

FCAT Science Performance Item Scoring Process

2006Aug07

Rick Tully

The FCAT Science is an incredibly complex attempt to measure the performance of Florida students in respect to the Sunshine State Standards in Science and to monitor improvement over time. Florida educators have been involved with each step of the process of the development, review, and scoring of the test. Unfortunately, not every teacher has had the opportunity of involvement. Consequently there is often misinformation and misunderstanding regarding various phases of the testing process.

What follows is an outline of one part of that process, the handscoring process. Every student response to a performance item, that is every written response, must be scored by a human, not a machine. These humans must assign scores in such a way that all scores are appropriate, consistent, and accurate. All readers (scorers) must be applying the same criteria as they assign scores to the papers they read.

Students, parents, teachers, and the public deserve a system that they believe is fair and on which they can rely. Having been trained in the scoring of two specific performance items, I am confident of the academic and intellectual integrity of this part of the test. Hopefully, the discussion that follows will give the reader some similar measure of confidence in this process.

Overview of Performance Item Scoring Process

- 196,000 papers were scored in 2004 for FCAT Science.
- Scoring occurred at 3 sites involving 5 sessions: Florida (1 session), Texas (2 sessions), and Iowa (2 sessions).
- Approximately 100 scorers were involved at each session.
- There were approximately 23 days of scoring.
- A limited number of different items were scored at each scoring site (2 items in Florida, 1 SR & 1 ER).
- Every individual student response was independently scored by at least 2 trained readers.
- The scores of the readers should be exact (the same) or adjacent (adjacent numeric scores).
- Discrepant scores must go to resolution.

Discrepant Score Resolution

- All scores on short-response (SR) items must be exact.
- Differing SR scores must go to resolution.
- In the event of differing but adjacent scores for an extended response (ER) item, the higher score will be applied (in Reading, Mathematics, and Science).
- Differing, non-adjacent ER scores must go to resolution.
- Resolution: A scoring supervisor will read the response and provide a third independent score. Assuming this aligns with one of the prior scores, the aligning score will be applied.

- Papers still needing resolution are referred to the scoring director and FCAT team leader.

Scoring Staff

- Florida Department of Education (DOE) Team: On site for training and first week of scoring.
- Pearson Educational Measurement, Performance Scoring Team: Manager, FCAT Team Leader, FCAT Lead Scoring Director for each subject.
- Pearson Scoring Site Staff: Site director, office staff, technology support staff
- Pearson Scoring Team:
 - Scoring Director
 - Assistant Scoring Director
 - Floating Supervisors (ad hoc tasks)
 - Supervisors (1 per team)
 - Readers (10-15 per team, ~100 total)

Reader Qualifications

- Bachelor's Degree.
- Placement test for FCAT Science.
- Participation in Reader Training
- Successful completion of Qualifying Exam.

Reader Training

- All reader candidates receive 2 days of training.
- The training is designed to provide the readers with the rubric and scoring guidelines created by the Florida Rangefinding Committee for each of the two items.
- Readers must achieve the “mindset” of the Florida educators on the Rangefinding Committee; they are to “embrace the rubric”.
- Readers are instructed repeatedly, “The state sets the standards and we apply the standards. There is no reason to disagree with the standards.”
- Holistic scoring is used rather than analytic scoring.
- With holistic scoring, the total response is considered when arriving at a score.
- There is no analytic rubric that specifically identifies necessary components of a response to a particular item.
- Rather than a rubric specific to each item, there are sets of actual student responses established as Rangefinding Sets to which the scorers frequently refer.
- When scoring holistically, nothing is irrelevant.
- Scoring is a matching process, new student papers are matched to papers that have been prescored and are contained in the Rangefinding documents.

Scoring Rubrics

- Generic rubrics have been established for both the SR and ER item types (please see the accompanying rubrics).
- These rubrics set the broad qualities that are necessary for each scoring level.

- Reference Sets of student responses have been created by the Rangefinding Committee to help the scorers identify the different levels of student responses.

Reference Sets of Responses

- Rangefinding Set –
 - The primary reference set of papers is the Rangefinding Set;
 - These papers have all been scored by the Rangefinding Committee and establish the standard that is to be applied throughout the scoring of any particular item;
 - It includes papers selected to demonstrate the range of responses acceptable at each score level;
 - Readers see these first without the Rangefinding scores and are asked to assign scores from their own interpretation of the generic rubric (Calibration Scoring).
- Horizontal Sets –
 - During training 2 Horizontal Sets of 10 papers each are used to practice scoring and the application of specific point values;
 - The reader independently scores each paper then compares those scores to true scores as assigned by the Rangefinding Committee (Practice Scoring);
 - These additional papers can be added to the reader’s training notebook as an aid to scoring.
- Qualifying Sets –
 - To qualify as a scorer, each candidate must complete 2 Qualifying Sets of 10 papers for each item;
 - An overall accuracy of 70% must be achieved on ER items and 80% must be achieved on SR items (Qualification Scoring);
 - These sets may also be added to the reader’s training notebook.
- Final Qualifying Set –
 - Candidates failing to qualify after the first two Qualifying sets receive additional training and then take a final Qualifying set;
 - This set is not returned to the reader.
- Training Materials –
 - The aggregate training materials will include 55 different student papers and will define the qualities that are appropriate for each score level.

Reader Bias

- Readers are instructed to disregard specific irrelevant qualities and other interferences when arriving at individual student scores.
- The scoring system is designed to manage for consistency and accuracy.
- Some potential scorer bias factors:
 - Experience with scoring other tests or other items
 - General appearance of a student’s response
 - Scorer’s knowledge of the topic
 - Reaction to style of response
 - Reaction to content
 - Fatigue of the scorer

Qualifying Procedure

- All candidates must qualify before doing any live scoring.
- Qualification requires the reader to score 10 of each of the 2 different items.
- A second set of 10 items of each type must also be scored.
- The candidate must have a total average of 70% correct score on the ER item and 80% correct score on the SR item.
- Candidates failing to accurately score at the necessary level of correctness then receive additional training in the interpretation and application of the scoring rubric. They will then have an additional opportunity to qualify.
- Candidates failing to qualify after this third attempt are released.

Electronic Image-Based Scoring

- All performance item responses are scanned and distributed electronically for reader scoring.
- Item responses are redistributed randomly for the second reading.
- Responses requiring additional reading to resolve discrepant scoring are routed to supervisors for resolution. These are scored “blind” – the supervisor has no knowledge of first and second scores.

Quality Control

- Every individual student response is independently scored by at least 2 trained readers.
- Items requiring score resolution are read by a supervisor.
- All scoring is continuously monitored by the Electronic Performance Evaluation Network (ePEN)
- Consistency of scoring (inter-rater reliability) is checked at least daily.
- Supervisors will periodically back read papers already scored by readers and will monitor for discrepancies.
- Throughout the day, validation items (items previously scored by the Rangefinding Committee) will be mixed with live items to monitor for accuracy.
- Validation items and live items will not be identifiable to the scorer.
- Each individual scorer is monitored for accuracy at least twice per day (~1 in 40 items).
- Daily reports are generated for Pearson and DOE review.
- Calibration sessions (retraining) are conducted regularly for large groups and as needed for individuals.

Test and Item Security

- Site security is closely maintained.
- Individually identifiable scoring notebooks are maintained by each reader and are not removed from the site.
- All readers agree to and sign confidentiality statements.

Student Response Alerts (Child Alerts)

- Some student writing reveals problems (abuse, neglect, threats, cheating).
- Readers alert scoring directors.
- Demographic information is retrieved.
- DOE Team Leader reviews paper and demographic information.
- Notification sent to school district for investigation.
- Relatively low incidence of alerts, ~150/600,000 papers in FCAT Writing.
- Even lower alert rate with FCAT Science, likely due to the more specific nature of the SR and ER items.