

CHAPTER 8: PERFORMANCE ASSESSMENT—TAKS, SDAA, AND EXIT LEVEL TAAS

TAKS

The TAKS writing tests at grades 4 (both the English and Spanish versions) and 7, as well as the English language arts (ELA) tests at grades 10 and 11, include a written composition component. In addition, short-answer questions are included on the grade 9 reading test and the ELA tests at grades 10 and 11. 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 measurement 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 ideas 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 preestablished criteria: focus and coherence, organization, depth of development, voice, and control of conventions. These criteria, explained in detail in the scoring rubrics for written compositions, 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 grades 10 and 11 ELA tests is designed to test the student's 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.

WRITTEN COMPOSITION SCORING

The written composition component of the TAKS writing and ELA assessments is scored by readers hired by Pearson Educational Measurement (PEM); these readers are organized into teams that are coordinated by scoring supervisors (formerly known as "team leaders"). All readers and scoring supervisors receive extensive training on materials related specifically to the writing prompts for each administration. Through various training and practice sessions, during which readers score papers that have predetermined scores, readers are required to demonstrate complete understanding of the scoring process and must agree with the a priori scores on the training papers. During the actual scoring of student responses, validation sets of papers are systematically distributed to readers to monitor whether they are consistently applying the criteria and whether there is any drift from the "true" score. Released scoring guides, which include rubrics and sample student responses, were sent to all Texas school districts and regional education service centers in the summer of 2003; the guides are also available on the TEA Web site.

TRAINING MATERIALS AND MONITORING OF PROJECT

The spring 2003 scoring of the written composition portion of the TAKS writing and ELA tests took place at three different Pearson Performance Scoring Center locations: grade 4 English responses were scored in Lawrence, Kansas; grade 7 responses were scored in Dallas, Texas; and grades 10 and 11 ELA and grade 4 Spanish responses were scored in Austin, Texas. The short-answer portion of the grade 9 reading test and the grades 10 and 11 ELA tests was scored at two Pearson Performance Scoring Center locations: grades 10 and 11 ELA in Albuquerque, New Mexico, and grade 9 reading in Dallas, Texas. Preparation for TAKS focused holistic scoring proceeded according to the following plan. PEM senior project staff and TEA staff, after independently scoring a selection of field-test responses, met in rangefinding sessions in November 2002, December 2002, and January 2003 to discuss those responses and to assign “true” scores—both holistic and analytic—to the compositions. The short-answer responses also went through this rangefinding procedure, although short-answer responses do not receive analytic scores.

After range finding, the responses and their “true” scores were provided to the scoring directors of the respective grades, who worked with senior staff in Austin, Dallas, and Lawrence to assemble the materials for training scoring supervisors and readers. Following TEA approval of the responses proposed for the scoring guides, the scoring directors assigned the remaining prescored responses to training sets and qualifying rounds used to certify scoring supervisors and readers. After the scoring directors annotated the guide responses and received TEA approval for the annotations, all training materials were photocopied.

TEA staff monitored training in Austin, Dallas, Albuquerque, and Lawrence; selected validity papers (see “Validity Packets” on p. 48); and worked with senior staff and the analytic coordinators in preparation for analytic scoring of the compositions. In addition, working with the 1/2 score verification specialist/analytic coordinator, TEA staff selected some “live” papers to include in 1/2 line and analytic training sets for the Austin grade 11 ELA project. Throughout the scoring process, senior PEM staff served as on-site monitors in Austin, Dallas, Albuquerque, and Lawrence.

MANAGEMENT-LEVEL STAFF

The PEM contract with TEA stipulates that all management-level staff at the scoring centers be approved by TEA. Accordingly, PEM submitted lists of scoring director candidates to TEA for approval, and TEA approved all candidates. All staff had extensive experience with the TAAS and/or TAKS programs and with numerous other large-scale state writing assessments.

RECRUITMENT OF READERS AND SCORING SUPERVISORS

Requirements for readers included a bachelor’s degree, preferably in English, education, or a related field; teaching experience was preferred. In addition, all applicants were required to write an essay and complete a proofreading exercise. Those applicants interested in scoring the grade 4 Spanish written compositions were required to complete a Spanish decoding/translation exercise.

In February 2003, scoring supervisors were selected by scoring directors. Scoring supervisors were chosen from experienced readers and/or past scoring supervisors. This selection process was based on scoring supervisors’ understanding of the criteria, their ability to apply the criteria consistently and accurately, their ability to articulate the criteria, and their demonstration of leadership skills. Except for the “floating” scoring supervisor, scoring supervisors were assigned teams of 10 to 12 readers. The floating scoring supervisor assisted the scoring director with various administrative and quality-control activities. In Albuquerque and Dallas, new scoring supervisor candidates were chosen through an interview process and underwent an advanced training session where they were evaluated by TEA and PEM senior project staff before beginning the normal scoring supervisor training specific to the project.

In Austin a total of 517 people were involved in grades 10 and 11 ELA scoring, grade 4 Spanish scoring, the SDAA projects, and TAAS exit level scoring. (For details on SDAA and TAAS composition scoring, please see the relevant sections.)

In Lawrence a total of 191 people were involved in the spring 2003 scoring project for grade 4 English. In Dallas a staff of 435, including readers, scoring supervisors, scoring directors, and coordinators, worked to complete the spring 2003 project for grade 7 composition scoring and for grade 9 short answer scoring. In Albuquerque, 302 people worked to complete the grades 10 and 11 short answer portion of the ELA test.

ADMINISTRATIVE ARRANGEMENTS

The PEM scoring centers each had a day shift and a night shift throughout the focused holistic scoring of the spring 2003 TAKS responses. Grade 7 TAKS in Dallas had a day shift only. The analytic scoring and the score verification of failing exit level compositions (in Austin only) were accomplished during the day shift.

TRAINING

Each of the hand-scored projects has a similar structure for its training. The material consists of a training guide, practice sets, and qualifying sets. The content of the training material varies according to the project. A description of the training follows.

TRAINING: TAKS WRITTEN COMPOSITION

GUIDES

There were a total of sixteen student responses: four annotated anchor responses representing each score point in order, from 1 to 4.

TRAINING SETS

There were three training sets delineating the 1/2 line, the 2/3 line and the 3/4 line that contained randomly mixed responses representing the selected scores. There were also two sets that contained ten randomly mixed responses representing the score points 1–4.

QUALIFYING SETS

There were three qualifying sets that contained fifteen randomly mixed responses representing the score points 1–4.

These training materials included 111 responses for each grade level tested. Following is a breakout by score point:

Grade 4	Grade 4 (Spanish)	Grade 7	Grade 10	Grade 11
24-1s	30-1s	22-1s	23-1s	24-1s
35-2s	32-2s	33-2s	33-2s	33-2s
30-3s	26-3s	34-3s	35-3s	34-3s
+22-4s	+23-4s	+22-4s	+20-4s	+20-4s
111	111	111	111	111

TRAINING SET CONTENTS: ANALYTIC SCORING

The following analytic categories were used to explain why responses that received a rating of 1 or 0 (nonscorable responses) were unsuccessful.

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

Training materials consisted of an eight-paper guide, the 1/2 line set from the holistic training, an eight-paper “Analytics” guide, and an explanation of the analytic categories and the numerical system used to assign the appropriate category or categories to each response. In addition, analytic readers received four ten-paper training sets representing the various categories and allowable combinations of categories as well as somewhat successful responses to indicate that the reader could identify these types of responses.

SCORING SUPERVISOR TRAINING

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 using the same model as described below. Throughout their training, scoring supervisors were encouraged to ask questions and to discuss any problems they had 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 were essentially the same. The specific steps were as follows:

1. Present the prompt in the exact form in which it was administered. For the grade 9 reading project and the grade 10 ELA and grade 11 exit level ELA projects, the reading selections were read by the trainees and any questions about the material were answered.
2. Read and explain the introduction section of the scoring guide.
3. Present 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 took the qualifying sets. Regardless of whether a scoring supervisor scored well enough on Set 1 to qualify, he or she took Set 2 and Set 3. Taking all the sets was important, since scoring supervisors were responsible for working directly with the readers; consequently, it was necessary for them to understand all the qualifying sets.

TEA and PEM monitors observed the scoring supervisor training and determined which of the candidates would best serve as scoring supervisors for the project. Those not chosen were retrained as readers and very often contributed as accurate readers during the project.

READER TRAINING

Before training began, readers signed their contracts and nondisclosure forms, and TEA representatives made introductory remarks.

The scoring director discussed the prompt, introduced the guide, and then explained each score point to the entire group of readers. The readers took each of the training sets and the scoring director discussed each response with the readers. The readers were encouraged to ask questions to clarify papers with which they had had difficulty. TEA staff monitored this entire process.

Like scoring supervisors, readers had to demonstrate accuracy in their scoring before they could begin reading packets of responses. Readers were allowed three opportunities to qualify. Any reader unable to meet the standards set by TEA was dismissed.

Training of the analytic readers for all grades followed a similar pattern, except that the training was performed by the respective coordinators.

ONGOING ROOMWIDE TRAINING

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

One method was the scoring and discussion of sets of three to five papers each. The scoring directors started accumulating copies of papers that were typical close (or “line”) calls. The scoring directors reviewed these papers with senior scoring staff and then circulated them among scoring supervisors to ensure team-to-team consistency on these difficult decisions. Both shifts used these sets. Discussion of these sets sometimes occurred roomwide and sometimes in teams.

While scoring papers and spot-checking the accuracy of readers’ scoring, scoring supervisors were instructed to collect various types of problematic papers. These papers were 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 agreed on the scores of these papers, the sets were administered to the readers. Discussion of these papers was conducted roomwide. Only one or two of these sets were needed, depending on the grade level. If individuals needed more help, the floating scoring supervisor worked with them.

Every Monday the scoring directors reviewed the rubrics with readers and had them reread their anchor papers, emphasizing any area that appeared to be giving readers problems.

MONITORING OF INDIVIDUAL READERS

In addition to the ongoing training methods mentioned above, the scoring centers employed a number of informal methods to identify individual reader scoring problems. Scoring directors and scoring supervisors relied on individual and small group retraining to ensure that readers were consistently applying the preestablished criteria when scoring. Scoring supervisors spot-checked and annotated reader packets throughout the project and then returned packets to the readers for their review. If necessary, the scoring supervisors would provide one-on-one assistance to a reader and discuss discrepant scores. Readers also flagged papers that were difficult for them to score. Scoring supervisors read these papers and then discussed each paper with the reader who had flagged it.

Early in the project, scoring supervisors closely monitored all readers, spot-checking according to the following: scoring trends identified from training results, reports of “true” score reliability (see “Validity Packets” below), and daily reader status reports (see “Data-Entry Procedures and Resulting Reports” on p. 50). The need to spot-check every reader decreased as it became clear which readers were consistently applying the scoring criteria and which needed additional support. At this point, scoring supervisors concentrated on readers who scored below 80% on the validity packets and/or who were below the room average on the daily reader status reports. They conducted 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 suspected that a certain reader might not be using the criteria properly was to obtain a regular packet that had been scored first by the floating scoring supervisor. Distribution of this type of packet was done routinely so attention would not be called to it as a training device. The reader’s scoring supervisor then compared the floating scoring supervisor’s scores with those of the reader. If there were a number of discrepant scores, the floating scoring supervisor or the reader’s scoring supervisor discussed 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 were retrieved and rescored by his or her scoring supervisor or by a reader at or above room average. The scoring supervisor then discussed with the reader the papers that had received discrepant scores. Any reader who could not be successfully retrained on the criteria was dismissed.

VALIDITY PACKETS

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

For the spring 2003 scoring project, ten validity packets containing ten papers each were used for grade 4 English, grade 7, grade 10, and grade 11; three validity packets were used for grade 4 Spanish. During the day, 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 had fewer hours and fewer readers, the packets were first- and second-read once each evening. All readers read one or more of these validity packets during the course of the scoring project.

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 reader drift almost as soon as it began.

ANALYTIC SCORING

At all grades each composition that received either a rating of 0 or a score of 1 was evaluated analytically to provide information about the specific weaknesses that caused it to be unsuccessful. Analytic readers were trained on all the analytic features simultaneously. Papers that exemplified the range of unsuccessful compositions and that TEA and senior scoring center staff agreed on in advance were selected as training papers. The scoring director first read and discussed the guide with the analytic readers. The guide included eight sample papers that were chosen to represent a variety of analytic scores. The analytic readers read, scored, and discussed four additional sets of ten papers each. The readers began "live" scoring when they were able to demonstrate accuracy on all analytic categories.

NONSCORABLE RESPONSES

During holistic scoring, if a reader believed that a paper may be nonscorable, the paper was flagged for the scoring director to read and score. If the scoring director found it to be nonscorable, the second reading was performed independently by the other scoring director or by the project monitor. Nonscorable responses were then evaluated by the analytic readers.

PROCEDURES

PAPER-FLOW AND RESOLUTION PROCEDURES

A scoring director supervised the day shift of readers for each grade; his or her counterpart supervised the evening shift. In Dallas all grade 7 scoring was accomplished in a day shift and the readers were divided between two scoring directors. 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 overlapped so that continuity could be maintained in those areas. In Dallas the two scoring directors worked closely to ensure the continuity between the two rooms.

The logistics of paper flow in the scoring centers was carefully planned and carried out. The answer documents were sent to the PEM Tech Ridge facility in Austin, where they were scanned. During the scanning process, the two lined pages on which students wrote their compositions were separated from the multiple-choice section of the answer document. The two sections of the answer document were linked by a unique number printed on each page so that the composition's score could be added to the student's record once scoring was complete. The writing pages were then assembled into packets containing 40 or fewer papers each. A packet header sheet was placed with the packet of papers, and the packet was 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 was 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, the readers had no knowledge of who the students were. The unavailability of identifying information on the papers helped ensure unbiased scoring.

The packets were then boxed by grade and shipped to the scoring centers in Austin, Dallas, and Lawrence. Whenever a scoring room needed additional papers, an aide carried packets to the room. The room aide and scoring supervisors handled all paper flow in the scoring rooms.

Each reader independently read an entire packet of papers, writing his or her reader number on both the packet envelope and the monitor sheet. The reader recorded the scores for the papers on the monitor sheet, on which the identification numbers of the essays in the packet had been preprinted. The completed first-reading monitor sheet was separated from the packet envelope before the packet was given to a second reader. The reader number on the packet envelope identified the reader's team as well as the individual to ensure that the same packet would 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 were produced. Only readers identified as being above room average in the accuracy of their scoring were allowed to be resolution, or third, readers. Early in the project they were 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 were invaluable in identifying the readers whose scoring accuracy was above room average. Designated third readers were not allowed to score third readings exclusively. Rather, they were required to score at least two 40-paper packets daily so that sufficient data could be collected to monitor their scoring on an ongoing basis. Any third reader whose perfect agreement rate on the daily status report dropped was confined to performing first and second readings. Occasionally a fourth reading of a student paper was necessary. When this occurred, the fourth-reading monitor sheets were matched to the packets and given to scoring directors for scoring.

Responses requiring analytic scores were identified on an analytic monitor sheet and delivered to the analytic scoring room.

DATA-ENTRY PROCEDURES AND RESULTING REPORTS

The packet monitor sheets were scanned at the scoring centers, and the scores were transmitted to PEM in Iowa City. After the scores for the first and second readings of a packet had been scanned, the resolution monitor sheet (third-reading monitor) was produced. PEM transmitted the data for third-reading monitor sheets (as well as fourth-reading, analytic, and specialist monitor sheets) to the PEM Performance Scoring Center's printer. The monitors were then printed and delivered to the warehouse.

The data also produced project status reports that gave senior staff and scoring directors up-to-date information on the progress of the entire project at all scoring centers. These reports provided a wealth of information about the scoring patterns of individual readers. In addition to the number of responses read by each reader, the reports included 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 was judged to be incorrect; consequently, the reports had three adjacent score categories, 1/2, 2/3, and 3/4. These showed the number of times the reader's incorrect scores were higher and/or lower for each of the adjacent score categories. The final columns on the reader status reports gave the readers' 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 was the year-to-date report, which had the same information but was cumulative for the project as of that date.

SCORE APPEALS

PEM rescores any TAKS written composition about which questions have been raised regarding the assigned score. Through a telephone call to the district contact person, PEM provides an individual analysis of the composition in question.

TRAINING: TAKS SHORT-ANSWER RESPONSES

For the grade 9 reading and grades 10 and 11 ELA short-answer responses, PEM used the ePEN (electronic Performance Evaluation Network) to manage the handling of responses. This system enabled the readers to read each response on a computer screen and then to select a score from a menu on the screen.

The training model for the short-answer portion of the grades 9, 10, and 11 tests and the monitoring of the project followed the training procedures outlined above. Exceptions to the procedures follow:

GUIDES:

The short-answer guides were divided into sections based on the items that were tested. There were three items 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—this item was known as the cross-over item. For each item, there were a total of sixteen student responses: four annotated anchor responses representing each score point in order, from 0 to 3.

TRAINING SETS:

For each item, there were three sets that contained ten randomly mixed responses representing the score points 0-3.

QUALIFYING SETS:

For each item there were two qualifying sets that contained fifteen randomly mixed responses representing the score points 0-3.

These training materials included either 75 or 76 responses for each grade level and item tested. Following is a breakout by score point:

Grade 9

Literary Item	Expository Item	Cross-Over Item
20-0	21-0	21-0
24-1	23-1	23-1
20-2	21-2	21-2
11-3	11-3	11-3

Grade 10

Literary Item	Expository Item	Cross-Over Item
21-0	20-0	21-0
23-1	24-1	23-1
21-2	20-2	21-2
11-3	11-3	11-3

Grade 11

Literary Item	Expository Item	Cross-Over Item
19-0	21-0	20-0
23-1	23-1	23-1
22-2	21-2	22-2
12-3	11-3	11-3

SCORING SUPERVISOR TRAINING

Scoring supervisor candidates in both Albuquerque and Dallas underwent an advanced training two weeks before the formal scoring supervisor training. This was done so that TEA and senior PEM project staff could evaluate the candidates to select the most qualified and to give the scoring supervisors additional experience with the rubrics.

After the scoring supervisors were chosen, they went through the regular training and were trained on all three items. This ensured that the scoring supervisors were able to assist all the readers if needed. The actual training model followed the TAKS written composition training as noted above. After the scoring supervisors were qualified, they were given a special training to familiarize themselves with the ePEN system.

READER TRAINING

Before training, the readers were divided into three groups. Each group was trained on and scored one of the items. This allowed each group to focus fully on a particular question without being distracted by the other items. After the readers were qualified, they were trained to use the ePEN system.

ONGOING ROOMWIDE TRAINING

To ensure continued accuracy and reliability, scoring directors started accumulating copies of papers that were typical close (or “line”) calls. These responses were printed from the computer system. The scoring directors reviewed these papers with senior scoring staff and then circulated them among scoring supervisors to ensure team-to-team consistency on these difficult decisions. Both shifts used these sets. Discussion of these sets occurred roomwide.

MONITORING OF INDIVIDUAL READERS

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

In addition, readers could send responses that were difficult to score to their scoring supervisor, who could respond to the reader or pass the question along to the scoring director or project monitor. This allowed the readers to receive constant feedback on their performance.

Responses scored by a reader identified as having difficulty applying the criteria were retrieved and rescored by his or her scoring supervisor or by a reader at or above room average. Any reader who could not be successfully retrained on the criteria was dismissed.

VALIDITY PACKETS

Instead of packets of validity responses, the ePEN system allowed the project staff to insert validity responses within the scoring cycle without the readers being aware that what they were scoring was a validity response. The scores of all validity responses were agreed on by the scoring directors and TEA staff. Proposed validity responses were shunted to a Validity Folder on the ePEN system. TEA staff had access to these files and approved or rejected the proposed responses. Once the responses were approved, they were placed in a validity queue. The validity responses were shunted into the scoring queue at a rate of 1 validity response for each 30 responses scored.

NONSCORABLE RESPONSES

During holistic scoring, if a reader believed that a response may be nonscorable, it was sent to a review queue for the scoring supervisor to review. If the scoring supervisor determined that the response was scorable, he or she scored it and then responded to the reader. If the scoring supervisor also believed the response to be nonscorable, he or she alerted the scoring director and left the response in the review queue. If the scoring director found it to be nonscorable, the second reading was performed independently by the other scoring director or by the project monitor.

PROCEDURES

PAPER-FLOW AND RESOLUTION PROCEDURES

A scoring director supervised the day shift of readers for each grade; his or her counterpart supervised the evening 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. TEA and PEM project monitors were able to view reports electronically for all readers to ensure continuity between shifts. The review process could be done locally on-site or remotely via the internet from computers with secure access.

The logistics of the flow of responses in the scoring centers was carefully planned and carried out. The answer documents were sent to the PEM TechRidge facility in Austin, where they were scanned. During the scanning process, the two pages on which students wrote their short answer responses were separated from the multiple-choice section of the answer document. The sections of the answer document were linked by a unique number printed on each page so that the short answer scores could be added to the student's record once scoring was complete. The short answer responses were then given a unique ePEN identifying number. The ePEN number was 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, the readers had no knowledge of who the students were. The unavailability of identifying information on the responses helped ensure unbiased scoring.

The responses were then grouped by grade and stored on an ePEN server; only qualified scoring directors, readers, and project monitors had access to this server. As the readers scored the responses, more responses were shunted into their scoring queues.

Each reader independently read a response, selecting a score from a menu on the computer screen. An employee number that identified the reader was electronically attached to the response in a way that only scoring supervisors, scoring directors, and project monitors could identify which reader read which response. After the reader had completed a first reading of the response, the response was shunted into a second reader's queue for an independent reading.

Following completion of both the first and second reading, responses needing an additional reading were indicated and shunted into a resolution queue. Only readers identified as being above room average in the accuracy of their scoring were allowed to be resolution, or third, readers. Early in the project they were 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 were invaluable in identifying the readers whose scoring accuracy was above room average. Designated third readers were not allowed to score third readings exclusively. Rather, they were required to score at least two hours daily on first and second readings so that sufficient data could be collected to monitor their scoring on an ongoing basis. Any third reader whose perfect agreement rate on the updated status report dropped was confined to performing first and second readings. Occasionally a fourth reading of a student response was necessary. When this occurred, the fourth readings were placed in an adjudication or fourth-reading queue and scored only by scoring directors or project monitors.

Short-answer responses did not receive analytic readings.

DATA-ENTRY PROCEDURES AND RESULTING REPORTS

After the scores for the first and second readings of a response had been processed, the ePEN system created the resolution readings (third readings and fourth readings) if needed.

The data collected for first, second, third, and fourth readings produced project status reports that gave senior staff and scoring directors up-to-date information on the progress of the entire project at all scoring centers. These reports provided a wealth of information about the scoring patterns of individual readers. In addition to the number of responses read by each reader, the reports included 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 was judged to be incorrect; consequently, the reports had three adjacent score categories, 0/1, 1/2, and 2/3. These showed the number of times the reader's incorrect scores were higher and/or lower for each of the

adjacent score categories. The final columns on the reader status reports gave the readers' distribution of score points—that is, what percentage of a particular reader's scores were 0s, 1s, etc. Accompanying the daily (or current) reader status report was the year-to-date report, which had the same information but was cumulative for the project as of that date.

SCORE APPEALS

PEM rescores any short-answer response about which questions have been raised regarding the assigned score. Through a telephone call to the district contact person, PEM provides an individual analysis of the response in question.

SDAA

SDAA writing assessments are administered to students enrolled in grades 4 or 7. These assessments are available at instructional levels K/1/2, 3/4, 5/6, and 7.

In the 2002–2003 testing year, approximately 63,618 responses were scored. More than 65 readers were involved, with the largest number scoring the Instructional Level 3/4 responses. Analytic readers were selected from the holistic readers. All groups were trained by Pearson Educational Measurement (PEM) personnel and supervised by project monitors. The majority of the scoring leadership has worked with the SDAA scoring program since the first field test in 1999.

SCORING GUIDES AND TRAINING FOR INSTRUCTIONAL LEVEL K/1/2

Preparation began when responses from the 2002 field tests were selected by the scoring leadership for the training paper selection meetings. These were held in Austin over several dates in December 2002. Senior PEM staff met with members of the TEA Student Assessment staff to score and discuss responses to the performance tasks and writing prompt.

After “true” scores were assigned to the responses, they were given to the scoring director, who assembled the training materials. Following TEA approval of the scoring guides, the remaining prescored responses were used to assemble practice sets. The scoring materials were photocopied for the readers.

TEA representatives observed reader training. Throughout the training and scoring process, PEM project directors served as on-site monitors. Immediately following scoring of the spring administration, readers scored the 2003 field test responses. Scoring of the field test prompts was completed in June.

TRAINING SET CONTENTS: INSTRUCTIONAL LEVEL K/1/2

Scoring Guide for Writing Numbers: 3 nonscorable responses and 3 responses at each performance level—emergent, developing, somewhat developed, and developed.

Sets A and B: Ten randomly mixed responses in each set, including nonscorable, emergent, developing, somewhat developed, and developed responses.

Qualifying set: Ten randomly mixed responses, including nonscorable, emergent, developing, somewhat developed, and developed responses.

Following is a breakout of responses used in training by performance level:

9	Nonscorable
8	Emergent
9	Developing
9	Somewhat Developed
10	Developed

45 total responses used in training

Scoring Guide for Writing Names: One nonscorable response, three emergent responses, and four responses at each additional level of performance—developing, somewhat developed, and developed.

Sets A and B: Ten randomly mixed responses in each set, including emergent, developing, somewhat developed, and developed responses.

Qualifying set: Ten randomly mixed responses, including emergent, developing, somewhat developed, and developed responses.

Following is a breakout of responses used in training by performance level:

1	Nonscorable
4	Emergent
16	Developing
13	Somewhat Developed
12	Developed

46 total responses used in training

Scoring Guide for Writing Letters: Three nonscorable responses and three responses at each performance level—emergent, developing, somewhat developed, and developed.

Sets A and B: Ten randomly mixed responses in each set, including nonscorable, emergent, developing, somewhat developed, and developed responses.

Qualifying set: Ten randomly mixed responses, including nonscorable, emergent, developing, somewhat developed, and developed responses.

Following is a breakout of responses used in training by performance level:

7	Nonscorable
9	Emergent
11	Developing
9	Somewhat Developed
9	Developed

45 total responses used in training

Scoring Guide for Writing Labels: Three nonscorable responses and three responses at each performance level—emergent, developing, somewhat developed, and developed.

Sets A and B: Ten randomly mixed responses in each set, including nonscorable, emergent, developing, somewhat developed, and developed responses.

Qualifying set: Ten randomly mixed responses, including nonscorable, emergent, developing, somewhat developed, and developed responses.

Following is a breakout of responses used in training by performance level:

9	Nonscorable
10	Emergent
7	Developing
11	Somewhat Developed
8	Developed
<hr/>	
45	total responses used in training

Scoring Guide for Narrative Writing: One response representing each language level in order from 0 to 6.

3 responses representing the analytic score for Attention to Prompt

3 responses representing the analytic score for Letter Formation/Spacing

3 responses representing the analytic score for Spelling, Capitalization, Punctuation

3 responses representing the analytic score for Development of Narrative

Sets A and B: Ten randomly mixed responses in each set representing different combinations of language levels and analytic scores.

Set C: Fifteen randomly mixed responses representing different combinations of language levels and analytic scores.

Qualifying set: Fifteen randomly mixed responses representing different combinations of language levels and analytic scores.

GUIDES AND TRAINING FOR INSTRUCTIONAL LEVELS 3/4, 5/6, AND 7

Preparation began when responses from the 2002 field tests were selected by the scoring leadership for the training paper selection meetings. These were held in Austin over several dates in December 2002 through January 2003. PEM staff met with members of the TEA Student Assessment staff to score and discuss the responses to the writing prompt.

After holistic and analytic “true” scores were assigned to the responses, they were given to the scoring directors and the analytic coordinator, who assembled the training materials. Following TEA approval of the scoring guides, the remaining prescored responses were used to assemble practice sets. The scoring guides were annotated, and the materials were photocopied for the readers.

TEA representatives observed holistic training and worked with the analytic coordinator. They also read selected responses for validity at Instructional Levels 3/4 and 5/6. Throughout the training and scoring

process, PEM staff served as on-site monitors. Immediately following scoring of the spring administration, readers scored the 2003 field-test responses. Scoring of the field-test prompts was completed in June.

TRAINING SET CONTENTS: HOLISTIC SCORING—INSTRUCTIONAL LEVEL 3/4

Written Composition Guide: Four annotated anchor responses at each score point in order from 1 to 4.

Sets A and B: Ten randomly mixed responses in each set representing score points 1–4. The number representing each score point varied from set to set.

Set C: Fifteen randomly mixed responses representing score points 1–4.

Qualifying sets I and II: Fifteen randomly mixed responses in each set representing score points 1–4. The number representing each score point varied from set to set.

Following is a breakout of responses used in training by score point:

25-1s
31-2s
18-3s
7-4s
<hr/>
81 total responses used in training

TRAINING SET CONTENTS: HOLISTIC SCORING—INSTRUCTIONAL LEVEL 5/6

Written Composition Guide: Four annotated anchor responses at each score point in order from 1 to 3. There were no score point 4 responses in the guide.

Sets A and B: Ten randomly mixed responses in each set representing score points 1–3. The number representing each score point varied from set to set.

Set C: Fifteen randomly mixed responses representing score points 1–3.

Qualifying sets I and II: Fifteen randomly mixed responses in each set representing score points 1–3. The number representing each score point varied from set to set.

Following is a breakout of responses used in training by score point:

33-1s
32-2s
12-3s
0-4s
<hr/>
77 total responses used in training

TRAINING SET CONTENTS: HOLISTIC SCORING—INSTRUCTIONAL LEVEL 7

Written Composition Guide: Four annotated anchor responses at each score point in order from 1 to 3. There were no score point 4 responses in the guide.

Sets A and B: Ten randomly mixed responses in each set representing score points 1–3. The number representing each score point varied from set to set.

Set C: Fifteen randomly mixed responses representing score points 1–3.

Qualifying set: Fifteen randomly mixed responses representing score points 1–3.

Following is a breakout of responses used in training by score point:

27-1s	
27-2s	
8-3s	
0-4s	
<hr/>	
62 total responses used in training	

TRAINING SET CONTENTS: ANALYTIC SCORING

The following analytic codes were used to explain why responses that were given a holistic score of 1 (scorable but unsuccessful) or 8 (nonscorable, except for blanks) received such scores:

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

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

SCORING SUPERVISOR TRAINING

INSTRUCTIONAL LEVEL K/1/2

For the 2003 SDAA project, the training for scoring supervisors was specific to the various instructional levels. For the Instructional Level K/1/2, the scoring supervisors were trained on all of 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 had 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 were essentially the same. The specific steps were as follows:

1. Present the item in the exact form in which it was administered.
2. Read and explain the introduction 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 somewhat developed rubric. Read and discuss each annotated somewhat developed paper.
 - d. 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, the scoring supervisors took the qualifying set.

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

INSTRUCTIONAL LEVELS 3/4, 5/6 AND 7

Scoring supervisors were trained specifically on the Instructional Level to which they were assigned. For Instructional Levels 5/6 and 7, the same scoring supervisors were used, but each scoring supervisor went through a separate training for the two levels. The model of the 3/4, 5/6 and 7 training follows.

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 had 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 were essentially the same. The specific steps were as follows:

1. Present the prompt in the exact form in which it was administered.
2. Read and explain the introduction 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 were responsible for working directly with the readers; consequently, it was necessary for them to understand all the qualifying sets.

TEA and Pearson monitors observed the scoring supervisor training and determined which of the candidates would best serve as scoring supervisors for the project. Those not chosen were retrained as readers and very often contributed as accurate readers during the project.

READER TRAINING

Before training, the readers were divided among the Instructional Levels K/1/2, 3/4, 5/6, and 7. All readers were trained using the scoring supervisor training model described above.

The readers assigned to the K/1/2 project were trained on all of 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 readers who scored the K/1/2 level went through the ePEN (Electronic Performance Evaluation Network) training since the students' responses were scored using the ePEN system. (For more information on the ePEN system, please refer to the section on the scoring of the TAKS short-answer items earlier in this chapter.)

PAPER FLOW AND RESOLUTION PROCESS

INSTRUCTIONAL LEVEL K/1/2

A scoring director supervised the day shift of readers for each instructional level. TEA and PEM project monitors were able to view reports electronically for all readers to ensure continuity between shifts. The review process could be done locally on-site or remotely via the internet from computers with secure access.

The logistics of the flow of responses in the scoring centers was carefully planned and carried out. The answer documents were sent to the PEM Tech Ridge facility in Austin, where they were scanned. The sections of the answer document were linked by a unique number printed on each page so that the score could be added to the student's record once scoring was complete. The responses were then given a unique ePEN identifying number. The ePEN number was 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, the readers had no knowledge of who the students were. The unavailability of identifying information on the papers helped ensure unbiased scoring.

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

Each reader read a response, selecting a score from a menu on the computer screen. An employee number that identified the reader was electronically attached to the response; only scoring supervisors, scoring directors, and project monitors could identify which reader read which response. The responses for K/1/2 received a single reading, but the scoring supervisors read behind the readers to ensure accuracy.

INSTRUCTIONAL LEVELS 3/4, 5/6, AND 7

Instructional Levels 3/4, 5/6, and 7 used paper scoring. The logistics of paper flow in the scoring centers was carefully planned and carried out. The answer documents were sent to the PEM Tech Ridge facility in Austin, where they were scanned. During the scanning process, the two lined pages on which students wrote their compositions were separated from the multiple-choice section of the answer document. The two sections of the answer document were linked by a unique number printed on each page so that the composition's score could be added to the student's record once scoring was complete. The writing pages were then assembled into packets containing 40 or fewer papers each. A packet header sheet was placed with the packet of papers, and the packet was 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 was 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, the readers had no knowledge of who the students were. The unavailability of identifying information on the papers helped ensure unbiased scoring.

The packets were then boxed by grade and shipped to the scoring center in Austin. Whenever a scoring room needed additional papers, an aide carried packets to the room. The room aide and scoring supervisors handled all paper flow in the scoring rooms.

Each reader independently read an entire packet of papers, writing his or her reader number on both the packet envelope and the monitor sheet. The reader recorded the scores for the papers on the monitor sheet, on which the identification numbers of the essays in the packet had been preprinted. The completed first-reading monitor sheet was removed from the packet envelope before the packet was given to a second reader. The reader number on the packet envelope identified the reader's team as well as the individual to ensure that the same packet would 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 were produced. Only readers identified as being above room average in the accuracy of their scoring were allowed to be resolution, or third, readers. Early in the project they were 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 supervisors' assessment of their "live" scoring. Later the daily reader status reports and validity reports were invaluable in identifying the readers whose scoring accuracy was above room average. Designated third readers were not allowed to score third readings exclusively. Rather, they were required to score at least two 40-paper packets daily so that sufficient data could be collected to monitor their scoring on an ongoing basis. Any third reader whose perfect agreement rate on the daily status report dropped was confined to performing first and second readings. Occasionally a fourth reading of a student paper was necessary. When this occurred, the fourth-reading monitor sheets were matched to the packets and given to scoring directors for scoring.

Responses requiring analytic scores were identified on an analytic monitor sheet and delivered to the analytic scoring room.

DATA-ENTRY PROCEDURES AND RESULTING REPORTS

The packet monitor sheets were scanned at the scoring centers, and the scores were transmitted to PEM in Iowa City. After the scores for the first and second readings of a packet had been scanned, the resolution monitor sheet (third-reading monitor) was produced. PEM transmitted the data for third-reading monitor sheets (as well as fourth-reading, analytic, and specialist monitor sheets) to the PEM Performance Scoring Center's printer. The monitors were then printed and delivered to the warehouse.

The data also produced project status reports that gave senior staff and scoring directors up-to-date information on the progress of the entire project at all scoring centers. These reports provided a wealth of information about the scoring patterns of individual readers.

INSTRUCTIONAL LEVELS K/1/2

After the score for the first reading of a response had been processed, the ePEN system held the score until the project monitor was satisfied with its accuracy; then it was uploaded as a final score.

INSTRUCTIONAL LEVELS 3/4, 5/6, AND 7

The data collected for first, second, third, and fourth readings produced project status reports that gave senior staff and scoring directors up-to-date information on the progress of the entire project at all scoring centers. These reports provided a wealth of information about the scoring patterns of individual readers. In addition to the number of responses read by each reader, the reports included 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 was judged to be incorrect; consequently, the reports had three adjacent score categories; 0/1, 1/2, and 2/3. These showed the number of times the reader's incorrect scores were higher and/or lower for each of the adjacent score categories. The final columns on the reader status reports gave the readers' distribution of score points—that is, what percentage of a particular reader's scores were 0s, 1s, etc. Accompanying the daily (or current) reader status report was the year-to-date report, which had the same information but was cumulative for the project as of that date.

SCORE APPEALS

PEM rescues any SDAA student response about which questions have been raised regarding the assigned score. Through a telephone call to the district contact person, PEM provides an individual analysis of the composition in question.

TAAS EXIT LEVEL: WRITTEN COMPOSITION SCORING

In the 2002–2003 school year, the TAAS exit level writing test continued to be offered to those students for whom TAAS is their graduation testing requirement.

The main administration of TAAS was in February 2003 (also known as the spring administration). In addition, three exit level retests were administered in the 2002–2003 school year. Two of these retests—those in October and July—were open to any exit level student who has previously failed one or more subject-area TAAS tests. The late April/early May retest is restricted to out-of-school students and those students who have enough credits to graduate but who have failed one or more subject-area TAAS tests.

All TAAS written compositions were scored at the Pearson Educational Measurement (PEM) scoring center in Austin. For each scoring session, TEA and senior PEM staff selected validity papers (see “Validity Packets”) and monitored training and scoring.

Fifty-four employees worked on the scoring of compositions for the main (February) administration of the test. Seventy-eight staff members worked on the October 2002 project, 19 worked on the May 2003 project, and 22 worked on the July 2003 project.

Scoring of the May 2003 TAAS retest took place within a one-week time frame. The daily schedule for this retest, as well as for the October 2002 and July 2003 retests, was the same as in spring 2003.

TRAINING MATERIALS

Each of the four TAAS exit level scoring projects required its own prompt-specific set of training materials.

GUIDES

There were a total of sixteen student responses: four annotated anchor responses representing each score point in order, from 1 to 4.

SPLIT SETS

Each split set contained four “close call” papers that defined the “line” between two score points. There was one split set for the 1/2 line, one for the 2/3 line, and one for the 3/4 line.

TRAINING SETS

Training Sets A, B, and C each contained ten randomly mixed responses representing score points 1–4. Training Set D contained 15 randomly mixed responses representing score points 1–4.

QUALIFYING SETS

Each of the three qualifying sets contained twenty randomly mixed responses representing score points 1–4.

These training materials included 133 responses. Following is a breakout by score point:

Exit Level (Spring)

29-1s

38-2s

38-3s

28-4s

133

The October 2002, May 2003, and July 2003 exit level retest scoring sessions used training sets with a similar composition, except that they contained slightly more 1s and slightly fewer 4s to reflect more accurately the types of responses the readers were more likely to see in those sessions.

TRAINING SET CONTENTS: EXIT LEVEL VERIFICATION SCORING

These training materials consisted of a guide using the 1s and 2s from the holistic guide and split set, along with ten-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[s] of the analytic readers and 1/2 score verification specialists, and TEA staff) rounded out the sets.

TRAINING SET CONTENTS: ANALYTIC SCORING

The following analytic categories were used to explain why responses that received a rating of 1 or 0 (nonscorable responses) were unsuccessful.

Holistic Score Point 1 Analytic Categories	Holistic Score Point 0 Analytic Categories
Lacks clarity	Off-topic response
Lacks language control	Indecipherable response
Lacks organization/structure	Insufficient response to specified task
Lacks support/elaboration	
Drifts from specified purpose	
Uses wrong purpose	
Drifts from specified topic	

Training materials consisted of an eight-paper guide and a four-paper split set (from the holistic guide and 1/2 split set) demonstrating the successful/unsuccessful holistic response line, a ten-paper “Analytics” guide, and an explanation of the analytic categories and the numerical system used to assign the appropriate category or categories to each response. In addition, readers received four ten-paper training sets representing the various analytic categories and allowable combinations of categories.

SCORING SUPERVISOR TRAINING

Effective reader training for the TAAS scoring sessions relied to a great extent on having knowledgeable, flexible scoring supervisors. The scoring directors depended on scoring supervisors to conduct small-group discussions with their teams; consequently, scoring supervisor training was critical to the success of the scoring effort.

Throughout their training, scoring supervisors (joined by the analytic coordinators and the score-verification specialist coordinator) were encouraged to ask questions and to discuss any problems they had 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 were essentially the same. The specific steps were as follows:

1. Present the prompt in the exact form in which it was administered.
2. Read and explain the introduction section of the scoring guide.
3. Present a “very 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. Read and discuss the 1/2 Split Set.
 - c. Read and explain the score point 3 rubric. Read and discuss each annotated score point 3 paper. Read and discuss the 2/3 Split Set.
 - d. Read and explain the score point 4 rubric. Read and discuss each annotated score point 4 paper. Read and discuss the 3/4 Split Set.
5. Score and discuss Training Sets A, B, C, and D.

After completing all the training sets, the scoring supervisors took the qualifying sets. Regardless of whether a scoring supervisor scored well enough on the first qualifying set to qualify, he or she still took the second qualifying set. Taking both sets was important, since scoring supervisors were responsible for working directly with readers; consequently, it was necessary for them to understand both qualifying sets. The third qualifying set was reserved for additional training as appropriate.

READER TRAINING

Only the most experienced TAAS exit-level readers were used for the relatively small (in terms of numbers of compositions scored) TAAS projects. Before training began, readers signed their contracts and nondisclosure forms, and TEA representatives made introductory remarks.

The scoring director discussed the prompt, introduced the guide, and then explained each score point to the entire group of readers. The readers took each of the training sets and the scoring director discussed each response with the readers. The readers were encouraged to ask questions to clarify papers with which they had had difficulty. TEA staff monitored this entire process.

Like scoring supervisors, readers had to demonstrate accuracy in their scoring before they could begin reading packets of responses. Readers were allowed three opportunities to qualify. Any reader unable to meet the standards set by TEA was dismissed.

Training of the analytic readers followed a similar pattern, except that the training was performed by the analytic coordinator.

EXIT LEVEL SCORE VERIFICATION

Since the spring 1992 exit level scoring session, TEA's contractors have used a score-verification procedure to further evaluate all responses that received a 1 during the holistic scoring process. A special team of readers are trained exclusively on the 1/2 line by using the 1s and 2s from the holistic guide and split set, along with ten-paper sets that make up the rest of the 1s and 2s from the holistic sets and qualifying rounds. Additional responses found in the live papers (approved by the scoring director, the coordinator of the analytic readers and 1/2 score verification specialists, and TEA staff) are selected to round out the training sets. If any response scored by a member of the specialist team is thought to be higher than a 1, it is read by the specialist coordinator. If the coordinator agrees, the response is then read by the scoring director. If the scoring director also agrees, the score is changed; if he or she disagrees, the response is read by the project monitor, who makes the final decision unless it involves an issue that should be brought to the attention of TEA. In that case the response is sent to TEA for a final scoring decision.

ONGOING ROOMWIDE TRAINING

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

One method was the scoring and discussion of sets of three to five papers each. The scoring directors started accumulating copies of papers that were typical close (or “line”) calls. The scoring directors reviewed these papers with senior scoring staff and then circulated them among scoring supervisors to ensure team-to-team consistency on these difficult decisions. Both shifts used these sets. Discussion of these sets sometimes occurred roomwide and sometimes in teams.

While scoring papers and spot-checking the accuracy of readers’ scoring, scoring supervisors were instructed to collect various types of problematic papers. These papers were 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 agreed on the scores of these papers, the sets were administered to the readers. Discussion of these papers was conducted roomwide. If individuals needed more help, the floating scoring supervisor worked with them.

Every Monday the scoring director reviewed the rubrics with readers and had them reread their anchor papers, emphasizing any area that appeared to be giving readers problems.

MONITORING OF INDIVIDUAL READERS

In addition to the ongoing training methods mentioned above, the scoring center employed a number of informal methods to identify individual reader scoring problems. Scoring directors and scoring supervisors relied on individual and small group retraining to ensure that readers were consistently applying the preestablished criteria when scoring. Scoring supervisors spot-checked and annotated reader packets throughout the project and then returned packets to the readers for their review. If necessary, the scoring supervisors would provide one-on-one assistance to a reader and discuss discrepant scores. Readers also flagged papers that were difficult for them to score. Scoring supervisors read these papers and then discussed each paper with the reader who had flagged it.

Early in the project, scoring supervisors closely monitored all readers, spot-checking according to the following: scoring trends identified from training results, reports of “true” score reliability, and daily reader status reports. The need to spot-check every reader decreased as it became clear which readers were consistently applying the scoring criteria and which needed additional support. At this point, scoring supervisors concentrated on readers who scored below 80% on the validity packets and/or who were below the room average on the daily reader status reports. They conducted 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 suspected that a certain reader might not be using the criteria properly was to obtain a regular packet that had been scored first by the floating scoring supervisor. Distribution of this type of packet was done routinely so attention would not be called to it as a training device. The reader’s scoring supervisor then compared the floating scoring supervisor’s scores with those of the reader. If there were a number of discrepant scores, the floating scoring supervisor or the reader’s scoring supervisor discussed 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 were retrieved and rescored by his or her scoring supervisor or by a reader at or above room average. The scoring supervisor then discussed with the reader the papers that had received discrepant scores. Any reader who could not be successfully retrained on the criteria was dismissed.

VALIDITY PACKETS

For the October 2002 and all 2003 exit level retests, time constraints required some procedural modifications. In place of validity packets, TEA and senior PEM staff signed-off on validity papers that were used in roomwide ongoing training exercises. At least one such exercise was given per week of the project. Since the May 2003 scoring session lasted only one week, no validity packets were used during that session.

ANALYTIC SCORING

Each composition that received either a rating of 0 or a score of 1 was evaluated analytically to provide information about the specific weaknesses that caused it to be unsuccessful. Analytic readers were trained on all the analytic features simultaneously. Papers that exemplify the range of unsuccessful compositions and that TEA and senior scoring center staff agreed on in advance were selected as training papers. The scoring director first read and discussed the guide with the analytic readers. The guide included eight sample papers that were chosen to represent a variety of analytic scores. The analytic readers read, scored, and discussed four additional sets of ten papers each. The readers began “live” scoring when they were able to demonstrate accuracy on all analytic categories.

NONSCORABLE RESPONSES

During holistic scoring, if a reader believed that a paper may be nonscorable, the paper was flagged for the scoring director to read and score. If the scoring director found it to be nonscorable, the second reading was performed independently by the project monitor. Nonscorable responses were then evaluated by the analytic readers.

PROCEDURES

PAPER-FLOW AND RESOLUTION PROCEDURES

A scoring director supervised the exit-level readers; because of the small number of readers needed for the exit-level test, only one scoring director was needed.

The logistics of paper flow in the scoring centers was carefully planned and carried out. The answer documents were sent to the PEM Tech Ridge facility in Austin, where they were scanned. During the scanning process, the two lined pages on which students wrote their compositions were separated from the multiple-choice section of the answer document. The two sections of the answer document were linked by a unique number printed on each page so that the composition’s score could be added to the student’s record once scoring was complete. The writing pages were then assembled into packets containing 40 or fewer papers each. A packet header sheet was placed with the packet of papers, and the packet was 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 was 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, the readers had no knowledge of who the students were. The unavailability of identifying information on the papers helped ensure unbiased scoring.

The packets were then boxed and shipped to the scoring center in Austin. Whenever a scoring room needed additional papers, an aide carried packets to the room. The room aide and scoring supervisors handled all paper flow in the scoring rooms.

Each reader independently read an entire packet of papers, writing his or her reader number on both the packet envelope and the monitor sheet. The reader recorded the scores for the papers on the monitor sheet, on which the identification numbers of the essays in the packet had been preprinted. The completed first-reading monitor sheet was separated from the packet envelope before the packet was given to a second reader. The reader number on the packet envelope identified the reader's team as well as the individual to ensure that the same packet would 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 were produced. Only readers identified as being above room average in the accuracy of their scoring were allowed to be resolution, or third, readers. Early in the project they were 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 supervisors' assessment of their "live" scoring. Later the daily reader status reports and validity reports were invaluable in identifying the readers whose scoring accuracy was above room average. Designated third readers were not allowed to score third readings exclusively. Rather, they were required to score at least two 40-paper packets daily so that sufficient data could be collected to monitor their scoring on an ongoing basis. Any third reader whose perfect agreement rate on the daily status report dropped was confined to performing first and second readings. Occasionally a fourth reading of a student paper was necessary. When this occurred, the fourth-reading monitor sheets were matched to the packets and given to scoring directors for scoring.

Responses requiring analytic scores were identified on an analytic monitor sheet and delivered to the analytic scoring room.

DATA-ENTRY PROCEDURES AND RESULTING REPORTS

The packet monitor sheets were scanned at the scoring center, and the scores were transmitted to PEM in Iowa City. After the scores for the first and second readings of a packet had been scanned, the resolution monitor sheet (third-reading monitor) was produced. PEM transmitted the data for third-reading monitor sheets (as well as fourth-reading, analytic, and specialist monitor sheets) to the PEM Performance Scoring Center's printer. The monitors were then printed and delivered to the warehouse.

The data also produced project status reports that gave senior staff and the scoring director up-to-date information on the progress of the entire project. These reports provided a wealth of information about the scoring patterns of individual readers. In addition to the number of responses read by each reader, the reports included 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 was judged to be incorrect; consequently, the reports had three adjacent score categories: 1/2, 2/3, and 3/4. These showed the number of times the reader's incorrect scores were higher and/or lower for each of the adjacent score categories. The final columns on the reader status reports gave the readers' 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 was the year-to-date report, which had the same information but was cumulative for the project as of that date.

SCORE APPEALS

PEM rescors any TAAS written composition about which questions have been raised regarding the assigned score. Through a telephone call to the district contact person, PEM provides an individual analysis of the composition in question.