

CHAPTER 17: PERFORMANCE ASSESSMENT— TAKS AND SDAA II

TAKS

The TAKS writing tests at Grade 4 (both the English and Spanish versions) and Grade 7, as well as the English language arts (ELA) tests at Grade 10 and exit level, include a written composition component. In addition, short-answer questions are included on the Grade 9 reading test and the ELA tests at Grade 10 and exit level. These performance tasks must be read and evaluated by carefully trained teams of readers as part of the TAKS scoring process.

The TAKS written composition is 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 writer 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 say, and maintaining a consistent control of written language.

A process called focused holistic scoring is used to assess TAKS written compositions. The scoring system is holistic in that the piece of writing is considered as a whole. It is focused in that the piece of writing 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 TAKS response is scored on a scale of 1 (low) to 4 (high). A rating of 0 is assigned to compositions that are nonscorable. In addition, all responses that receive a rating of 0 or a score of 1 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 for individual students and aggregations of analytic designations on the Written Composition Analytic Information Summary Report for individual campuses and districts.

The short-answer component of the Grade 9 reading test and the Grade 10 and exit level ELA tests 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 or 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.

SDAA II

SDAA II writing assessments are administered to students enrolled in Grades 4, 7 and 10. These assessments are available at Instructional Levels K/1, 2, 3/4, 5, 6/7, 8/9, and 10. In addition, short-answer questions are included on an Instructional Level 9 reading test and 10 ELA test. At Instructional Level K/1, students respond to five specific performance tasks: writing numbers, writing their name, writing letters, labeling pictures of common objects, and writing a composition. Together, the five performance tasks are designed to provide a snapshot of the degree to which each student has developed the foundational skills of written English. At Instructional Levels 2, 3/4, 5, 6/7, 8/9, and 10, students write a composition.

The short-answer component of the Instructional Level 9 reading test and the Instructional Level 10 ELA test is designed to assess 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 or 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.

For details on the rubrics used to evaluate the various tasks, refer to the SDAA II released test scoring guides for each of the instructional levels. The scoring guides are available on the TEA Student Assessment Division website at <http://www.tea.state.tx.us/student.assessment>.

Exit Level TAAS

Four times during the 2004–2005 school year, the exit level TAAS writing test was offered to those students for whom TAAS is their graduation testing requirement. The exit level TAAS writing test includes a written composition, which is scored based on the TAAS rubric. Released TAAS composition scoring guides from previous years include those rubrics and are available on the Student Assessment Division website.

Scoring Facilities

Because of the staffing and facility requirements necessary to score the large number of responses submitted for the TAKS, SDAA II, and exit level TAAS performance assessments, several scoring centers must be used. During 2004–2005, exit level TAKS ELA compositions and short-answer responses, Grade 4 TAKS compositions, SDAA II responses at all instructional levels, and exit level TAAS compositions were scored at PEM's Austin, Texas, facility; Grade 4 TAKS English compositions, Grade 7 TAKS compositions, and Grade 9 TAKS short-answer reading responses were scored at Pearson's Dallas, Texas, facility; and Grade 10 TAKS ELA written compositions and short-answer reading responses were scored at Pearson's facility in Albuquerque, New Mexico.

Scoring Staff

The Pearson Educational Measurement contract with TEA stipulates that all management-level staff at the scoring centers, including the scoring directors for the various projects, be approved by TEA. All management-level staff have extensive experience with the TAKS, SDAA II, and/or TAAS programs and with numerous other large-scale writing assessments. Throughout the scoring process, senior PEM staff serve as on-site monitors at each of the three scoring centers.

All performance assessments are scored by readers hired by PEM. These readers are organized into teams of approximately 10 to 12 readers per team for the larger projects. Each team is coordinated by scoring supervisors under the leadership of a scoring director. Scoring supervisors are chosen from experienced readers and past scoring supervisors. This selection process is based on the candidates' understanding of the scoring rubrics (i.e., the scoring criteria), their ability to apply the criteria accurately and consistently, their ability to articulate the criteria, and their demonstration of leadership skills. Each project also employs a "floating" scoring supervisor, whose duties include assisting the scoring director with various administrative and quality-control activities.

In order to be considered for employment as a reader, applicants must have at least a bachelor's degree, preferably in English, education, or a related field; teaching experience is preferred. In addition, during the interview process all applicants are required to write an essay and complete a proofreading exercise. Those applicants interested in scoring Grade 4 Spanish written compositions are required to complete a Spanish decoding/translation exercise.

In spring 2005 1,958 people were involved in the scoring projects for TAKS and SDAA II: 802 at the Austin Scoring Center, 783 at the Dallas Scoring Center, and 373 at the Albuquerque Scoring Center.

Prompt Development, Field-Testing, and the Rangefinding Process

Each year PEM management-level staff, in collaboration with TEA staff, develop numerous TAKS and SDAA II writing prompts, as well as TAKS short-answer items, for field-testing. Field-testing takes place in early spring, and responses are generated by representative samples of Texas students. The field-test responses are scored each summer after all live responses for that year have been scored; the scoring process is the same as that used for the live responses. Following the scoring of the field-test responses, PEM 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, variety of approaches used, 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 each summer for their discussion and comment. The field-test responses serve as the basis for

assembling training materials once the live prompts/short-answer items for the following school year's assessments have been selected by TEA. (Note: At this point, no new exit level TAAS prompts are being developed; current exit level TAAS prompts are selected from a bank of prompts previously field-tested.)

In late 2004 and early 2005, TEA and PEM staff independently scored samples of the field-test responses to the prompts/short answer items that were to be used on the live TAKS and SDAA II assessments in spring 2005. This scoring was in addition to the scoring already done by the field-test readers in the summer of 2004. TEA and PEM 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 were assigned both holistic and analytics scores. The scoring directors then selected responses from the rangefinding sessions to be included in each scoring guide. After TEA approval of these selections, the scoring directors wrote annotations for the guide responses; all annotations were reviewed and approved by TEA staff. The scoring directors then assigned the remaining prescored responses from the rangefinding sessions to "practice" sets and "qualifying" sets for use in reader training.

Reader Training Process

Overview

All readers and scoring supervisors receive extensive training on materials based on the prompts and/or short answer items related to each TAKS, SDAA II, and exit level TAAS assessment. Readers attend various training and practice 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 actual scoring begins.

After the scoring and discussion of the practice sets, 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 actual, or "live," student responses. The TAKS and SDAA II scoring guides used to train the readers during the 2004–2005 school year were released in the summer of 2005. These scoring guides were sent to districts and educational service centers; the guides are also available on the TEA website.

All responses, both compositions and short-answer responses, are scored based on the "perfect agreement" model. During holistic scoring, responses receive scores from two independent readers. If the two scores do not agree, a third reader who is especially experienced scores the response. If at that point perfect agreement has not yet been reached, the response is given to the scoring director for resolution.

For the projects dealing with written compositions, two sets of readers are employed: holistic readers and analytics readers. Holistic readers score the compositions, and analytics readers review all of the unsuccessful papers in order 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 1/2 line specialists, are trained to provide a score-verification procedure to further evaluate all responses that received a "1" (unsuccessful) score during the holistic scoring process. This step is taken before the responses are sent to the analytics group. If the 1/2 line score verification specialists determine that a particular response may be higher than a 1, the specialist coordinator also evaluates the paper; if the specialist coordinator agrees, the response is then 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 holistic readers, analytics readers, and at the exit level, 1/2 line specialists. In addition, TEA representatives select validity responses (see Chapter 15: Validity) and work with PEM staff and analytics coordinators in preparation for analytics scoring and, at the exit level, 1/2 line specialist scoring. Throughout the scoring project, TEA staff are consulted on "decision papers," which are responses that are highly unusual or that require a policy decision from TEA.

The ePEN System

In spring 2005 the following responses were scored using the electronic Performance Evaluation Network (ePEN) system: Grade 10 TAKS written compositions, all TAKS short-answer responses (Grades 9, 10, and exit level), and all SDAA II responses at all instructional levels. The ePEN system enables readers to read the scanned response on a computer monitor and then select a score for the response from a menu on the screen. It is important to remember that the ePEN system is simply an alternative way of handling the flow of responses to readers. Like the readers who read responses on paper answer documents, the readers who work on an ePEN project still read the responses exactly as the students wrote them and still make scoring judgments using the applicable rubrics. The readers receive the same rigorous and thorough training in the application of the relevant scoring criteria whether they work on a paper document project or on an ePEN project.

Training

TAKS Written Composition: Holistic Scoring

Guides: Each written composition scoring guide contained a total of 16 student responses: four annotated anchor responses representing each score point in order from 1 to 4.

Training Sets: Three training sets delineating the 1/2 line, the 2/3 line, and the 3/4 line contained randomly mixed responses representing the selected scores. In addition, readers used three sets containing 10 randomly mixed responses representing all score points.

Qualifying Sets: The three qualifying sets contained 15 randomly mixed responses representing all score points.

TAKS Written Composition: Analytics Scoring

The following analytics categories were used to explain why responses that were given a holistic score of 1 (scorable but unsuccessful) or 0 (nonscorable) received these scores:

Holistic Score Point 1 Analytics Categories	Holistic Score Point 0 Analytics Categories
Weak focus and coherence	Off-topic response
Weak or illogical organization	Indecipherable response
Weak development	Insufficient response
Little or no sense of voice	Blank response
Little or no control of conventions	

Training materials consisted of an abbreviated version of the holistic guide that included only the score point 1 and score point 2 papers, the 1/2 line set from the holistic training, an eight-paper analytics guide, and an explanation of the analytics categories and the numerical system used to assign the appropriate category or categories to each response. In addition, readers received four 10-paper training sets representing the various categories and allowable combinations of categories, as well as a number of somewhat successful responses to verify that readers could identify these types of responses.

TAKS Written Composition: Exit Level Verification Scoring

These training materials consisted of a guide using the 1s and 2s from the holistic guide, along with 10-paper sets that constituted the rest of the 1s and 2s from the holistic training and qualifying sets. Additional responses found in the “live” papers (approved by the scoring director[s], the coordinator of the analytics readers, the coordinator of the score verification specialists, and TEA staff) rounded out the sets.

TAKS Short-Answer Responses

Guides: The short-answer guides are divided into sections based on the items that were tested. Three short-answer items were tested at each grade level: an objective 2 item based on the literary selection, an objective 3 item based on the expository selection, and an objective 3 item based on both the literary and the expository selections, known as the crossover item. For each item, the short-answer scoring guides contain a total of 16 student responses: four annotated anchor responses representing each score point in order from 0 (insufficient) to 3 (exemplary).

Training Sets: For each item, there were three sets containing 10 randomly mixed responses representing all score points.

Qualifying Sets: For each item there were two qualifying sets that contained 15 randomly mixed responses representing all score points.

SDAA II: Instructional Level K/1

Performance Tasks—Writing Numbers, Writing Letters, and Writing Labels

Guide: For each task, the scoring guide contains four nonscorable responses and four scorable responses at each performance level (emergent, developing, and developed).

Training Sets: Training Sets A and B each contain 10 randomly mixed responses, including nonscorable, emergent, developing, and developed responses.

Qualifying Set: The qualifying set contains 10 randomly mixed responses, including nonscorable, emergent, developing, and developed responses.

Performance Task—Writing Names

Guide: The scoring guide contains four nonscorable responses, and four responses at each additional level of performance (emergent, developing, and developed).

Training Sets: Training Sets A and B each contain 10 randomly mixed responses in each set, including emergent, developing, and developed responses.

Qualifying Set: The qualifying set contains 10 randomly mixed responses, including emergent, developing, and developed responses.

Composition—Narrative Writing

Guide: The scoring guide for narrative writing contains one response representing each language level in order from 0 to 6 and three responses representing all possible analytics scores for each category—Attention to Prompt; Letter Formation/Spacing; Spelling, Capitalization, Punctuation; and Development of Narrative.

Training Sets: Training Sets A and B each contained 10 randomly mixed responses representing different combinations of language levels and analytics scores. Training Set C contained 15 randomly mixed responses representing different combinations of language levels and analytics scores.

Qualifying Set: The qualifying set contains 15 randomly mixed responses representing different combinations of language levels and analytics scores.

SDAA II Holistic Writing Scoring: Instructional Levels 2, 3/4, 5, 6/7, 8/9, and 10

Guide: Each written composition scoring guide contains a total of 16 student responses: four annotated anchor responses representing each score point in order from 1 to 4.

Training Sets: Training Sets A, B, and C each contain 10 randomly mixed responses representing all score points.

Qualifying Sets: The Qualifying Set contains 12 randomly mixed responses representing all score points.

SDAA II: Analytic Writing Scoring

The following analytics codes were used to explain why responses that were given a holistic score of 1 (scorable but unsuccessful) or 0 (nonscorable), received these scores:

Holistic Score Point 1 Analytics Categories	Holistic Score Point 0 Analytics Categories
Lacks clarity	Off-topic response
Lacks language control	Indecipherable response
Lacks support/elaboration	Insufficient response
Drifts from specified purpose	Blank response
Uses wrong purpose	
Drifts from specified topic	

The analytics guide consisted of the 1s and 2s from the holistic guide. There were four training sets with 10 responses each, drawn both from the holistic training materials and from “live” scoring.

Training Procedures

Scoring Supervisor and Reader Training: TAKS Written Composition

The scoring directors conduct the scoring supervisor and reader training for holistic scoring. However, 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, using the model described below. Throughout their training, scoring supervisors are encouraged to ask questions and to discuss any problems they had with the guide and the training sets. They are required to annotate their sets of training papers and to practice explaining their annotations to the rest of the group. Through this procedure the scoring supervisors develop confidence in their ability to explain why a paper has been given a particular score.

The guidelines for scoring supervisor and reader training are essentially the same. The specific steps are as follows, with the explanations and discussions being led by the scoring director.

1. Present the prompt in the exact form in which it was administered. For the Grade 10 and exit level ELA projects, the reading selections are first read by trainees, and any questions about the material are answered before training on the papers begins.
2. Read and explain the introductory section of the scoring guide.

3. Present an example of a “highly effective” paper (one that received a 4) from the scoring guide.
4. Proceed through the guide in the following manner:
 - a. Read and explain the score point 1 rubric. Read and discuss each annotated score point 1 paper.
 - b. Read and explain the score point 2 rubric. Read and discuss each annotated score point 2 paper.
 - c. Read and explain the score point 3 rubric. Read and discuss each annotated score point 3 paper.
 - d. Read and explain the score point 4 rubric. Read and discuss each annotated score point 4 paper.
5. Score and discuss the training sets for the 1/2 line, 2/3 line, 3/4 line and the two mixed sets.

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, he or she takes Set 2 and Set 3. Taking all the sets is important since scoring supervisors are responsible for working directly with the 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 set by TEA is dismissed.

Training of the analytics readers for all grades and of the 1/2 line 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 1/2 line specialist training, readers can begin “live” scoring only when they are able to demonstrate accuracy.

Scoring Supervisor and Reader Training: TAKS 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.

Scoring Supervisor and Reader Training: SDAA II

Instructional Level K/1

For the 2004–2005 SDAA II project, the training for scoring supervisors was specific to the various instructional levels. For Instructional Level K/1, scoring supervisors were trained on all the student tasks (writing numbers, writing names, writing letters, writing labels, and narrative writing). Throughout their training, scoring supervisors were encouraged to ask questions and to discuss any problems they were encountering with the guide and the training sets. They were required to annotate their sets of training papers and to practice explaining their annotations to the rest of the group. Through this procedure the scoring supervisors developed confidence in their ability to explain why a paper had been given a particular score.

The guidelines for scoring supervisor and reader training are essentially the same. The specific steps are as follows:

1. Present the item in the exact form in which it was administered.
2. Read and explain the introductory section of the scoring guide.
3. Proceed through the guide for each of the student tasks in the following manner:
 - a. Read and explain the “emergent” rubric. Read and discuss each annotated “emergent” paper.
 - b. Read and explain the “developing” rubric. Read and discuss each annotated “developing” paper.
 - c. Read and explain the “developed” rubric. Read and discuss each annotated “developed” paper.
4. Score and discuss Training Sets A and B.

After completing all the training sets, both scoring supervisors and readers take the qualifying set.

After scoring supervisors and readers are qualified, they are given a special training to familiarize themselves with the ePEN system.

Instructional Levels 2, 3/4, 5, 6/7, 8/9, 10

Scoring supervisors were trained specifically on the Instructional Level to which they were assigned. For Instructional Levels 2 and 3-4, the same scoring supervisors were used, as were IL 5, 6/7, and 8/9, but each scoring supervisor went through a separate training for the levels.

The scoring directors conducted the scoring supervisor and reader training. However, to ensure that the scoring supervisors were prepared to answer reader questions during and after the training and to ensure that the scoring supervisors were highly qualified to perform their roles during the scoring process, scoring supervisor candidates were trained before the readers. Throughout their training, scoring supervisors were encouraged to ask questions and

to discuss any problems they were encountering with the guide and the training sets. They were required to annotate their sets of training papers and to practice explaining their annotations to the rest of the group. Through this procedure the scoring supervisors developed confidence in their ability to explain why a paper had been given a particular score.

The guidelines for scoring supervisor and reader training are essentially the same. The specific steps are as follows:

1. Present the prompt in the exact form in which it was administered.
2. Read and explain the introductory section of the scoring guide.
3. Present a good paper (one that received a 4) from the scoring guide.
4. Proceed through the guide in the following manner:
 - a. Read and explain the score point 1 rubric. Read and discuss each annotated score point 1 paper.
 - b. Read and explain the score point 2 rubric. Read and discuss each annotated score point 2 paper.
 - c. Read and explain the score point 3 rubric. Read and discuss each annotated score point 3 paper.
 - d. Read and explain the score point 4 rubric. Read and discuss each annotated score point 4 paper.
5. Score and discuss Sets A, B, and C.

After completing all the training sets, the scoring supervisors took the qualifying sets. Regardless of whether a scoring supervisor scored well enough on Set 1 to qualify, he or she took Set 2. Taking the sets was important, since scoring supervisors are responsible for working directly with the readers, making it necessary for them to understand all the qualifying sets. After the scoring supervisors were qualified, they were trained to use the ePEN system.

When the readers arrived, they were divided among the Instructional Levels K/1, 2, 3/4, 5, 6/7, 8/9, 9 reading, 10 ELA. All readers were trained using the scoring-supervisor training model described above.

The readers assigned to the K/1 project were trained on all the student tasks (writing numbers, writing names, writing letters, writing labels, and narrative writing) and then broken out to focus on the task on which they demonstrated the highest skill. In addition, the SDAA II readers went through ePEN training since the students' responses were scored using the ePEN system.

Scoring Supervisor and Reader Training: SDAA II Short-Answer Responses

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.

Reader Training: Exit Level TAAS

The scoring director presented the guide papers. Then practice sets were taken and discussed.

Readers had to demonstrate that they could score accurately before they could begin reading “live” responses. Readers were allowed three opportunities to qualify. Any reader unable to meet the standards set by TEA was dismissed.

Training of the analytics readers and of the 1/2 score verification specialists followed a similar pattern as TAKS. Although no qualifying sets were used with these groups, readers could begin “live” scoring only when they were able to demonstrate accuracy.

Ongoing Roomwide Training: All Projects

After initial training, ongoing training is provided routinely to prevent “drift” and to ensure high reader agreement. Scoring directors plan for at least three ongoing training sessions a week. These methods are described in the following paragraphs.

One method is the scoring and discussion of sets of three to five papers each. The scoring directors start accumulating copies of papers that are typical close (or “line”) calls. The scoring directors review these papers with senior scoring staff and then circulate them among scoring supervisors to ensure team-to-team consistency on these difficult papers. Both shifts use these sets. Discussion of these sets sometimes occurs roomwide and sometimes in teams.

While scoring papers and spot-checking the accuracy of readers’ scoring, scoring supervisors are instructed to collect various types of problematic papers. These papers are reproduced and put into small sets for readers to score. After both scoring directors, the project monitor, and, in the case of a “decision” paper, a TEA representative agree on the scores of these papers, the sets are administered to the readers. Discussion of these papers is conducted roomwide. Only one or two of these sets are needed, depending on the grade level. If individuals need more help, the floating scoring supervisor works with them.

Every Monday 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: Paper Document Projects

In addition to the ongoing training methods mentioned above, the scoring centers employ a number of informal methods to identify individual reader scoring problems. Scoring directors and scoring supervisors rely on individual and small-group retraining to ensure that readers

are consistently applying the pre-established criteria when scoring. Scoring supervisors spot-check and annotate reader packets throughout the project and then return packets to the readers for their review. If necessary, the scoring supervisors provide one-on-one assistance to a reader and discuss discrepant scores. Readers also flag papers that are difficult for them to score. Scoring supervisors read these papers and then discuss each paper with the reader who flagged it.

Early in the project, scoring supervisors closely monitor all readers, spot-checking according to the following: scoring trends identified from training results, reports of “true” score reliability, and daily reader status reports. (See “Data-Entry Procedures and Resulting Reports.”) The need to spot-check every reader decreases as it becomes clear which readers consistently apply the scoring criteria and which need additional support. At this point, scoring supervisors concentrate on readers who score below 80 percent on the validity packets and/or who are below the room average on the daily reader status reports. Scoring supervisors conduct hands-on retraining by identifying problem papers, having readers articulate their reasoning for assigning a particular score and reinforcing the rubric and training papers to improve readers’ accuracy.

Another method used when a scoring supervisor suspects that a certain reader might not be using the criteria properly is to obtain a regular packet that has been scored first by the floating scoring supervisor. Distribution of this type of packet is done routinely so attention will not be called to it as a training device. The reader’s scoring supervisor then compares the floating scoring supervisor’s scores with those of the reader. If there are a number of discrepant scores, the floating scoring supervisor or the reader’s scoring supervisor discusses the papers with the reader to help him or her apply the criteria consistently.

Packets scored by a reader identified as having difficulty applying the criteria are retrieved and rescored by his or her scoring supervisor or by a reader at or above room average. The scoring supervisor then discusses with the reader the papers that have received discrepant scores. Any reader who cannot be successfully retrained on the criteria is dismissed.

Monitoring of Individual Readers: ePEN Projects

In addition to the ongoing training, readers are closely monitored by their scoring supervisor, the scoring director, and the project monitor. The computerized system of ePEN allows an up-to-date (updated approximately every 15 minutes) evaluation of the readers’ performance.

In addition, readers can 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.

As in the paper projects, responses scored by a reader identified as having difficulty applying the criteria are retrieved and rescored by his or her scoring supervisor or by a reader at or above room average. Any reader who cannot be successfully retrained on the criteria is dismissed.

Validity System

Paper Document Projects

As a method of detecting whether a room of readers is drifting from the scoring criteria, packets of prescored student responses, called validity packets, are assembled for each grade. For each scoring session, these papers are chosen from packets that have been read and assigned “true” scores by two scoring supervisors, the appropriate scoring director(s), and the TEA representative(s).

For the spring 2005 scoring project, 10 validity packets containing 10 papers each were used for TAKS Grade 4 English, Grade 7, and exit level; three validity packets were used for TAKS Grade 4 Spanish. During the day shift for the first two weeks of the project, each packet received first and second readings both in the morning and in the afternoon. After the second week of scoring, each packet received first and second readings once each day. Because the evening shift (Grade 4 English and exit level only) had fewer hours and fewer scorers, the packets were first- and second-read once each evening. All scorers read one or more of these validity packets during the course of the scoring project. Because the exit level TAAS project was completed in a week with few readers, validity packets were not used; instead, calibration sets effectively maintained scoring accuracy.

For each validity packet, PEM printed multiple monitor sheets listing each composition’s unique preprinted identification number. At the end of each shift, completed monitor sheets were processed, reports were printed, and new monitor sheets were inserted for the next shift’s scoring. Thus, senior staff had access to validity packet reports twice daily and could detect room drift and/or scorer drift almost as soon as it began.

ePEN Projects

Instead of packets of validity responses, 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 scores of all validity responses are agreed on by the scoring directors and TEA staff. Proposed validity responses are transferred to a Validity Folder on the ePEN system. TEA staff members have access to these files and have approved or rejected the proposed responses. Once the responses are approved, they are placed in a validity queue. The validity responses are shunted into the scoring queue at a rate of one validity response for each 30 responses scored.

Nonscorable Responses

Paper Document Projects

During holistic scoring, if a reader believes that a paper may be nonscorable, it is flagged for the scoring director to read and score. If the scoring director finds the paper to be nonscorable, the second reading is performed independently by the other scoring director or by the project monitor. Nonscorable responses are then evaluated by the analytics readers.

ePEN Projects

During holistic scoring, if a reader believes that a response may be nonscorable, it 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 then 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 are then sent to the analytics queue for evaluation by the analytics readers.

Paper-Flow and Resolution Procedures

Paper Document Projects

A scoring director supervises the day shift of readers for each grade; his or her counterpart supervises the evening shift. In 2004–2005, exceptions were Grade 4 TAKS Spanish and exit level TAAS, which were each scored by one room of readers during a day shift. Continuity between the day shift and the evening shift was maintained in a number of ways, including a 2 1/2-hour overlap in the work schedule of the scoring directors. The schedules of supervisors in the data-entry room and warehouse also overlapped so that continuity could be maintained in those areas.

When first received from school districts, student answer documents are scanned. During the scanning process, the two lined pages on which students wrote their compositions are separated from the multiple-choice section of the answer document. The two sections of the answer document are linked by a unique number printed on each page so that the composition's score can be added to the student's record once scoring is complete. The writing pages are then assembled into packets containing 40 or fewer papers each. A packet header sheet is placed with the packet of papers, and the packet is stapled together and put into an envelope with two scoring monitor sheets. As a result of this process, the only identifying information on the student papers is the six-digit identification number preprinted on the answer document. 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 unavailability of identifying information on the papers helps ensure unbiased scoring.

Each reader independently reads an entire packet of papers, writing his or her reader number on both the packet envelope and the monitor sheet. The reader records the scores for the papers on the monitor sheet, on which the identification numbers of the essays in the packet have been preprinted. The completed first-reading monitor sheet is separated from the packet envelope before the packet is given to a second reader. The reader number on the packet envelope identifies the reader's team as well as the individual to ensure that the same packet will not be read by another reader on the same team as the first reader.

Following scanning of both the first- and second-reading monitors, third-reading monitor sheets identifying responses needing an additional reading are produced. Only readers identified as being above room average in the accuracy of their scoring are allowed to be resolution, or third, readers. Early in a project they are selected on the basis of their performance in training, such as their scores on training sets and the caliber of their questions and comments, along with their scoring supervisor's assessment of their "live" scoring. Later the daily reader status reports and validity reports are invaluable in identifying the readers whose scoring accuracy is above room average. Designated third readers are not allowed to score third readings exclusively. Rather, they are required to score at least two 40-paper packets daily so that sufficient data can be collected to monitor their scoring on an ongoing basis. Any third reader whose perfect agreement rate on the daily status report drops is confined to performing first and second readings. Occasionally a fourth reading of a student paper is necessary. When this occurs, the fourth-reading monitor sheets are matched to the packets and given to scoring directors for scoring.

Responses requiring analytics evaluation are identified on an analytics monitor sheet and delivered to the analytics room.

TAKS ePEN Projects

A scoring director supervises the day shift of readers for each grade. His or her counterpart supervises an evening shift. Continuity between the day shift and the evening shift is maintained in a number of ways, including a 2 1/2-hour overlap in the work schedule of the scoring directors. Texas Education Agency and Pearson Educational Measurement project monitors are able to view reports electronically for all readers to ensure continuity between shifts. The review process can be done on-site or remotely via the Internet from computers with secure access.

When first received from school districts, student answer documents are scanned. During the scanning process, the pages on which students wrote their 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 that the performance-task scores can be added to the student's record once scoring is complete. The performance-task responses are then 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 unavailability of identifying information on the responses helps ensure unbiased scoring.

The responses are then 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. An employee number that identifies the reader is electronically attached to the response in such a way that only scoring supervisors, scoring directors, and project monitors can identify which reader read 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 being above room average in the accuracy of their scoring are allowed to be resolution, or third, readers. Early in the project they are selected on the basis of their performance in training, such as their scores on training sets and the caliber of their questions and comments, along with their scoring supervisor's assessment of their "live" scoring. Later the daily reader status reports and validity reports are invaluable in identifying the readers whose scoring accuracy is above room average. Designated third readers are not allowed to score third readings exclusively. Rather, they are required to score at least two hours daily on first and second readings so that sufficient data can be collected to monitor their scoring on an ongoing basis. Any third reader whose perfect agreement rate on the updated status report drops is confined to performing first and second readings. Occasionally a fourth reading of a student response is necessary. When this occurs, the fourth readings are placed in an adjudication, or fourth-reading, queue and scored only by scoring directors or project monitors. Grade 10 compositions receiving a holistic 1 or holistic 0 (nonscorable) are shunted to the analytics queue for evaluation by the analytics group.

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

SDAA II ePEN Projects

A scoring director supervises the day shift of readers for each instructional level. Texas Education Agency and Pearson Educational Measurement project monitors are able to view reports electronically for all readers to ensure continuity between shifts. The review process can be done on-site or remotely via the Internet from computers with secure access.

When first received from school districts, student answer documents are scanned. The sections of the answer document are linked by a unique number printed on each page so that the score can be added to the student's record once scoring is complete. The responses are then 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 unavailability of identifying information on the responses helps ensure unbiased scoring.

The responses are then grouped by task and stored on an ePEN server to which only qualified scoring directors, readers, and project monitors have access. As readers score responses, more responses are shunted into their scoring queues.

Each reader reads a response and selects a score from a menu on the computer screen. An employee number that identifies the reader is electronically attached to the response; only scoring supervisors, scoring directors, and project monitors can identify which reader read which response. The responses for K/1 receive a single reading, but the scoring supervisors read behind the readers to ensure accuracy. The compositions at Instructional Levels 2, 3/4, 5, 6/7, 8/9, 9 reading, and 10 ELA receive second readings and, if necessary, third, fourth, or analytics readings, as described previously in the "TAKS ePEN Projects" section.

Data-Entry Procedures and Resulting Reports

Paper Document Projects

The packet monitor sheets are scanned at the scoring centers, and the scores are transmitted to PEM in Iowa City. After the scores for the first and second readings of a packet have been scanned, the resolution monitor sheet (third-reading monitor) is produced. Pearson Educational Measurement transmits the data for third-reading monitor sheets (as well as fourth-reading, analytics, and specialist monitor sheets) to the PEM Performance Scoring Center's printer. Monitor sheets are then printed.

Project status reports are produced so that senior staff and scoring directors have up-to-date information on the progress of the entire project at all scoring centers. These reports provide a wealth of information about the scoring patterns of individual readers. In addition to the number of responses read by each reader, the reports include the following for each reader: number of third readings completed, percentage of responses read in perfect agreement with the other scorer, and percentage of responses read in perfect agreement with the other scorer in combination with responses read in perfect agreement with the resolver. In every resolution reading, one reader's score is judged to be incorrect; consequently, the reports have three adjacent score categories: 1/2, 2/3, and 3/4. These show the number of times the reader's incorrect scores are higher and/or lower for each of the adjacent score categories. The final columns on the reader status reports give the reader's distribution of score points—that is, what percentage of a particular reader's scores were 1s, 2s, etc. Accompanying the daily (or current) reader status report is the year-to-date report, which has the same information but is cumulative for the project as of that date.

ePEN Projects

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. (For SDAA II K/1, after the score for the first reading of a response has been processed, the ePEN system holds the score until the project monitor is satisfied with its accuracy. Then it is uploaded as a final score.)

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. These reports provide a wealth of information about the scoring patterns of individual readers. In addition to the number of responses read by each reader, the reports for short-answer reading items include the following for each reader: number of third readings completed, percentage of responses read in perfect agreement with the other scorer, and percentage of responses read in perfect agreement with the other scorer in combination with responses read in perfect agreement with the resolver. In every resolution reading, one reader's score is judged to be incorrect; consequently, the reports have three adjacent score categories, 0/1, 1/2, and 2/3. These show the number of times the reader's incorrect scores are higher and/or lower for each of the adjacent score categories. The final columns on the reader

status reports give the readers’ distribution of score points—that is, what percentage of a particular reader’s scores are 0s, 1s, etc. Accompanying the daily (or current) reader status report is the year-to-date report, which has the same information but is cumulative for the project as of that date.

Score Reliability and Validity Information: TAKS

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.

The following two tables summarize reader agreement rates (reliability) by grade and results of the use of validity packets by grade. Reader agreement rate is expressed in terms of absolute agreement (Reader 1’s score equals Reader 2’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.

Table 18. Summary of Scorer Agreement (Reliability)

Grade	Number of Responses Read	Agreement Rate (%) After 2 Readings	Number of Third Readings	Agreement Rate (%) After 3 Readings
4 (English)	275,131	72	76,402	98.4
4 (Spanish)	18,386	75	4,678	99.3
7	300,163	72	83,763	98.7
9	1,036,473	75	265,696	99.3
10 WC	278,326	65	94,757	97.3
10 SA	834,978	70	249,124	98.9
11 WC	256,391	73	68,218	98.7
11 SA	769,173	72	214,124	98.7

Table 19. Summary of Validity Packet Results

Grade	Agreement Rate (%)
4 (English)	70.6
4 (Spanish)	78.1
7	77.8
9	87.6
10 WC	74
10 SA	85
11 WC	77
11 SA	83

It should be noted that 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.

Score Reliability and Validity Information: SDAA II

Throughout the years, TEA has reported on the reliability and the 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 responses selected and examined by TEA staff.

The following two tables summarize reader agreement rates (reliability) by instructional level and results of the used of validity by instructional level. Reader agreement rate is expressed in terms of absolute agreement (reader 1’s score equal Reader 2’s score). Validity is expressed in terms of perfect agreement between the score assigned by a give reader and the “true” score assigned by TEA.

Table 20. Summary of Scorer Agreement (Reliability)

Instructional Level	Number of Responses Read	Agreement Rate (%) After 2 Readings	Number of Third Readings	Agreement Rate (%) After 3 Readings
*K/1: Essay	11,183	NA	NA	NA
*K/1: Numbers	11,183	NA	NA	NA
*K/1: Labels	11,183	NA	NA	NA
*K/1: Letters	11,183	NA	NA	NA
*K/1: Names	11,183	NA	NA	NA
2 WC	15,556	77	3,548	99.6
3/4 WC	29,898	77	6,698	99.6
5 WC	11,025	73	2,743	99.0
6/7 WC	12,719	75	3,044	99.4
8/9 WC	1,788	80	333	99.7
10 WC	4,413	87	538	99.9
9 SA	11,250	78	2,399	98.8
10 SA	13,239	82	2,384	99.2

* K/1 responses receive only 1 holistic read.

Table 21. Summary of Validity Results

Instructional Level	Agreement Rate (%)
K/1: Essay	*LL (97.1), AtP (95.4), LF/S (84.6), SCP (81.3), DoN (76.7)
K/1: Numbers	97.1
K/1: Labels	86.5
K/1: Letters	98.0
K/1: Names	83.3
2 WC	82.8
3/4 WC	82.4
5 WC	71.9
6/7 WC	73.6
8/9 WC	86.2
10 WC	80.4
9 SA	92.6
10 SA	88.6

* K/1 essay validity is given in five traits: Language Level (LL), Attention to Prompt (AtP), Letter Formation and Spacing (LF/S), Spelling, Capitalization and Punctuation (SCP), and Development of Narrative (DoN).

It should be noted that student response scores are based on the score that has been agreed upon independently by at least two of the three readers (K/1 responses receive only one holistic read). 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: TAKS and SDAA II

After all live scoring was completed in late spring 2005, small groups of experienced readers were selected to score the responses generated by representative samples of students during field-testing in February. 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 live 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 a live 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 actual assessment.

Because of mandated changes to the SDAA assessment program, the new program, SDAA II, was field-tested in spring 2004. SDAA II is an updated assessment that is more aligned with TAKS. The new tests that include performance tasks are based on the following instructional levels: K/1, 2, 3/4, 5, 6/7, 8/9, and 9 reading, and 10 ELA.

Appeals: All Projects

Pearson Educational Measurement rescues any TAKS, SDAA II, or exit level TAAS response about which questions have been raised regarding the assigned score. If PEM 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, PEM provides an analysis of the response in question, to determine whether or not the score is changed.