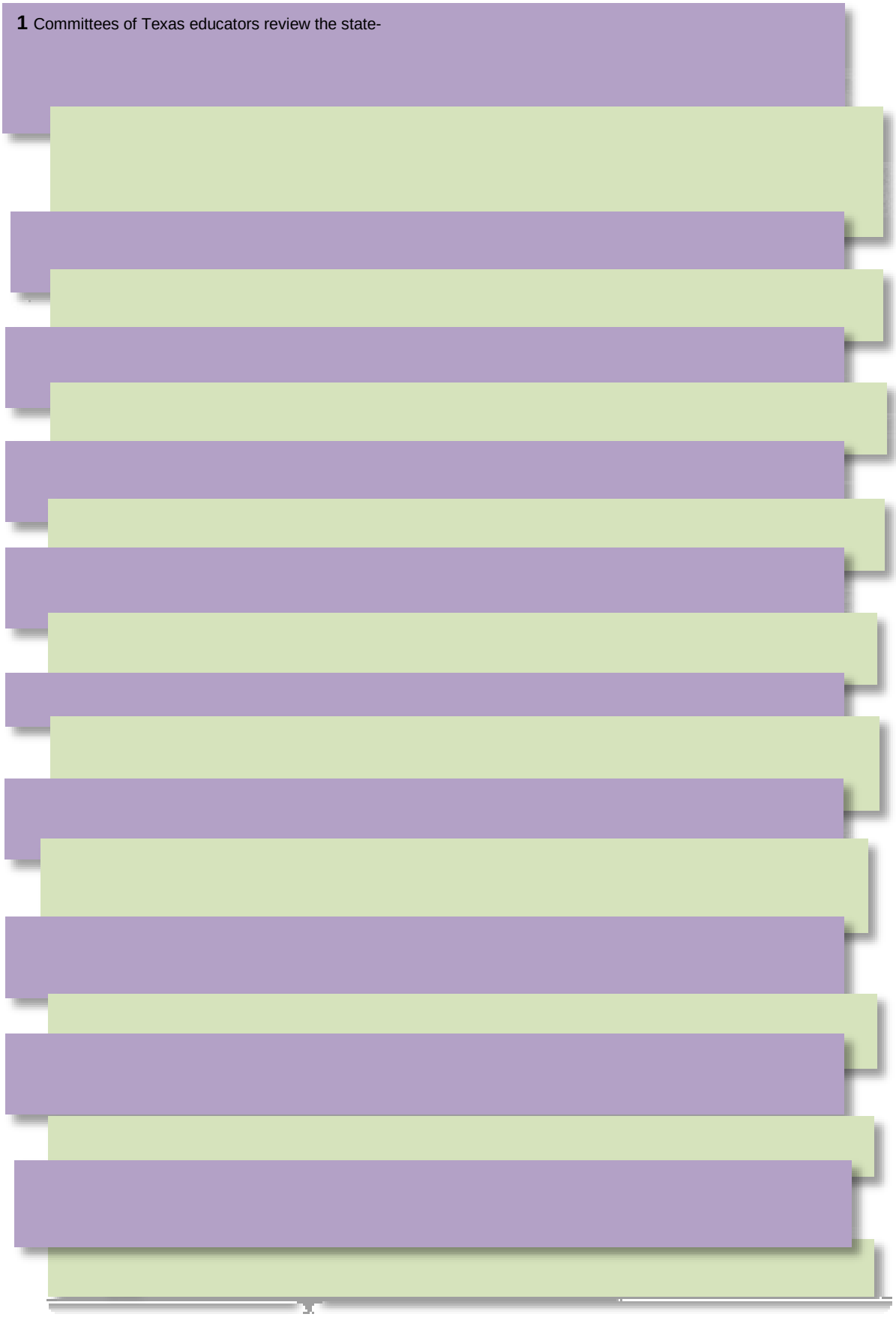
Figure 2.1. Test-Development Process



Item Writers

Pearson and its subcontractors employ item writers who have extensive experience developing items for standardized achievement tests, large-scale criterion-referenced measurements, and English language proficiency tests. These individuals are selected based on their content-area knowledge, their teaching or curriculum development experience in the relevant grades, or their experience teaching students with special needs and English learners.

For each STAAR, STAAR Alternate 2, and TELPAS assessment, TEA receives an item inventory that indicates the number of test items to be developed for each reporting category and TEKS student expectation (for STAAR and STAAR Alternate 2 assessments) or ELPS student expectation (for TELPAS assessments). Item inventories are used throughout the item-review process. If necessary, additional items are developed by the vendor to provide the requisite number of items per student expectation.

For the TELPAS Alternate assessment, the observable behaviors were developed by Texas educators during a series of TEA-led meetings. The educators were guided by Pearson and TEA staff to develop an inventory of items that aligned to the ELPS and covered the alternate Proficiency Level Descriptors (PLDs).

Training

Pearson provides extensive training for item writers prior to item development. During these trainings, Pearson reviews in detail the content expectations and item guidelines and discusses the scope of the testing program; security issues; adherence to the measurement specifications; and avoidance of possible economic, regional, cultural, gender, or ethnic

Table 2.1. Item-Review Guidelines

Passage and Item-Review Guidelines	
Reporting Category/Student Expectation Item Match	• The item measures what it is supposed to assess. • The item poses a clearly defined problem or task.
Appropriateness (Interest Level)	• The item or passage is well written and clear. • The point of view is relevant to students taking the test. • vd cobl opreve

Pilot Testing

The purpose of pilot testing is to gather information about test-item prototypes and administration logistics for a new assessment and to refine item-development guidelines as needed. Pilot testing can be conducted to accomplish varying objectives. If the purpose is to pilot items of differing types and ranges of difficulty, piloting might occur before the extensive item-development process described on the preceding pages. If the purpose is to pilot test administration logistics, the pilot might occur after major item development but before field testing. There was no pilot testing in the 2020–2021 school year.

Field Testing and Data Review

Field testing is conducted prior to a test item being used on an operational test form. However, when there are curriculum changes, newly developed items that have not been field-tested may be used on an operational test form. This is referred to as operational field testing, which was seldom used on the Texas Assessment Program.

Field-Test Procedures

Whenever possible, TEA conducts field tests of new items by embedding them in multiple forms of operational tests so that the field-test items are randomly distributed to students across the state. This results in a large representative sample of responses gathered on each item. Periodically, TEA conducts standalone field tests of new items (e.g., writing prompts) by administering them to a purposefully selected representative Texas student sample. Experience has shown that embedded field testing yields

To ensure that each item is examined for potential ethnic bias, the sample selection is designed so that the proportions of African American and Hispanic students in the samples are representative of their respective total student populations in Texas. Data obtained from the field test include

- the number of students by ethnicity and gender in each sample;
- the percentage of students choosing each response;
- the percentage of students, by gender and by ethnicity, choosing each response;
- point-biserial correlations to determine the relationship between a correct response on a particular test item and the score obtained on the total content-area test;
- Rasch statistical indices to determine the relative difficulty of each test item; and
- Mantel-Haenszel statistics for dichotomous items and standardized mean difference (SMD) for Constructed Response (CR) items to identify greater-than-expected differences in group performance on any single item by gender and ethnicity.

Data-Review Procedures

After field testing, TEA curriculum and assessment specialists provide feedback to Pearson on each test item and its associated data regarding reporting category and student expectation match; appropriateness; level of difficulty; and potential gender, ethnic, or other bias; and then recommend acceptance or rejection of each field-test item. Items that pass all stages of development—item review, field testing, and data review—are placed in the item bank and become eligible for use on future test forms. Rejected items are marked as such and eliminated from consideration for use on any summative assessment.

Item Bank

ETS and Pearson each maintain an electronic item bank for their respective portion of the assessment program. The item banks store each test item and its accompanying artwork.

Each electronic item bank also stores item data, such as the unique item number (UIN), grade or course, subject, reporting category, TEKS or ELPS student expectation measured, dates the item was administered, and item statistics. Each item bank also warehouses information obtained during data-review meetings, which specifies whether a test item is acceptable for use. TEA, ETS, and Pearson use the item statistics and other information about items during the test-construction process to maintain constant test

other items in the bank. Consequently, items are selected not only to meet sound content and test-construction practices but also to ensure that tests are approximately comparable in difficulty from administration to administration. Refer to chapter 3, "Standard Technical Processes," for detailed information about Rasch scaling.

Tests are constructed to meet a blueprint for the required number of items o6 (r)-6 (ed num)-5.9 (be)10.6 o

Security

TEA places a high priority on test security and confidentiality for all aspects of the statewide assessment program. From the development of test items to the construction of tests, and from the distribution and administration of test materials to the delivery of students' score reports, special care is taken to promote test security and confidentiality. TEA ensures that every allegation of cheating or breach of confidentiality are properly investigated.

Maintaining the security and confidentiality of the Texas Assessment Program is critical for ensuring valid test scores and providing standardized and comparable testing opportunities for all students. TEA has implemented numerous measures to strengthen test security and confidentiality, including the development of various administrative procedures and manuals to train and support district testing personnel.

Test Administration Manuals

Test security for the Texas Assessment Program has been supported by an aligned set of test administration documents that provide clear and specific information to testing personnel. In response to the statutes and administrative rules that are the foundation for policies and documentation pertaining to test security, TEA produces and updates detailed information about appropriate test administration procedures in the test administrator manuals.

MANUALS

The Coordinator Resources and test administrator manuals, including the TELPAS Rater Manual, provide guidelines on how to train testing personnel, administer tests, create secure testing environments, and properly store test materials. They also instruct testing personnel on how to report to TEA any confirmed or alleged testing irregularities that might have occurred in a classroom, on a campus, or within a school district. Finally, the manuals provide training and guidelines relative to test security oaths that all personnel with access to secure test materials are required to sign. The manuals give specific details about the possible penalties for violating test procedures. In addition, TAC §101.3031 includes specific language detailing the requirements of school districts and charter schools to maintain security and confiiha (l)2.6 (i)2.6 (t)-6.6 (s)-2 ( )1

T

Erasure information and descriptive statistics for each group (usually by grade level in

T

Performance Assessments

The STAAR and TELPAS tests include constructed-response items, which require scoring by trained human raters on the following operational assessments:

- STAAR grade 4 and 7 writing
- STAAR Spanish grade 4 writing
- STAAR English I, English II, and English III
- TELPAS grades 2–12 speaking

The Texas Assessment Program uses written compositions on STAAR and STAAR Spanish, which are a direct measure of the student's ability to synthesize the component skills of writing; that is, the composition task requires the student to express ideas effectively in writing for a specified purpose. To do this, the student must be able to respond in a focused and coherent manner to a specific prompt while organizing ideas clearly, generating and developing thoughts in a way that allows the reader to thoroughly understand what the writer is attempting to communicate, and maintaining a consistent control of the conventions of written language.

For the STAAR and STAAR Spanish assessments, the types of writing required vary by grade and course and represent the learning progression evident in the TEKS.

Written compositions for STAAR are evaluated using a holistic scoring process, meaning that the essay is considered as a whole. It is evaluated according to pre-established criteria: organization/progression, development of ideas, and use of language/conventions. These criteria, explained in detail in the writing scoring rubrics for each grade and type of writing, are used to determine the effectiveness of each written response. Each essay is scored on a scale of 1 (a very limited writing performance) to 4 (an accomplished writing performance). A rating of 0 is assigned to

by demonstrating a high level of mastery before any student responses are scored. Pearson’s content supervisors 0598 0 Td( )Tj-0.005 Tc -0.005 Tw 0.52 0 Td[(s.5 t(a)70.5 ffh)]TJ 0 Tc 0

Raters first complete a rubric overview training before training on item specific responses. The Item Specific Training (IST) modules are a set of student compositions that have already been scored by assessment specialists and TEA staff. The training materials are selected to clearly differentiate student performance at the different rubric score points and to help raters learn the difference between score points. The training materials also contain responses determined to be borderline between two adjacent score points to help raters refine their understanding of the rubric and the [(to)-1 (w)-1 (o)] TJ 0 -1.207 Td [(adjacen

N

the remaining pre-scored responses from the rangefinding sessions to training sets and qualifying sets for use in future rater training. Educators assist in the review and make recommendations to reach a consensus on the scores. Prior to scoring, TEA staff review and approve all scoring guides and training sets.

Score Reliability and Validity Information

Throughout the years, TEA has reported on the reliability and validity of the performance-scoring process. Reliability has been expressed in terms of rater agreement (percentage of exact agreement between rater scores) and correlation between first and second ratings. Validity has been assessed by the inclusion of validity responses throughout the operational-scoring process. It is expressed in terms of exact agreement between the score assigned by the rater and the “true” score assigned by ETS and approved by TEA.

Appeals

If a district has questions about the score assigned to a response, a rescore can be requested through submission of the appropriate request form. ETS provides rescore results by posting an updated STAAR Report Card (SRC) to the STAAR Assessment Management System, only if the score has changed. If the score does not change, there is a fee that districts pay. If the score changes, that fee is waived. If a district files a formal appeal with TEA related to scores reported on the consolidated accountability file, an analysis of the response in question that explains the final outcome of the appeal, and whether the rater.