives for a specific grade and/or subject test. Educators provide advice on a model or structure for assessing the particular subject that aligns with good classroom instruction.
2. Educator committees work with TEA to prepare draft test objectives, which are distributed widely for review by teachers, 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.
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 is developed that establishes the length of the test and the number of test items measuring each objective.
 - *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, 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.
 - *17. Stringent quality control measures are applied to all stages of printing, scanning, scoring, and reporting.
 - 18. Texas Assessment of Knowledge and Skills (TAKS), Reading Proficiency Tests In English (RPTE), and State-Developed Alternative Assessment II (SDAA II) tests are released to the public in accordance with state law.
 - 19. The State Board of Education uses impact data and the statewide opportunity-to-learn study, along with additional information, to set a passing standard for each new test.
 - *20. A technical digest that provides verified technical information about the tests to schools and the public is developed.
- *These steps are repeated annually to ensure that tests of the highest quality are developed.

Item Development and Review

This section describes the item writing process used during the development of all TAKS, SDAA II, RPTE, and EOC test items. In 2005–2006, Algebra I End-of-Course (EOC) item development as well as TAKS–Alternate (TAKS–Alt) development were begun. While Pearson Educational Measurement (PEM) assumes the major role for item development, many subcontractors and agency personnel are involved in the item development process. Items remain the property of TEA.

Item Guidelines

Item guidelines developed for TAKS, SDAA II, RPTE, and EOC are strictly followed by item writers to ensure the accurate measurement of the TEKS student expectations.

Item Writers

Pearson Educational Measurement and its subcontractors employ item writers who have extensive experience developing items for standardized achievement tests and large-scale criterion-referenced measurements. PEM and its subcontractors select item writers for their specific subject-area knowledge and for their teaching or curriculum development experience for 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 PEM or its subcontractors to complete the requisite number of items per objective.

Training

Pearson Educational Measurement and its subcontractors provide extensive training for each item writer prior to item development. During these training seminars PEM 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 PEM 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 fairness of the items as they may impact various demographic groups. PEM also instructs reviewers to consider such additional issues as the alignment match between the items and the test objectives, range of difficulty, clarity, accuracy of correct answers, and plausibility of distractors. PEM 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 then examined by PEM editorial staff before they are submitted to TEA for review.

TEA Review

Staff from TEA and PEM/subcontractor personnel 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 content-to-specification match, item appropriateness for the grade level being assessed, clarity of wording, plausibility of the distractors, and any potential economic, regional, cultural, gender, and ethnic bias.

Educator Committee Review

During the 2005–2006 school year, as it has done since statewide assessment began in Texas in 1980, the TEA Student Assessment Division convened committees composed of teachers, curriculum directors, principals and other district professionals, and administrators from regional education service centers 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, education service center executive directors and staff members, subject-area specialists in TEA's Division of Curriculum, and other agency divisions. Nomination forms are provided to districts and education service centers by the Student Assessment Division and are available on the TEA website. With review by TEA, PEM builds the educator review committees and selects committee members based on their established expertise in a particular subject area. Committee members represent the 20 education service center 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 during 2005–2006 to review all newly developed test items and all new field-test data for the TAKS, SDAA II, RPTE, and EOC tests. In addition, educator committees were convened to review the newly developed materials for the new special education assessment, TAKS–Alt. A total of 73 item-review and 68 data-review meetings were held in Austin between August 1, 2005, and July 31, 2006. Appendix 11 provides a listing of the meetings held during the 2005–2006 school year. Appendix 12 provides samples of materials produced for these meetings. The composition of these committees is shown in the tables on the following page.

Table 6. Texas Educator Review Committees' Demographic Data*

		Number	Percent
Gender	Female	1,747	79%
	Male	469	21%
	Total	2,216	100%

		Number	Percent
Ethnicity	African American	240	11%
	Hispanic	800	36%
	White	1,107	50%
	Other	69	3%
	Total	2,216	100%

* The demographic data presented in the tables include information about attendees at the 2005 item review committee meetings and attendees at the 2006 data review committee meetings.

Item-Review Committees

The Texas Education Agency Student Assessment Division staff, along with PEM, Educational Testing Service and/or BETA staff, train committee members on the proper procedures and the criteria for reviewing newly developed items. Committees are composed of Texas teachers, TEA curriculum and assessment specialists, principals, and superintendents. Committee members judge each item for appropriateness, adequacy of student preparation, and any potential bias. Appendix 12 contains a sample of the item judgment form used by committee members to review items. Before items are field-tested, committee members discuss each test item and recommend whether the item should be field-tested as written, revised, 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 still finds an item to be inappropriate after reviewing and revising it, the item is removed from consideration for field testing.

After the educator committee meetings, PEM 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 all 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 in order to prepare a field test for a new assessment area and to refine measurement specifications as needed. 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 2005–2006 two pilot tests were conducted: the grade 2 online pilot for the second edition of the Reading Proficiency Tests in English (RPTE II) and the TAKS grade 8 science innovative item online pilot test.

Field Testing and Data Review

Before a test item can be used on a live test form, it must be field-tested. Field testing was conducted during the 2005–2006 school year for the TAKS, SDAA II, RPTE, and EOC tests. Appendix 13 contains the field-test schedule for the 2005–2006 school year.

Sampling Procedures

TEA uses two approaches to administer field-test items to samples of students: embedded items and separate field-test forms. Whenever possible, TEA embeds field-test items in multiple forms of live 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 TEA to collect statistical data on a large number of field-test items in an authentic 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 listed among the item-analysis data presented to the data-review committees.

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 (open-ended responses and/or compositions) required; these field tests are conducted with a sample of students from across the state. The SDAA II field tests and certain TAKS field tests (Spanish version) 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, PEM designs the sample selection program in such a way that the proportions of African American and Hispanic students in the samples are representative of their total student populations in Texas. Districts are notified of which campuses and classes are chosen for the administration of each test form so that 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, principals, and superintendents. 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 PEM compiles for each field-test item. Pearson Educational Measurement and its subcontractors supply psychometricians, content experts (usually former teachers 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 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 so noted and are precluded from use on any test form.

Statistics Used

Pearson Educational Measurement uses various statistical analyses, including classical measurement theory and item response theory (Rasch model measurement), to obtain field-test data. These data are representative of the student population in Texas and are of sufficient quantity (currently, responses to the majority of field-test items are obtained from thousands of students).

Appendix 12 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. 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, that 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.

Item Bank

Pearson Educational Measurement maintains a computerized item bank for the Texas Student Assessment Program's tests. The item bank stores each test item and its accompanying artwork. In addition, TEA and PEM 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 PEM personnel for test booklet production and printing.

Pearson Educational Measurement maintains a second computerized item bank that 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 also warehouses information obtained during the data-review committee meetings regarding whether a test item is acceptable for use or unacceptable. TEA and PEM use the item statistics during the test construction process to calculate and adjust for differential test difficulty and to check and adjust the test for content coverage and balance. 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 the agency's curriculum staff. TEA constructs the tests to

- represent 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 test forms from the pool of items deemed eligible for testing by the educator committees that participated in 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 specifications 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. However, the tests are constructed to measure a variety of TEKS student expectations and are representative of the range of content eligible for each objective being assessed.

In addition to annual educator reviews, item reviews, and data reviews, panels composed of university-level experts in the fields of English language arts, mathematics, science, and social studies meet each year in Austin to review the content of each of the high school level TAKS 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 possible. A content validation review is considered necessary at the high school grades (9, 10, and 11) because of the advanced level of content being assessed.

Webb Alignment Analyses

Alignment is central to the validity of the TAKS testing program. Demonstrating that every item on TAKS can be matched to a skill required by the Texas Essential Knowledge and Skills (TEKS) curriculum standards is not enough to ensure strong alignment. An evaluation of alignment should also address the degree to which an assessment reflects the full range and balance of the curriculum standards as well as the degree of cognitive complexity of the standards.

To examine the alignment of TAKS to the grade-level TEKS standards, the Texas Education Agency contracted with Dr. Norman L. Webb to conduct an independent alignment study. Dr. Webb's alignment process, known as a Webb alignment, provides a set of procedures states can use to conduct an in-depth analysis of the alignment between their state curriculum standards and state assessments. In March 2006, a Webb alignment institute was conducted for the following grades and subjects:

- Mathematics, grades 3–8 and 10
- Spanish Mathematics, grades 3–6
- Reading/English Language Arts, grades 3–8 and 10
- Spanish Reading, grades 3–6
- Science, grades 5, 8, and 10
- Spanish Science, grade 5

Dr. Webb’s alignment study addressed four specific alignment criteria:

- Categorical Concurrence — the extent to which the same or consistent categories of content appear in the standards (TEKS) and assessments (TAKS)
- Depth-of-Knowledge Consistency — the cognitive complexity required by the standards and the assessments, with four levels possible
- Range of Knowledge Correspondence — the breadth or span of knowledge required by the assessments
- Balance of Representation — the degree to which the objectives that fall under a specific standard in the curriculum are given relatively equal emphasis on the assessment

As part of the alignment institute, reviewers were trained to identify the depth of knowledge of the skills and assessment items. This training included reviewing the definitions of the four depth-of-knowledge levels and then reviewing examples of each. Then the reviewers participated in 1) a consensus process to determine the depth-of-knowledge levels of the skills and 2) individual analyses of the assessment items. Following individual analyses of the items, reviewers participated in a debriefing discussion in which they reviewed the degree to which they had coded particular items or types of content to the skills.

To derive the results from the analyses, the reviewers’ responses were averaged. Any variance among reviewers was considered legitimate, with the true depth-of-knowledge level for the item falling somewhere in between the two or more assigned values.

Implications and Recommendations of the Alignment Studies

While the Categorical Concurrence, Range of Knowledge Correspondence, and Balance of Representation were found to be acceptable for all assessments, the Webb alignment study suggested that the Depth-of-Knowledge Consistency of the tests could be enhanced with minor changes. Since the current item development process for TAKS does not systematically capture depth-of-knowledge characteristics of test items, this element will be added.

In addition, the cognitive complexity level of the TEKS needs continued study. Evaluating the depth of knowledge of the TEKS expectations will directly address the degree of cognitive complexity inherent in the content standards. This evaluation will also allow the TAKS content specialists to focus on an attribute of the curriculum that has, as yet, not been formally considered as they review the items for “match to the standards.” At the time, depth of knowledge levels should be added to the TAKS objectives and item development process, the cognitive complexity of items is considered. Strengthening the depth of knowledge match between the TEKS and the TAKS will enhance the public’s understanding of the level of performance students need to demonstrate to be academically successful in the classroom and on the test.

Further information about the Webb alignment study, including the DOK levels assigned to the standards and individual test items, can be found in the *Texas Alignment Study Report* in Appendix 14.