BUROS
CENTER FOR TESTING

2013 Audit III Report:
Scoring of the FCAT 2.0 Writing Assessment

Prepared by:

Kurt F. Geisinger, Ph.D.

Anja Römhild, M.A.

Robert A. Spies, Ph.D.

Stephen G. Sireci, Ph.D.

Patrick M. Irwin, Ph.D.

Consultants to the Florida Department of Education

May, 2013

With questions or comments, please contact:
Kurt F. Geisinger, Ph.D.
kgeisinger@buros.org
(402) 472-6203

Introduction

As part of an annual review of selected components of Florida's statewide assessment system, the Buros Center for Testing (Buros) conducted a targeted review of the FCAT 2.0 Writing handscoring process from February through April 2013. The review activities included a review of the *Handscoring Specifications* document and operational handscoring statistics, and extensive on-site monitoring during seven 3-day site visits. A site visit was conducted to observe the document preparation and scanning of the FCAT 2.0 Writing test documents at the Pearson scanning facility in Iowa City, Iowa from March 4 to 6. Two site visits were made to each of the three FCAT 2.0 Writing scoring centers in Jacksonville, FL (Grade 4); Auburn, WA (Grade 8); and Tucson, AZ (Grade 10) to observe the initial scorer candidate training from March 12 to 14, and subsequently to monitor ongoing operational scoring processes during March and April. Buros also participated in weekly (February 6 to 27) and daily conference calls (March 4 to April 29) with staff from Florida Department of Education (FDOE), Florida's Test Development Center (TDC), and Pearson. This report summarizes our observations and findings concerning the quality and integrity of operations for the scoring of FCAT 2.0 Writing responses and concludes with several recommendations and comments for FDOE to consider. Individual site visit reports are provided in the Appendix.

Test Document Preparation and Scanning

Buros conducted a site observation at Pearson's Iowa City, Iowa facility March 7-8 to observe the scanning of FCAT 2.0 Writing Test. As stated in the scanning report (Audit I) Buros found Pearson to be operating at or above industry standards. Buros did, however, have a question for FDOE related to transcribing the Braille responses. The current process requires Pearson to send the Braille documents to an external contractor to transcribe the Braille to print

format. In some cases the schools sent a print and Braille version of the students' work, but Pearson was still required to send the Braille work to be transcribed once again. It may be worth FDOE investigating the effort it would take the schools to send to Pearson both the print and Braille versions of the students' work for scoring. If FDOE still requires sending the documents to an external contractor, they could verify the two documents (print and Braille) instead of transcribing, or the transcription step could be completely eliminated. Pearson would still need to modify the documents to fit on the student answer sheets, but this step could save time (sending the documents to be transcribed) and resources. The FDOE contractor checklist for site visits is included in the Appendix.

Scoring Time Frame and Timeline

Operational scoring activities for FCAT 2.0 Writing were scheduled to begin March 4, 2013 and to be completed by April 26. Training of scoring supervisors occurred during the first week of the schedule from March 4 to 6, followed by the initial training of scorer candidates during the second week from March 12 to 14. Full operational scoring began around March 21 when the first large group of scorer candidates had passed the qualifying rounds. The scoring of Grade 8 and Grade 10 essays was completed within the scheduled scoring window. Grade 4 scoring required an additional full day plus weekend overtime to complete. Overall, the Pearson staff in collaboration with the TDC was proactive throughout the project to ensure timely completion of all activities. Pearson and TDC monitored completion rates on a daily basis throughout the project and implemented steps to stay on track. Early in the scoring process, additional training waves were held at the Jacksonville, FL and Tucson, AZ sites to increase the number of qualifying scorers. In addition, overtime was offered at all sites to highly qualified scorers and supervisors. With those measures in place, Pearson was able to stay within schedule.

Scorer Recruitment and Training

As in recent years, the scoring centers experienced an initial shortfall of scorer candidates able to meet the stringent qualification standards for FCAT scoring. Although all sites met or exceeded the target number of scorer candidates invited to the training, lower than expected numbers passed the qualifying rounds. For Grade 8 scoring in Auburn, WA, the qualification number was close to the original target of 172 with approximately 67% or 155 of the 230 candidates who attempted the qualifying scoring being able to pass. Fewer candidates qualified for the Grade 4 (96 of 214) and Grade 10 (103 of 283) scoring, which prompted a second wave of training potential raters for those grades during the fourth week (March 24 to 28) of the scoring window. The second training wave added 22 qualifying scorers to the Grade 10 pool, an addition that was deemed sufficient to keep project completion on track. For Grade 4, the somewhat unusual decision was made to conduct a third training wave at the Auburn, WA site to which 45 Grade 8 scorers were recruited from among the Grade 8 scorers. The third training wave (April 8 to 10) was very effective with a qualifying rate of over 75%. Thus, this "third wave" added 33 scorers and 3 supervisors to the pool. The 11 scorers who did not qualify for Grade 4 scoring went back to scoring Grade 8 after a brief period in pseudoscoring. Buros supports the actions that Pearson took in this regard to complete the scoring in a timely and appropriate fashion.

It is difficult to hypothesize what factors cause one training to be more successful than another as all training waves appear to have been conducted consistently, with the same quality materials, and by the same experienced scoring directors. It is likely that the Auburn site was more successful this year in recruiting experienced scorer candidates than the Jacksonville site. Fluctuating local economic conditions no doubt influence the availability of qualified scorer

candidates each year, and Pearson has little control over such factors. Given these circumstances, Buros believes Pearson has attempted to recruit scorers in an optimal fashion by using effective advertising strategies to reach a maximum of potential candidates.

The training process is generally highly standardized across the three sites and consists of a morning orientation that addresses logistical issues and provides an overview and context for the scoring project. All candidates were required to sign nondisclosure forms and were given instructions in a serious fashion on how to keep test materials secure. The lead scoring directors discussed the training materials, the training process, and introduced the quality management plan that details the qualification and maintenance requirements for operational scoring. Before introducing the grade-specific prompt and scoring rubric, instructions were given on reader bias. Prompt-specific training followed that introduced the prompt and the range of allowable interpretations, the holistic scoring method used to score FCAT 2.0 Writing, and gave a thorough explanation of the rubric and the four rubric elements (Focus, Organization, Support, and Conventions). Scorer candidates were also given instructions on how to approach the scoring training and scoring process successfully with reminders of what to do and what not to do. Buros found the segment that covered successful scoring strategies, such as reading supportively and considering the intent of the student writer, quite valuable. It was our observation that the scoring leaders reminded candidates of those strategies during training, but emphasizing this strategy even more may prove useful.

At the conclusion of the morning training, the lead scoring directors introduced the prompt-specific set of anchor papers consisting of 3 student essays per score point for a total of 18 essays. In most cases, the 3 essays for a given score point indicated the range for that given score point. For example, the three anchor papers for a score of “4” included a low 4, a modal 4,

and a high 4. The Grade 8 set included 17 essays with only two examples at score point 6. This minor reduction was not a concern for Buros because this occurred at the highest score point where suitable examples of student work can be difficult to find. Overall, the selection of anchor papers appears to have been performed with great care and seemingly resulted in clear examples of student work exhibiting a logical progression of student performance.

Each anchor paper was accompanied by a set of annotations that describe the rationale for the assigned score point in terms of the four rubric elements. Anchor papers and their annotations were read out loud by a scoring director who also provided additional commentary and clarifications. Once the complete anchor paper set had been introduced and discussed, candidates were given their first opportunity at essay scoring. The first three practice sets each consisted of five essays representing a limited score point range (1 to 3, 4 to 6, and 2 to 4). Candidates were informed of the score range in advance. The final two practice sets were longer and represented the entire range. After each practice round, score sheets were collected by supervisors who returned them to candidates with the percent of agreement (perfect and perfect plus adjacent) recorded on the sheet. Each practice paper set was then discussed in plenum by the scoring directors. These scoring directors had prepared notes with discussion points similar to annotations at their disposal to guide the review. During these discussions, questions from candidates were always encouraged.

Although the overall training process can be quite lengthy and repetitive at times, we found that scoring candidates generally remained attentive and alert, interested in the student essays, asking appropriate questions, and taking notes throughout. They appeared motivated to succeed at this task. Scoring directors responded competently and willingly to candidates' questions and kept an appropriate pace. Overall, Buros found that the entire training process was

very well organized and executed with great success. As an additional measure for improvement, Buros suggests that Florida and Pearson consider developing and administering a survey of candidates' experiences at the end of training. Such a survey should be administered to both successful and unsuccessful candidates and should focus on identifying the training elements that they found to be effective and those not so effective. The results should then be used to make further improvements to the current training format.

This year, a modification was introduced to the training process that resulted in extended practice opportunities for scorer candidates. After receiving instructions on the scoring process and completing five practice sets of student essays, candidates were given an additional prequalification set that mimicked the length of the qualification sets. In addition, candidates went through a minimum of 4 hours of pseudoscoring in the ePEN system allowing them to practice essay scoring on the computer and gaining familiarity with the tools available in ePEN. During the second and third wave trainings, candidates were no longer taking the prequalification set on paper. Instead they spend several hours pseudoscoring in ePEN where they practiced scoring of essays from the prequalification sets and from several calibration paper sets. Unlike previous years, the three qualification sets were also administered within ePEN. This change offers the advantage that scoring conditions during qualifying are more closely aligned with those during operational scoring. Buros is very supportive of the decision to include additional practice scoring opportunities in ePEN and to administer the qualification sets in the computer environment. Although ePEN may make scoring slightly more challenging due to the fact that scorers cannot revise already-scored essays, which is not the case when scoring on paper, the benefits gained through increased fidelity with operational conditions clearly outweigh this concern.

Operational Scoring Process

All three scoring sites met or exceeded the project-wide quality criteria for interrater reliability (IRR), set at 60% for Grades 4 and 8 and 55% for Grade 10, and for validity agreement, which is set at 70% across all three grades. Both the Grade 4 and Grade 8 scoring sites met the IRR target of 60% and each exceeded the requirement for validity agreement with 76%. Scoring for Grade 10 was similarly successful exceeding both IRR (57%) and validity agreement (78%) targets. It is particularly noteworthy that these targets were reached fairly early in the scoring process and were sustained throughout – an indication that the quality of the scoring process was very stable.

The validity agreement rate is also used as a criterion to evaluate and monitor the scoring accuracy of individual scorers. Throughout the scoring project, scorers are expected to maintain a minimum validity agreement rate of 60% exact score match and a minimum of 90% exact or adjacent score agreement. Scorers who fall below these standards are issued a warning and receive remediation training. If scoring accuracy does not improve within the next 10 validity papers, scorers receive a 10-paper scorer exception set, which they must pass with a 70% exact and 100% exact/adjacent score agreement. Scorers who do not pass the scorer exception set are released from the project and their scores are reset. Buros believes this latter score adjustment is a strong statement in terms of how concerned both Florida and Pearson are about the accuracy of scoring.

Approximately one in every seven essays read by a scorer is a validity paper. Toward the end of this year's scoring project, the rate of insertion of validity papers in the essay queue was extended to 1 in 15 essays after it was apparent that the validity agreement rate had been consistently above the quality standard of 70% for all three grade level assessments. It is the

impression of Buros that the change in the insertion rate is done foremost as a measure to alleviate some of the effort required to update continually the pool of validity papers with new papers; some of them are prepared with annotations. However, Buros is concerned that reducing the rate of insertion of validity papers makes it more difficult to monitor overall scoring accuracy, even when it can be assumed that scoring accuracy has been stable for some time. Research has shown that the further scorers are from training, the more likely they are to "stray" from the scoring rubric. Buros therefore suggests that the validity paper insertion rate remain constant throughout the project. Such a recommendation, of course, must be balanced against the need for timeliness in completing the scoring process, and Buros understands that Florida must consider this recommendation against that factor.

A second concern regards the frequency with which the validity paper pool is updated. During this year's scoring, some validity papers remained in the queue for too long and were consequently read more than twice by individual scorers. FDOE staff were generally very diligent about monitoring the total number of validity paper reads and Pearson responded promptly by removing papers when concerns were voiced. However, these measures could not altogether prevent that some papers exceeded the acceptable number of reads by individual scorers. Perhaps a more automated process could be implemented in ePEN in the future, which creates alerts or even prevents individual scorers from reading a validity paper more than a specified number of times.

Besides the validity agreement rate, Pearson uses a number of additional quality monitoring and training devices to ensure that scorers use the scoring rubric appropriately and with accuracy. Scoring supervisors monitor their scorers' interrater reliability (exact or adjacent score agreement) and backread approximately 5% of scored essays. Scoring directors monitor

the work of supervisors and scorers alike. They also identify specific calibration needs for the total pool of scorers and for smaller groups of scorers requiring targeted training. As a first activity during each day of the scoring process, focused anchor paper reviews were conducted by the scoring directors. In addition, calibration papers were administered via ePEN or on paper to all scorers or to smaller targeted groups. This year, scorers seemed to respond positively to one-on-one training with their scoring supervisors and scoring directors. As a result, scoring directors were quick to implement it as a regular training tool for providing targeted training where needed. Overall, Buros commends Pearson and its scoring directors for their exceptional efforts of providing quality training and instruction to the scorers throughout the project. It is clear that those efforts greatly contributed to the high scoring accuracy and consistency rates achieved at all three scoring sites.

During the reporting of the calibration set results on the daily conference calls, Buros was made aware that in a few rare instances a large majority of the scorer pool – in one case as many as 70% – converged on a score point that was different from the pre-assigned calibration score. For those rare instances, Buros suggests that the calibration paper be given a second review by one or more content specialists in order to determine if the pre-assigned calibration score is, in fact, accurate. It could also be useful to specify the conditions under which such a review is triggered, for example, by setting a threshold of 65% of the entire scorer pool before a second review is initiated.

Scoring Conditions and Security

As in previous years, all operational scoring takes place within the environment of the ePEN system, which is loaded onto desktop or laptop computers. During the site visits, Buros observed that scoring conditions appeared to be somewhat more optimal when scorers worked on

desktop computers (in Tucson) rather than on laptops due to the larger screen size available for viewing the essays and for using the ePEN tools. To the extent that is practical and fiscally feasible, Buros recommends that the full-size monitors be used on scorers' work spaces instead of laptops.

All three scoring centers as well as the two off-site scorer training locations were set up in spacious facilities that allowed for quiet and pleasant work environments. All sites provided separate break rooms to scorers. Entrance ways and exits were continually monitored by the scoring center site staff who also enforced the security procedures at the site. All scoring personnel as well as Pearson project team members and visitors were required to wear id badges while on the premises. Sign-in and sign-out procedures ensured that only authorized individuals entered the sites. During scorer training, additional precautions were taken by keeping test materials securely locked overnight and by requiring nondisclosure agreements from all scorer candidates. Overall, Buros found that the security procedures were fully implemented and enforced ensuring the security of the test materials and test scores. Buros did not evaluate the security of the ePEN system, however.

FCAT 2.0 Writing Scores

In 2013, the overall performance of Florida's students on the FCAT 2.0 Writing assessment improved when compared to the 2012 administration of the FCAT Writing. The percentage of students scoring at or above 3.5 increased by 2 percentage points in Grades 8 and 10 and increased by 9 percentage points in Grade 4. Similar improvements within grades are also seen when the percentage of students scoring at or above 4.0 is considered (increases of 2 percentage points in Grades 8 and 10, and 10 percentage points in Grade 4).

Buros considers the observed across-year improvements in scores to be well within the range of score fluctuations that can be expected of assessments that are human scored. Furthermore, it is probable that an increase from 45 minutes to 60 minutes of allowable time to respond to the writing prompt contributed to the improved writing of the students. No other changes to the operational scoring process were made as far as Buros was able to observe, and the same stringent scoring criteria were implemented as in 2012. Nevertheless, the equating of essay scoring year-to-year is not possible in the manner it is with tests composed of multiple-choice questions. A few years ago, scores on the Writing Test dropped a few points from one year to the next. This year's increase, of course, is a more positive result than such a decline, but the state of the art and science of essay scoring is simply not at such a degree of accuracy that the Florida Department of Education can declare that writing has improved just because scores have. It is possible to investigate the impact of the 15-minute time extension on student performance in 2013 through a small research study in which the score distributions obtained from the field test administration and from an equally sized random sample from the state population are compared. Buros would recommend such a study as providing information that might be useful for setting future time limits. Because the field test administration of the 2013 prompt was conducted without the extra 15 minutes, any observed score differences would likely be attributable to the time extension, if the two groups are truly comparable.

Procedures for Exceptional Papers

Buros has observed in their operational procedures and in practice that the Florida DOE has developed procedures whereby essays written on specialized paper by students with disabilities can be read and scored as part of this process in a transparent manner. Buros believes that such a practice is exactly in keeping with the best intents of No Child Left Behind and Buros commends

the State for this practice. Buros also notes that procedures have been developed whereby a scorer who encounters an essay written by a troubled youth (e.g., one considering suicide) is given special attention and the identity of the individual student is quickly brought to the attention of the Florida DOE. Buros also commends this process as one that helps safeguard the schoolchildren of the State. Buros expects that essays that demonstrate a contemplation of violence would be similarly identified.

Conclusions and Recommendations

Now in its fourth year of conducting an annual review of the handscoring process of Florida's statewide writing assessment, the Buros Center for Testing continues to be impressed by the dedication and professionalism of the staff from the Florida Department of Education, the Test Development Center, and Florida's contractor Pearson to ensure that all Florida students are accurately and fairly assessed on their writing ability. We found that the scoring process was effectively organized, maintained high standards of scoring accuracy, and generally met the standards of best practices in the field. Overall, there are only a few minor recommendations and suggestions Buros has to offer that may be helpful in improving individual elements of the scorer training and quality monitoring of the scoring process.

<u>Recommendations for document scanning</u>:

1. Consider transcribing Braille test documents at the Pearson scanning facility in order to save time and create efficiency for document transfer and storing.

<u>Recommendations for scorer candidate training and qualification</u>:

2. Consider making instruction on scoring strategies a more frequent feature of the training process. Examples of scoring strategies are to read supportively and to

consider a student's intent. It may also be helpful to discuss unproductive scoring strategies.

3. Continue using the ePEN system for practice scoring and for the delivery of the qualification papers sets. Using the ePEN environment allows scorer candidates to become familiar with the ePEN tools during practice and to experience the scoring process as it occurs in operational conditions.
4. Consider the use of a survey to obtain feedback from successful and unsuccessful scorer candidates concerning effective and less effective training elements.

Recommendations for the operational scoring process:

5. Consider implementation of an alert system that tracks the number of reads of validity papers by individual scorers and creates alerts when the number of reads exceeds a specified maximum. Such a system should also be accompanied by increased efforts to keep the pool of validity papers continually updated with new papers.
6. Keep the validity paper insertion rate constant for the duration of the scoring project.
7. Consider implementing a review process for calibration papers that have been assigned a score by a majority of scorers that is not the pre-assigned calibration score.
8. To the extent practicable, use full-size monitors on scorer's workspaces to ensure conditions for viewing and using the ePEN tools are optimal.

Appendix

1) Site visit 1: Document preparation and scanning (Iowa City, Iowa, March 4-6, 2013)

2) Site visit 2: Scorer Training (Jacksonville, FL, March 12-14, 2013)

3) Site visit 3: Scorer Training (Auburn, WA, March 12-14, 2013)

4) Site visit 4: Scorer Training (Tucson, AZ, March 12-14, 2013)

5) Site visit 5: Operational Scoring (Jacksonville, FL, April 2-4, 2013)

6) Site visit 6: Operational Scoring (Auburn, WA, April 16-18, 2013)

7) Site visit 7: Operational Scoring (Tucson, AZ, April 17-19, 2013

FCAT 2.0 CONTRACTOR CHECKLIST

CONTRACTOR: **Pearson**	LOCATION: **Iowa City, IA**
ADMINISTRATION: **Spring 2013 FCAT 2.0 Writing**	DATES: **03/07/2013 –- 03/08/2013**
CHECKLIST COMPLETED BY:	Patrick Irwin
STAFF ASSISTING:	Brian Chase, Toby Trail, Curt Miller
REVIEWED:	Document preparation, scanning, and editing processes

TASKS/ACTIVITIES	RFP REF	Reviewer Initials and Date	COMMENTS
REVIEW DOCUMENTS			

Review Disaster Recovery Plan:

Document:

Pearson Technology Global Services Iowa City Data Center Incident Response and Recovery Plan Version 2012.1 Release date 10/8/2012

The document is Iowa City specific and includes 229 pages of detailed information on how to proceed if there is an actual disaster. The reason it resides in Iowa City is due to the data center also being housed in Iowa City. The document is very specific to how Pearson would proceed if a variety of disasters were to happen.

Review procedures to safeguard Personally Identifiable Information Section 7.11, Appendix H

As part of their continual efforts to improve in protecting Personally Identifiable Information Pearson is taking the appropriate steps towards obtaining the ISO-27001 (Information Security Standards) and BS-25999 (Business Continuity Management) certifications. As part of their operational excellence initiative Pearson's State Services Division is working towards obtaining the ISO-9001 (Quality Management System) certification.

Appendix H Test Security Requirements, Statute, and Rule

Restrictions on Printing for Security Purposes

1. All test negatives and plates must be maintained under lock and key by the printing supervisor.
2. Unauthorized personnel must not be permitted access to the test negatives, plates, or copies.
3. All plates and negatives must be destroyed by the contractor upon completion of this contract.
5. All press pull-ins, trim, and waste material must be shredded at the end of each day's press run by a person authorized to do so by the contractor.
6. Each production run must be made under close supervision of the printing supervisor

Florida Test Security State Board of Education Rule: 6A-10.042 Maintenance of Test Security

(a) Test questions shall be preserved in a secure manner by individuals who are developing and validating the tests. Such individuals shall not reveal in any manner, verbally or in writing, the test questions under development.

Any communication within Pearson is done on a secure network and any communication with FDOE is done on the FSTP secure file transfer protocol Florida has in place. All items and tests are stored on a secure network and database.

(g) Each person who has access to tests or test questions during the development, printing, administration, or scoring of the tests shall be informed of specifications for maintaining test security, the provisions in statute and rule governing test security, and a description of the penalties for breaches of test security.

Pearson has all employees sign a confidentiality agreement upon hiring and security issues are also covered in the Pearson employee handbook.

OBSERVE CHECK-IN PROCESS			

Observe to-be-scored documents being processed (removed from boxes, de-humidified, palletized, etc.) Sections 4.1, , 5.3, Appendix B

Documents:

IC REC Dock Sort Spec FCAT 2.0 Spring 2013 Writing (Document Number 49260640 Version-5 Pages-3)
Data Prep Project Information Requirement Spring 2013 Writing 620-169 Version-1 2/6/2013 3:43 PM
FCAT 2.0 Spring 2013 Writing 620-169 What if? Resolution Version-1 2/6/2013 3:45 PM
Disposition of Additional Materials

FedEx truck arrived and all the pallets were forklifted from the truck, all boxes were removed from the original pallet, scanned in and re-palletized and shrink wrapped. The pallet was stored in a holding area until the Data Prep department was ready for the materials. The boxes were unloaded from the pallet and opened up and placed on carts by Data Prep Department. The carts were then sent to the dehumidified room for a minimum of 8 hours.

Review Drop Stack Procedure Appendix B

Document:

AU & IC Operations Drop Stack Procedure (66551877 Version-1 Pages-1)

Very simple document describes how to handle a dropped stack. The procedures to follow when a stack is dropped are outlined in the document including notifying a regular full-time employee and the Data Prep supervisor.

Review scanner calibration frequency reports, Blue dot runs Appendix B

Documents:

AU-IC SCN Scanner Quality Check Report WI – Has the program Specs and Requirements listed. For writing 166 Program is 958 and SSN 003.
PUMS (Performance Utilization Management System) Report – Provides quality control checks on ALL scanners scanning the FCAT writing project.
Scan Program Requirements FCAT Writing Spring 2013 / 620-169
Scan Program Requirements FCAT Writing Spring 2013 VOIDS P620166S (Security Barcode Scanning)
Inbound label report – Each box is scanned and tracked based on school and district. This occurs when the boxes arrive from FedEx and are re-palletized.

WIP Report (Work In Progress) – Pearson assigns a bar code (unique number) to each answer sheet/test sent out. This is tracked in Oracle. Pearson is able to account for every answer sheet/test in the WIP system. Pearson knows what has been sent out, but is unable to account for or see what has been returned until the answer sheet/tests have been scanned. When a returned answer sheet/test is on the dock, or in Data Prep, dehumidifying, or cutting they are unable to track the answer sheet/test. Once the documents are scanned they know exactly where the information is located or being stored.

Security and Resolution Report – As mentioned above each answer sheet/test form (writing answer sheet) sent out has a unique barcode on the document. The document is assigned to a school and/or a district. All documents are to be returned even if the answer sheet was never used. The unused documents are called

"voids." The districts are to send back all voids to be tracked and scanned for accountability purposes. Each district is sent out approximately 15% overage on answer sheet/test forms.

Everything is initially sent to the district but is boxed by school (however, the school uploads the kids in the system Pearson Access).

OBSERVE HAND-EDITORS			

Review training materials and procedures for hand-editing of documents that do not scan properly Appendix B

Documents:
AU-IC Flatbed Scanning Procedures (Document Number 15666634 Version-7 Pages-3)
AU IC EDT NCount Procedure (Document Number 16285826 Version-6 Pages-2)
AU-IC EDT Accuracy Check Procedure (Document Number 15666572 Version-5 Pages-3)

Review procedures for double key-entry (2 people) for item data not scanned Appendix B

Document:
AU-IC EDT Online Editing Procedure (Document Number 15667789 Version-5 Pages-3)

Review alert processes Appendix B **Review resolution processes Appendix B**

Documents:
Image Writing Edit Specifications FCAT 2.0 Spring 2013 Writing 2/6/2013 10:02 AM
Image Program Requirements FCAT Writing Spring 2013 PGM-SSN-620-169 (Version-1.1 Pages-16)

OBSERVE TRANSCRIPTIONS			

Review tracking log for special documents Sections 4.12, 5.3, Appendix G

I was able to watch some of the Braille and large print forms being accepted. Before the Braille is sent out to be transcribed, the materials were checked in (accounted for) and cross referenced with the student data base (Pearson Access) to verify the correct student information and answer sheets were returned. Once the materials were verified they were prepared to be sent out to the transcriber. The materials were boxed up and sent with a tracking number one day mail to the transcriber. The transcriber was sent an email to alert her materials to be transcribed were sent and the quantity. The tracking numbers were stored in a separate database for large print and Braille only. When the transcribed answer sheets return from being transcribed they will then follow the same scanning procedure as all the other student answer sheets. The purpose is for the scorers not to know the students used the Braille accommodation.

Large print also followed the same check-in procedure, but then the student materials were photo reduced to 64%. The photo copy was then cut and taped into a regular test booklet. The photo-reduced booklet will then follow the same scanning procedure as all the other student answer sheets. The purpose is for the scorers not to know the students used the large print accommodation.

Review alert processes Sections 4.12, 5.3, Appendix G Review resolution processes Section 5.3, Appendix B

The alert and Resolution process for a transcription is the same process as all other scanned materials once the materials have been transcribed. The only alert and resolution process that is different for transcriptions is the materials are transcribed from a modified answer sheet to a nonmodified sheet with writing photo reduction and with FCAT R, M, & S, all items are currently multiple choice and the items are just transcribed (a regular sized answer booklet is filled in based on the students' original responses).

Review training materials and procedures for transcription of LP documents Sections 4.12, 5.3.1, Appendix G

Documents:
Project Information Requirements FCAT & FCAT 2.0 Accommodated Documents Transcribing Specifications (Version-1 02/11/2013 Pages-3)

Slightly different for writing they photo reduce and paste in answer booklet. Material double checked for quality control and then scanned and processed normally.

OBSERVE SCANNING PROCESS			

Compare scan and post edit files to information in answer documents Section 5.3, Appendix B

I was able to compare the computer or scanned image to the original student written document. Perfect match.

Review alert processes Section 5.3, Appendix B Review resolution processes Section 5.3, Appendix B

Document:
Scan Program Requirements Florida Comprehensive Assessment Test (FCAT Writing Spring 2013 / 620-169 (Presented in scanner calibrations above)

Review procedures for storing documents after scanning (preparations for shipment to warehouse) Section 4.4

Document:
Storing Documents Version 5 IC REC Strapping Procedures

Once the answer sheets have been scanned and verified they are strapped (bound together) based on their cart, batch, and stack numbering. Then the strapped stacks are bound, labeled, placed into storage boxes. These materials are scanned into the system and tracked by their location in the box (use a Z pattern) to identify placement in the boxes. Boxes are then scanned, labeled, and placed in the Gaylords (giant metal storage racks) that are labeled for easy location. This makes finding a single test document possible and accomplished in a timely manner.

FCAT 2.0 Scorer Training Observations

Grade 4 Writing, Jacksonville, FL

Stephen G. Sireci

March 12-14, 2013

Introduction

I observed the scorer training for the FCAT 2.0 Grade 4 Writing assessment. The scorer training occurred March 12-14 in Jacksonville, Florida. In this report, I summarize my observations regarding security, quality of training, and execution of the training process.

There were several personnel involved in the training, including staff from the Florida Department of Education (FDE) and Pearson.

Staff from the FDE included Steve Ash (Director of Test Development), Victoria Ash, and Elizabeth Tricquet.

Staff from Pearson included Dana Stimpert (Site Manager), Robert Owen (Site Aide), Ron Gibbs (Site Tech), Danny Lunde (Assistant), Mary McIntyre (Chief Scoring Director), Nancy Shafter (Scoring Director), Janice Tarleton (Scoring Director), Wendi Winkle (Project Manager), and 12 Supervisors/Assistants (Donna Adamczyk, Sandra Angus, Jerry Cohn, Sherry Czerniejewski, Brenda Ford, Chris Harper, George Johnston, George Leibig, Maureen McGrath, Wayne Miller, Charles Smith, Shannon Sumner).

In addition to FDE and Pearson staff, there were 237 scoring candidates who were being trained to score the essays. There seemed to be similar numbers of men and women, and some ethnic diversity among the candidates was evident, with approximately 10-20% appearing to be people of color. After training, the candidates are to be evaluated through two rounds of qualification scoring. It is expected that about 170 of the 237 candidates will be selected for operational scoring.

Security

Comprehensive security procedures were in place, and were adhered to throughout my 3-day visit. Prior to the first day, all scoring candidates had previously registered at Pearson's scoring center and were photographed. Picture ID badges were created, and on the first day (3/12), each scorer checked in at one of four security tables in the morning, signed in, and picked up their picture ID. They were then sent to prearranged seats. Appropriate seating arrangements were made for individuals with disabilities. The scoring candidates and all other personnel were required to wear their IDs at all times. There were at least four Pearson staff monitoring access to the building and the training room to make sure everyone had a visible ID badge.

On the second and third days, scorers and all other personnel were required to wear their badges to access the building and training hall. They were required to sign in at their respective table. Again, Pearson staff monitored all participants throughout the day.

The meeting materials were also secured through a comprehensive system. Each scoring candidate was assigned a numbered three-ring binder that contained PowerPoint slides and other descriptive information, anchor essays, practice essays, and essays that were used for prequalification and qualification. These binders were placed at the seats assigned to each candidate using a system that numbered the tables and each seat at a table. The materials were collected at the end of training and locked up overnight.

Training

The training was organized in a "conference-like" style, in a large auditorium. The Chief Scoring Director, Mary McIntyre, was situated behind a podium on a stage in front of the scoring candidates. The candidates were arranged across multiple tables, all of which had appropriate views of the stage and within adequate listening distance. There were a few problems with the Chief Scoring Director's microphone on the first day, but the system was quickly fixed, with minimal loss of training time. When candidates had questions during a training discussion, they were given a microphone so that everyone could hear their questions. The set-up for training was appropriate in terms of facilitating quality discussion among the trainers/scoring directors and scoring candidates, and the training appeared well organized.

The training was facilitated using PowerPoint slides. Some of the slides were difficult to see from the last row of tables, but all participants had printed copies of the slides. The slides, other training materials, and discussion focused on (a) expected personal characteristics (punctuality, courtesy), (b) security/confidentiality, (c) the processes that led to the essays selected for training (range finding, etc.), (d) importance of accurate scoring, (e) summary of quality management plan, and (f) discussion of reader bias. The discussion of reader bias included a definition of bias, what scorers should look for and what they should ignore, consideration of their own personal biases (e.g., forget other rubrics they may have used in the past), and other elements of reader bias prevention.

The training included discussions of holistic scoring (the method in general, and guidelines for holistic scoring), scoring criteria, and a review of the rubric to be used in grading the essays. In general, the introductory training materials seemed very good for explaining the process and instructing candidates how to validly score the essays.

After the initial overviews, training involved reviewing the anchor set as a group, and then giving the candidates time to score small sets of essays. The training materials were comprehensive and included the set of anchor papers (with bulleted annotations), practice sets, prequalification sets, and qualification sets. After candidates scored a set of essays, the essays were discussed as a group. Practice Sets 1 and 2 included 5 essays and were discussed on Day 1. Subsequent practice sets involved 10-20 essays and were discussed on Days 2 and 3. The question-and-answer periods in which the essays were discussed was facilitated well and addressed specific questions candidates had about the specific essays they had just scored. This

"hands-on" training approach seemed to help the candidates understand the rubric and scoring process.

The training featured extensive discussion of anchor and practice papers. In addition to discussing the 6-point rubric, specific issues associated with specific student responses were addressed through question-and-answer discussions, and general guidelines were provided. Examples of guidelines that were generated through discussion were "Don't worry about factual inaccuracies, just judge quality of writing," "Don't score for creativity," and "Don't score for voice." Some discussion focused on papers that could represent the border of two score points, and whether a response was off-topic (e.g., third-person response is okay). Guidelines for fourth-grade grammar were also discussed.

Breaks were given once in the morning at the end of a training or practice scoring session and for lunch (45 minutes). The sessions promptly restarted on time after breaks and lunch periods. The candidates seemed engaged throughout the training process.

Recording of Data and Evaluation of Candidates

After the candidates scored the practice sets of essays, they handed in their scores (on a paper scoring sheet) to a scoring supervisor/assistant. If the scores were not the same as the score assigned to the essay, the scoring supervisors/assistants wrote the correct score in the column next to the candidate's score. If the candidate's score differed by more than one point, the assigned score was circled to highlight the discrepancy. The percentage of scores that exactly match the assigned score (% exact) and the percentage within one score point (% adjacent) were recorded for each candidate on each set of anchor papers. The rates of discrepancies for specific student responses were tracked to identify trends and issues for discussion. Those anchor papers that had the most discrepancy from the assigned scores were highlighted for group discussion. The data collection, entry, and tracking procedures appeared fully consistent with the processes delineated in the FCAT 2.0 2013 *Handscoring Specifications* document (pp. 26-30, pp. 37-40).

The training process included "validity papers," which were prescored student responses selected by Pearson to represent specific points on the rubric. These validity papers were reviewed and approved by the Florida DOE. Approximately one in seven essays in the practice and qualification sets were validity papers.

To qualify for operational scoring, candidates must have at least 70% exact agreement on one of the two qualification sets of essays, no less than 60% exact agreement on the second qualification set, and no more than one discrepant score that deviates by more than one point from the assigned score.

Other Observations

I had a chance to observe the ePEN system for viewing essays. I was surprised at how clear the resolution was. Not only was the students' handwriting clear, one could also tell where erasures occurred. I was impressed with the system and believe it will facilitate valid scoring of the students' written responses. I was also impressed with the skills of the Lead Scoring Director. She handled questions well and patiently reviewed anchor papers and other student responses to help the candidates understand the scoring rubric.

Conclusion and Recommendations

No flaws or inefficiencies were evident in what I observed. The process was extremely well-organized and was executed as planned. Security procedures were in place and were adhered to. My conclusion is that the training of the 4th grade essay scorers was conducted efficiently, successfully, and appropriately.

One recommendation for the future is to survey the scoring candidates with respect to their impressions of the training experience. The training seemed good to me, but perhaps some areas were repetitive or conducted too quickly, which would be difficult for an observer to detect. Candidate survey results could help us to better understand which elements they thought went well, and which could use improvement.

When I inquired about surveying candidates, Pearson staff reported that they survey candidates who qualify to ask them about their training experience. I have requested a copy of that survey. It may be helpful for the Florida DOE to review the results from the survey and to consider adding additional questions if necessary. It may also be helpful to survey all candidates, including those who do not get selected, before they are informed of the hiring decision. Comparing the results of such a survey across qualified and unqualified candidates may also prove interesting.

FCAT 2.0 Scorer Training Observations

Grade 8 Writing, Auburn, WA

Robert A. Spies

March 12-14, 2013

Introduction

During the course of the above days, I observed the first wave of scorer training for the FCAT 2.0 Grade 8 Writing assessment. The formal training was conducted away from the Pearson Scoring Center (PSC) in Auburn, Washington, primarily due to the required size of the training class. In this report, my observations are summarized into the sections of security, training, additional observations, and conclusions derived from this specific training period.

There were a number of individuals involved in the initial training process that included both the Florida Department of Education (FDE), Test Development Center (TDC), and Pearson Scoring Center site and visiting personnel.

Visiting Staff from TDC consisted of Renn Edenfield (English/Language Arts Coordinator).

Visiting Staff from Pearson included Bob Sanders (Vice President – Assessment and Instruction), Kenna Fries (Content Specialist), and Molly Less-Peterson (Human Resources Manager).

Site Staff from Pearson included Rob Heinzman, Susan Blake, and Leah DeHoet (Scoring Directors), Bonnie Bizzell (Site Manager), Cassie Chinn (Site Aide), Susan Cochran (Site Tech) and working in an administrative role during this time, Sharon Park, Gregg Rice, and Laura Stark (Scoring Supervisors).

In addition to the above FDE, TDC and Pearson staff, Pearson recruited the specified 18 supervisors and the exact number of 265 scoring candidates for 8th grade scoring listed in the Resource Allocation (4.1.4, 4.1.5) of the FCAT 2.0 *Handscoring Specifications.* After training was completed over the course of the 3 specified days, candidates transitioned to the Pearson Scoring Center in Auburn, WA. In order to qualify to serve as scorers, candidates were engaged in qualification rounds that were expected to yield approximately 172 operational scorers.

Security

Procedures necessary for the secure admission of eligible candidates only into the Puyallup Center were consistently maintained throughout the 3 days of this observation. During the days leading up to initial candidate training, photographs and identifying information were processed for the majority of candidate scorers. For those few candidates receiving authorization just prior to the start of training, temporary badges were issued.

On the first day, candidates signed in at the front desk and were issued their Pearson Photo ID. All candidates and staff were instructed to keep their identification in plain sight while inside the Puyallup Center. After admission, candidates walked to their assigned seating in preparation for official instruction to begin. Pearson site staff (listed above) positioned themselves and rotated during scheduled break periods so that they could consistently monitor access to the Puyallup Center entrances, exits, and the materials storage room. Candidates were presented with general scorer guidelines, security, and confidentiality requirements related to essay scoring including nondisclosure rules, limitations with secure materials, and access to personal electronic devices.

In the second and third day of this training period, only candidates with their badges on physical display were directly admitted to the Puyallup Center. An additional screening procedure was required for any candidates who had forgotten their badges. When candidates failed to display their badges properly, they were called on (often by their name) and asked to place the badge so that it was clearly visible. After candidates were admitted into the training room, they signed in at their desks to register both time and attendance. At the beginning of the second day, security procedures were tightened so that only one entrance and one exit (excluding the elevator) were permitted into and out of the training site, and signs were placed on doorways to clearly identify their use. Staff were positioned at tables to enforce this rule.

Training materials were kept in a storage room that had been rekeyed for purposes of enhanced security. All FCAT training materials once distributed were either kept on tables in front of individual candidates or were collected (when completed) by supervisors. Training materials consisted of three-ring binders for each of the candidates containing PowerPoint slides (with procedural and policy rules reviewed the first day), and with FCAT anchor essays (two or three examples at each of six score points). Binders were kept in front of individual work areas. All candidates could be specifically identified at their table and chair. Materials were secured at night and distributed again in the morning. Overall, Pearson's Auburn staff consistently followed the Off-Site Training security procedures identified in the FCAT 2.0 2013 *Handscoring Specifications* (pp. 31-33).

Training

The Puyallup Fair and Event Center was the site of the first 3 days of FCAT training due to its comparatively large rectangular size that could likely accommodate just over 300 candidates and staff. The Scoring Directors (primarily Rob Heinzman) provided instruction from a built-up stage area with podium and a large overhead screen. The training space was constructed so that candidates could easily hear and see the formal instruction from all sections of the room. During 3 days of training at this site, both the audio and visual equipment worked without any apparent difficulties. To facilitate the training experience, candidates with questions during training would either have microphones brought to them or the scoring directors would repeat the candidate's question before providing an answer. Overall, the setup of the site offered scoring directors with the maximum opportunity for training to be seen and heard by candidates.

Initial instruction for candidates was accomplished through projection of PowerPoint slides and supplemented by the same text being reproduced in the candidates' three-ring binders. Among the expectations for candidate behavior discussed in the early stages of training were maintaining a professional and respectful demeanor, timeliness to training class each day, and the importance of security throughout the scoring process as required by Florida statute and rule. Candidates were reminded on numerous occasions of the importance of accurate scoring to individual students, parents, and school districts. Avoiding bias was presented as essential to the accuracy of all essay scores, and examples were provided (e.g., reactions to style and content, overall appearance) so that candidates had concrete illustrations of bias interfering with score decisions. Scoring directors often stated that their goal was to help all candidates qualify as FCAT scorers, and specified the type of candidate questions that would assist or detract from this overall purpose.

Instruction on the first day included a complete discussion of the holistic nature of essay scoring. Descriptions of the characteristics of the four writing elements (focus, organization, support, and conventions) were provided to candidates along with examples of high versus low scores on each element. Training in the four elements and illustrations of high versus low scores appeared to provide candidates with a very effective path to understanding the four writing elements without confusing the overall importance of integrating these elements into one holistic score.

The FCAT scoring rubric was then described in sufficient detail for candidates to gain a basic conceptual understanding behind each of the six score points. Training proceeded with examples of anchor essays discussed at each score point followed by three sets of practice essays limited to a specific score range (score points 1-3, 4-6, 2-4). After these scoring exercises were completed, candidates moved to larger numbers of essays and the complete range of potential score options in the second and third days of training for their practice and prequalification sets.

When candidates finished their practice essays, the scoring directors read each essay aloud, asked candidates for a show of hands indicating their score for the essay, provided candidates with the assigned score, explained the assigned score, and opened the floor to general questions. Despite the best efforts of scoring directors, in training sessions last year, a small number of candidates became agitated by their failure to understand the scoring rubric. This year, the scoring directors were particularly effective at limiting any disagreements by carefully explaining the rationale behind each score point, by avoiding any defensive tone to the discussion, by addressing questions rather than comments, and by repeatedly focusing on the overall goal of the training process (i.e., to qualify the maximum number of candidates as scorers).

Candidates received regular breaks at mid-morning and mid-afternoon during each of the 3 days of observation. A lunch break of 45 minutes was typically scheduled around noon with small time variations based on the training schedule. Training was restarted in a timely manner.

During the specified period of scorer training above, the Auburn PSC followed the general guidelines and specific procedures described in 7.1 and 7.3 of the FCAT 2.0 2013 *Handscoring Specifications*.

Data Recording and Candidate Evaluation

As part of the training process, candidates recorded their scores during practice scoring and prequalification rounds on two paper forms. When these scores were completed during the time limit specified by the scoring directors, supervisors collected and scored the forms, passing one set back to the candidates and the other to Kenna Fries (Pearson Content Specialist). Records were kept for purposes of general record keeping and identifying latent trends and training needs. Forms tracked the percentage of correct scores listing exact match, adjacent scores, and nonadjacent scores compared to the assigned score for all practice and prequalification rounds.

Other Observations

The materials and organization of the candidate qualification process at the Puyallup Center exceeded the appropriate standards set in the previous year. During a substantial amount of the available training time, candidates were actively engaged in learning the fine details of scoring essays. Few moments of administrative downtime were apparent. The Scoring Directors (particularly Rob Heinzman) provided candidates with a variety of strategies designed to assist their development of essay scoring expertise. Active instruction coupled with the paced introduction of new skills, consistent reassurance of candidates (especially during intervals of greater learning frustration), and the frequent injection of humor established a productive learning environment for candidates to learn the required 8th grade scoring rubric.

Conclusions

The first round of training at the Puyallup Center consistently followed the specifications outlined by the FCAT 2013 *Handscoring Specifications* during the 3 days of this observation. The Auburn PSC met the expected candidate target (265) essential to employing a sufficiency of FCAT essay scorers. Based on statements made by Pearson's visiting and site staff, individual scorers were recruited in numbers that considerably exceeded the actual numbers needed. Both personnel and scoring directors were able to examine potential scorers on a variety of selection criteria that increased candidate chances of being successful.

Secondly, the anchor sets prepared during field-testing clearly demonstrated both between-point (1-6) and within-point (low, middle, high) scoring differences that could be consistently and coherently explained by scoring directors to inexperienced candidate scorers. During the observed training period, few candidates expressed any confusion at their inability to understand either the anchor sets or the scoring decisions in FCAT practice or prequalification essay scoring.

Finally, security was consistently maintained during all aspects of training at the Puyallup Center. Staff remained in their specified positions until relieved by other staff members so that no security breakdowns could occur. The training materials were not subject to compromise at any specific points and only authorized candidates for training were admitted to the facility as stipulated by the FCAT 2.0 2013 *Handscoring Specifications*.

FCAT 2.0 Scorer Training Observations

Grade 10 Writing, Tucson, AZ

Anja Römhild

March 12-14, 2013

Introduction

I observed the scorer candidate training for the 2013 FCAT 2.0 Grade 10 Writing assessment at the Pearson Scoring Center in Tucson, Arizona from March 12 to March 14, 2013. This report summarizes my observations and evaluation of the training process and the training materials I reviewed. Some information in this report is based on conversations with Sally Rhodes, the representative of Florida's Test Development Center, and members of the Pearson project team including the program manager Ross Holstein, site manager Martina Teran, and the Pearson scoring directors Alexandra Atrubin, Milton Eichacker, and Anita Cook.

Security

Pearson implemented multiple security measures that ensured only authorized personnel were able to enter the facilities and that kept sensitive test materials secure throughout the training. The security procedures appeared to be very appropriate for the site and were consistently adhered to by Pearson staff. Entrance to the site was continually monitored by the front desk staff. All scorer candidates were issued a photo-id badge on the first day of training and were required to wear the badge visibly throughout the day. Sign-in and sign-out sheets logged visitors to the site. Visitors were also required to wear name badges. At the beginning of training, all scorer candidates signed nondisclosure agreement forms and were instructed not to copy, remove, or discuss any test materials, individual student responses, or information pertaining to the scoring project outside the scoring center. Materials given to scorer candidates were collected at the end of each training day and kept secure in a locked space.

Training Site and Logistical Arrangements

The scorer training in Tucson, AZ was attended by approximately 310 scorer candidates who had been prescreened by Pearson to meet educational degree requirements. Many of the invited candidates also had previous scoring experience. The scorer training took place on the fourth floor of a former furniture store that had been leased by Pearson for the duration of the scoring project. The training space consisted of a large open area in which rows of desks were set up at which the candidates were to sit. Additional screens were set up to project the presentation

materials to the back rows. Speakers and microphones ensured that everyone was able to hear the presenters and candidates asking questions. Overall, the space was quite pleasant and conducive for the purpose of training such a large group.

On the first day of training, all scorer candidates were checked in and assigned a desk space, which was theirs for the remainder of the training. Presentation materials were distributed in a binder and included the orientation PowerPoint slides, several materials with information about the scorer qualification standards, the holistic scoring method and scorer bias, the writing prompt with allowable interpretations, the scoring rubric, and lastly, the anchor paper set with annotations. Once the training transitioned into practice scoring, candidates also received separate packages with practice papers and one prequalification paper set. All materials were collected at the end of the day by scoring supervisors and stowed away in a locked room.

During the training, candidates were allowed two 15-minute breaks and one 45-minute lunch break. In general, the Pearson scoring directors were very good at keeping candidates on schedule.

Training Process

Training began with a 2-hour morning orientation led by the lead scoring director, Alexandra Atrubin. The orientation started with a reminder to scorer candidates of the security and confidentiality of the training materials and gave instructions concerning expectations of professionalism during the training process. An overview of the lifecycle of the FCAT 2.0 Writing assessment was given by Sally Rhodes from the Test Development Center, which provided useful background information to contextualize the scoring task. Scorer candidates were then introduced to the various training materials (anchor, practice, prequalification and qualification sets) and their functions, which was followed by a discussion of the scoring qualification standards specified in the quality management plan. A tutorial on reader bias concluded the general overview section of the orientation. The following presentation focused on introducing the writing prompt and the holistic scoring approach. A fair amount of time was devoted to introducing and discussing the scoring rubric and its four writing elements. In particular, the rubric element support was discussed in more depth in order to develop an understanding in candidates of different levels of support that can be found in students' writing. The orientation concluded with a reminder to scorer candidates of how to successfully apply the holistic scoring method using a variety of useful strategies. I found that the discussion of these strategies was quite useful. It may be helpful to candidates if those strategies were addressed more frequently throughout the training. Overall, the morning orientation flowed well and provided a thorough introduction to the scoring process. Scorer candidates appeared to follow along attentively, with many of them diligently taking notes and marking their materials.

After the morning orientation, a thorough introduction of the anchor paper set was given. The lead scoring director read out loud each individual anchor paper and the annotations that had been developed for it. Additional clarifications were made as needed, and scorer candidates were encouraged to ask questions. In the afternoon of the first training day, candidates took the first two practice paper sets, which each consisted of five prescored student essays that represented a score point range from 1 to 3 in the first set and 4 to 6 in the second set. Candidates were informed in advance of the score point range of each set. Three additional practice sets were given on the following 2 training days. The third set included 5 essays in the score range from 2 to 4. The last two sets included 12 and 15 essays each and spanned the entire score range from 1 to 6. A special prequalification set consisting of 20 essays and representing all score points was also given on the third day of training. This set was assembled under the same parameters as the qualification sets and was intended to familiarize candidates with the task of scoring 20 essays at once. Because the qualification sets were going to be administered in the ePEN system and not on paper, however, it may be beneficial if scorer candidates can practice scoring the prequalification set in the ePEN system as well.

Once scoring of a practice set was completed, supervisors collected the score sheets from each candidate. The practice scores were entered into a spreadsheet to track the distribution of scores per essay as well as the scoring accuracy rates of individual candidates and the candidate pool. Those accuracy rates (exact score agreement and exact or adjacent score agreement) were marked onto each candidate's score sheet and then returned to the candidates. For the review of each practice set, the scoring director read each individual paper and commented on the paper's score using prepared discussion points. These discussions typically focused on specific characteristics in the paper and related them to rubric elements. In addition, the scoring director also referenced a specific anchor paper for comparison. Scorer candidates were able to ask many questions during each review, which the scoring directors willingly addressed.

During the site visit, I attempted all five practice sets and the prequalification set. I found that scores within each set were quite evenly distributed so that essays with the same score point would not appear back-to-back in a set. This observation influenced how I assigned scores. For example, knowing that I scored the previous essay with a score point of 3, I would be swayed to give the next essay a score other than 3. Such strategies, whether productive or not, would be disabled if the order of essays within a set were better randomized and allowed for the same score point to occur back-to-back within a set.

Conclusions

The scorer training for the Grade 10 FCAT 2.0 Writing assessment was competently executed and provided an effective structure for candidates to learn the complex task of scoring student essays. The strong emphasis on practice scoring during the training is particularly

valuable and I welcome the fact that this year additional practice opportunities were created with the prequalification set and the advanced session of pseudoscoring in ePEN.

The scorer training was also faithfully executed in accordance with the *Handscoring Specifications* and with the security of the test materials consistently maintained. Throughout the training, the Pearson scoring directors provided helpful guidance and instruction and created a supportive environment for candidates. In turn, scorer candidates appeared to be focused and genuinely concerned about giving fair and accurate scores to students' work. Overall, I found that the scorer training accomplished its purpose with great success.

FCAT 2.0 Observations of Operational Scoring

Grade 4 Writing, Jacksonville, FL

Stephen G. Sireci

April 2-4, 2013

Introduction

I observed the 4th grade essay scoring from Tuesday April 2, 2013 through Thursday April 4, 2013 in Jacksonville, Florida. In addition to observations, I discussed the scoring processes and progress with Pearson staff, particularly Scoring Directors Mary McIntyre and Nancy Shafter. I also interviewed one experienced scoring supervisor and two scorers. The interviewed scorers were a convenience sample of $n = 2$.

In addition to the aforementioned Scoring Directors, Pearson staff on site included Robert Owen (Site Aide), Dana Stimpert (Site Manager), Ron Gibbs (Site Tech), Janice Tarleton (Scoring Director), and several Supervisors/Assistants (Donna Adamczyk, Sandra Angus, Sherry Czerniejewski, George Johnston, George Leibig, Maureen McGrath, Wayne Miller, Shannon Sumner).

There were approximately 72 scorers, which represents about half of the desired number. Apparently, many of the 237 scoring candidates were not successful in achieving the criteria required to be certified for scoring. Pearson ran a second scoring training/recruiting series to address this problem, recruiting about 70 new candidates, but again the acceptance rate was low, with only about 10 candidates qualifying. On the positive side, the low number of certified scorers demonstrates that the strict quality control standards for selecting quality scorers were adhered to. Clearly, the scorer selection process is criterion-referenced, adheres to high standards in scoring, and does *not* become norm-referenced if there is a shortage of qualified scorers. The approximately 72 scorers who were hired met the requirements for beginning to score FCAT essays.

Security

As with scorer training, security at the site was well organized. All scorers were given picture ID badges, and these badges were checked at the door each morning by Pearson staff. There was only one entry point for the building and it was continuously monitored by at least one Pearson staff member. The ID badges for all personnel were clearly visible. Visitors such as I were required to sign in at the door. Scorers sign in each morning with their assigned supervisor. They were assigned a seat at a scoring table, and given their binder that contained the training materials (anchor sets, etc.). All scorers were also assigned a laptop on which the essays they were to score were presented. I reviewed the sign-in sheets and all processes appeared to be appropriately followed.

Materials and Arrangements

Each scorer worked on a laptop to read and score the essay responses. The electronic images of the essays were clearly displayed on the screen. Anchor sets – discussed thoroughly during training and then briefly each day during scoring – were provided to each scorer in a three-ring binder. Most scorers referred to the anchor sets during the scoring process and had written extensive notes, highlighted text, and annotations on them. Additional materials, such as new calibraton papers for discussion, were handed out as needed, and the scorers could insert them into their binders.

The Scoring Supervisors sat facing their team. If scorers had a question or issue, they simply raised their hand and the supervisor came over to assist them. Each supervisor had between 5 and 12 scorers on their team, which was appropriate, given the smaller number of scorers than anticipated. One positive consequence of the smaller group was that all scoring and discussions occurred in a single room, which facilitated discussion of calibration papers, anchor papers, etc. Discussions were held at the beginning of each day, and following breaks when needed. The microphones, lighting, seating, and other logistics all functioned well.

The day began with a discussion of specific essays (either anchor, calibration, and/or validity) designed to address a problem the Supervisors noted the previous day (e.g., a "4/5 line"). These discussions typically ran 45 minutes to an hour. Scorers then graded continuously with the exception of 15-minute breaks in the morning and afternoon, and a lunch break for 30 minutes in the middle of the day. Due to the lower number of scorers than desired, scorers who had high validity ratings could qualify for "overtime," where they could come in an hour early and stay an hour later to score essays. The daily conference calls documented the numbers of scorers who took advantage of overtime.

Monitoring and Feedback

The Supervisors and Scoring Directors appeared to be continuously monitoring and supporting scorers. Feedback and training were provided in several different ways including group discussions and personal feedback. Supervisors provided feedback to scorers through personal conversation or via notes that could be sent to scorers over the ePEN system. These notes could be positive reinforcement for scorers based on an improvement that was observed or could be suggestions to correct a problem noted by the Supervisor. Scorers also received assistance based on annotations that were placed on the validity papers they scored.

The quality control and scorer monitoring procedures included *validity papers*, *calibration papers*, and *backreading*. Validity papers represent prescored and preselected essays that were agreed upon by Pearson and the FDOE. Scorers must score 70% of these papers exactly as the assigned score to continue to qualify for scoring. Scorers must also demonstrate 60% interrater reliability, which is defined as scoring 60% of their essays the same as the second score assigned to these essays by other scorers. Scoring Supervisors also backread and scored essays scored by their team members, focusing particularly on members whose validity or interrater reliability statistics suggest they are falling off the rubric. At least 5% of a scorer's essays are backread; but Supervisors are likely to read more for scorers for whom they are most

concerned. Calibration papers focus on specific "lines" in the score rubric and are chosen to coach scorers for specific problem areas noted during recent scoring (e.g., the previous day). Calibration sets are typically 1-3 essays, and Supervisors and Scoring Directors discuss scorers' responses to these sets to help them improve assigning scores to essays at the relevant score points on the rubric.

Monitoring of scorers' performance was facilitated by the ePEN system. I viewed this system and spent some time viewing the reports available to the Scoring Directors and Supervisors. I was impressed. The numbers and percentages of operational, calibration, backread, and validity essays scored by each scorer, scoring teams, and the total group of scorers could be easily reported. The statistics calculated were understandable and were presented in a manner in which Supervisors and Scoring Directors could easily identify several indices to flag a scorer. Statistics were also produced for the essays scored by the Supervisors, whose performance was also monitored. These reports are described in the FCAT 2.0 2013 *Handscoring Specifications* document, but examples of the reports are not included in that document. In addition to the "17a and 17b" reports just described, Supervisors and Scoring Directors could also access "validity disagreement reports," which provide data to evaluate scorers' performance on the validity essays. All reports could be produced on-the-fly in real time, with new data continuously populating the reports as scorers entered their scores.

General Observations

In addition to the security and logistic observations already noted, my impressions were that the scorers appeared to be concentrating on scoring the essays at all times, and were generally engaged in the group discussions. The scoring and monitoring processes seemed entirely consistent with the descriptions in the FCAT 2.0 2013 *Handscoring Specifications* document (pp. 26-30, pp. 37-40).

Interviews

Over the 3-day period I had several conversations with the Scoring Directors and Site Manager, and I requested an interview with a Scoring Supervisor, which was granted. All Pearson staff members were extremely helpful in answering questions and providing any information I needed. In addition, during one of the breaks, I introduced myself and described my role to two scorers, and I asked if I could ask them a few confidential questions about their experiences. These interviews contributed to my observations. The information I received from these different interview sources led me to conclude the operation was proceeding as expected, with the exception of the disappointing number of scorers who qualified. The scorers noted that they particularly valued annotations on the validity papers, which helped them a lot, and the discussions of the calibration and validity papers. The scorers did not find their individual statistical reports very helpful for improving their scoring. When asked if they could recognize a validity paper they previously scored, they said they could, but they were not able to remember

the previous scores they assigned to these essays, and so remembering the essay did not affect the score they assigned the next time around.

Conclusion

Based on my observations, interviews, and review of ePEN reports, it appears the scoring of the 2013 FCAT 2.0 Grade 4 Writing Assessment is being appropriately conducted and adheres to the quality control criteria specified in the FCAT 2.0 2013 *Handscoring Specifications* document. The team of Pearson staff appears to be working very hard to support the scorers to help them to continue to reach the high validity standard that will enable them to continue to score. The system is set up so that the scorers who start "drifting" the most are quickly identified and intervention commences to get them back on track.

Recommendations

Although the quality control and scoring processes were deemed appropriate, there are some minor recommendations that can be offered based on the observations.

1) Include sample monitoring reports from the ePEN system in the appendices of the FCAT 2.0 2013 *Handscoring Specifications* document. Doing so will help future auditors understand the statistics supervisors use in monitoring scorers, and how the Supervisors and scoring sites are monitored.

2) Investigate reasons why the success rate in acquiring qualified scorers was much lower than expected, and lower than previous years. Given that the training appeared effective, changes in the candidate pool, and the recruiting strategies employed, should be evaluated.

3) Continue to provide annotations to validity papers and other essays graded by scorers that are used to coach them towards improvements. The scorers seem to like these annotations and use them in grading.

FCAT 2.0 Observations of Operational Scoring

Grade 8 Writing, Auburn, WA

Robert A. Spies

April 16–18, 2013

Introduction

The second observation period for the 8th grade essay scoring in Auburn, WA began April 16, 2013 and was completed April 18, 2013. During the time period, I held a number of conversations with supervisors and staff from the Auburn Pearson Scoring Center (PSC) and conducted systematic observations.

Of the initial 265 candidates who began scorer training, 29 scorers dropped out prior to the qualification process being completed. After the qualification sets were finished, the remaining 236 candidates were reduced to 161 qualifying scorers. This number of qualifying scorers was nearly the 172 expected (FCAT 2.0 2013 *Handscoring Specifications*, pp. 34). By the time of the second visit, 11 scorers had voluntarily left FCAT scoring for a variety of reasons (e.g., prior employment reinstatement, new job position, family emergencies) but only one had been discharged due to the scorer exception process. Over the 3 days of this current observation, between 100 and 106 Grade 8 scorers were on-site along with 12 Grade 8 supervisors.

During the week previous to this visit, 45 scorers and 3 supervisors in the Auburn PSC had been reassigned to work on the Grade 4 scoring project. These Grade 4 scorers and supervisors underwent the same process previously required of Grade 8 qualifying staff. Auburn PSC Grade 4 scorers and supervisors worked in a separate, partitioned-off section of the building. When the anchor reviews were conducted for Grade 4 and Grade 8 at the beginning of each day, Grade 4 scorers and supervisors moved to a separate section of the Auburn PSC to prevent any disruption of the anchor review process for either group. Of the 45 original Grade 8 scorers who were redesignated to work on Grade 4 essays, 30 scorers and 3 supervisors (plus Nancy Shafter from the Jacksonville PSC) participated during this observation period with the Grade 4 scoring. Scorers not qualifying for Grade 4 scoring were eligible to return to the Grade 8 FCAT scoring process.

Security

The organization of security at the Auburn PSC was consistent with the procedures previously implemented at the Puyallup Fair and Event Center during the initial training period. One entry and one exit (the same) were accessible to scorers, supervisors, and staff. Two emergency exits also existed within the building, but alarms were set to sound immediately if these doors were opened. All scorers, supervisors, and staff displayed photo ID badges that were on physical display. If the badge was not immediately apparent, one of the three staff members at the front desk (Cassie Chin, Susan Cochran, or Bonnie Bizzell) would request the individual to

display their badge more prominently. The front desk was continually staffed. There were no occasions where secure entrance to or exit from the Auburn PSC was compromised. Scorers typically signed into the PSC at their supervisor's desk in the morning and out when leaving for the day. On the occasion when a scorer forgot to sign in, their direct supervisor would bring the sheet over for their signature. I examined the sign-in sheets during this visit and found the number of official scorers to be appropriately listed on the form, in their assigned desk space, and within the visual observation of their supervisor. No phones, recorders, cameras, or similar electronic devices were permitted (or observed) in the scoring area.

Materials and Arrangements

All scorers for the Grade 8 project in the Auburn PSC worked from Dell 260/280 desktop computers with 15-inch flat screens that easily could be adjusted to increase or decrease the display size of the FCAT essay. The majority of desktop computers used ethernet ports to connect to the Pearson mainframe, with no access to other websites available. Scorers and supervisors accessed the ePEN system via their assigned login and password. A minority of desktop computers in the Auburn PSC accessed ePEN through wireless (wifi) access. Although all wifi access was restricted to the Pearson website, complications would sometimes arise when resetting wireless computers (through the simultaneous press of the control, alt, delete keys) because it would first restart the Windows (XP) Operating System. The Dell computers also contained USB ports, but without the required administrative access, ePEN will (by report of the Auburn site tech) not allow recognition or any type of systems access.

Each day began with an anchor review typically selected based on an examination of the scoring results (using Report 34A, 34B) from the previous day. On the first day of the week, scoring directors reviewed the full anchor set to reintroduce scorers to the full anchor point range. On subsequent mornings of the week, scoring directors would focus on one score point or the borderline range between two score points. These reviews would typically require between 30 and 60 minutes depending on the number of score points and follow-up questions being discussed. During the anchor review process, individual scorers were observed examining their anchor sets, rechecking additional materials in their three-ring binders, or visually following the scoring directors as anchor sets were being read aloud. After anchor set reviews were completed, scorers proceeded directly to scoring FCAT essays.

Scorers and supervisors were all within adequate visual and auditory range to be understood during these anchor reviews. In addition, building temperature, lighting, and other environmental factors at the Auburn PSC were within the expected comfort range. Scorers received two regularly scheduled 15-minute breaks in the morning and afternoon, along with a 30-minute lunch break. If needed, scorers were also permitted short individual breaks. Due to the size of the break room at the Auburn PSC, all breaks were split so that only half of the scorers would be away from the scoring area at one time.

Monitoring and Feedback

During the observation period, scoring directors were consistently engaged in training scorers in large groups or reviewing the scoring process to evaluate potential problems with Grade 8 scoring accuracy. Individual score points for training were frequently identified using Reports 17A, 17B, 34A, and 34B. In addition, scoring directors acted to support scoring supervisors in those instances where individual scorers would suddenly demonstrate variability in their score accuracy. Overall, scoring directors worked through scoring supervisors to impact individual scorers.

Each scoring supervisor was observed to have direct responsibility for between 7 and 10 scorers. Scorers would typically raise their hands to request the assistance of their supervisors when any questions or scoring difficulties were encountered. The primary responsibility of scoring supervisors included monitoring individual scorers through a variety of tools available through ePEN. Supervisors often used electronic messages to contact their assigned scorers. When an individual scorer had improved their accuracy, supervisors would send a message of encouragement. When problems were noted, scorers would receive a message to visit their supervisor. Supervisors indicated a strong preference for personal discussions when scorers were at risk of being discharged or when small adjustments were needed to maintain scoring accuracy.

Four discrete quality control procedures were designed to ensure the accuracy of FCAT scoring. First, the ePEN system automatically maintains a validity check on the quality of FCAT scorer performance. Validity is measured by comparing the exact agreement of a scorer's rating against essays rated by specialists from the Test Development Center and Pearson. Approximately every seventh essay represents a validity paper in the scorer's queue. In order to meet the expected standards, scorers must exactly match the previously established rating on 70% of these essays. Secondly, because all FCAT essays are scored twice, scorers must match the second scorer's assigned rating (i.e., interrater reliability or IRR) 60% of the time. Both daily and cumulative validity and IRR scores are reported during regular FCAT telephone conferences. In addition, supervisors are required to examine at least 5% of their individual scorers' ratings through a system of "backreading" scores. This scoring requirement provides supervisors with a means to evaluate individual scorer tendencies to consistently miss the score point rating (e.g., too high, too low) or to mistake a specific score judgment (e.g., 3 versus 4) on their assigned essays. Finally, calibration papers offer scoring directors and supervisors the opportunity to coach scorers on particular problem areas of essay scoring that have often proved difficult to score. These essays are sometimes found as a result of scorer questions or on observed inconsistencies between scores that then became the subject of specific training.

During the 3-day visit, there were a number of opportunities to examine the ePEN system in regular use. Scoring directors made training decisions based on cumulative score point errors that could be easily identified by accessing specific reports (e.g., Report 34A, 34B). Supervisors used ePEN to read essays, to review the individual ratings of scorers, to identify scorers demonstrating variance in their application of the scoring rubric, and to communicate directly with scorers. The ePEN system appeared user-friendly and was continually in use by supervisory staff.

Interviews

Interviews were conducted during the two visits with Bonnie Bizzell (site manager), Rob Heinzman, Susan Blake, and Leah DeHoet (scoring directors for Grade 8), and Nancy Shafter (scoring director for Grade 4). In addition to this supervisory staff, I also held discussions with Susan Cochran (site tech), Cassie Chinn (site aide), Heidi Miller (human resources), several experienced scoring supervisors, and four essay scorers.

General Observations

During the two observation periods of this year's Grade 8 FCAT 2.0 Writing scoring, it was apparent that both the Florida TDC and Pearson had made substantive efforts to organize this year's FCAT scoring consistent with the specifications listed in the *Handscoring Specifications of the Florida Comprehensive Assessment Test FCAT Writing 2013*. For example, this year's Grade 8 persuasive writing prompt appeared to provide an effective forum for students' written communication. Several Grade 8 scorers with previous Grade 8 FCAT scoring experience volunteered that the anchor papers presented by scoring directors progressed in a consistent and logical order for them to effectively evaluate FCAT essays. In addition, Pearson Human Resources appeared very successful at recruiting high caliber scorers at the Auburn PSC.

The scorer qualification process this year used a variation from the previous year with the pseudoscoring of essays occurring **before** (rather than after) the qualification sets used to determine the eligibility status of scorers. This change of sequence provided increased training opportunities for scorers and may have allowed more skilled individuals to qualify for FCAT scoring. In addition, because computer monitors were used for the qualification process rather than paper, the task was more analogous to the actual process being used during FCAT essay scoring. Both changes may have contributed to the higher number of scorers qualifying this year compared to the previous year. Of the 236 scorers who took these qualification sets, 161 (67%) became eligible to score Grade 8 essays. However, it must be noted that different results were observed at Grade 4 and Grade 10 scoring PSC sites.

In previous years, small scorer groups at risk of being dismissed were gathered together for additional scorer instruction and calibration review. This remedial process, although well intentioned, may have contributed to an unnecessarily large number of scorers becoming discouraged about their continuing employment and their ability to score essays. During Grade 8 scoring this year, underperforming scorers were not directly identified to their peers and instead received electronic notices and targeted remedial calibration training sent directly to their computer screens. Scoring directors believed this specific revision to last year's procedures was an effective tool that raised morale while maintaining a larger number of active scorers. Based on the numbers of individuals who continued to score Grade 8 essays, sufficient evidence exists for continuing this practice.

Scorers easily accessed the annotated essays in their computer library, which provided increased assistance (in addition to the anchor essays in three-ring binders) for making difficult score point discriminations. Although some scorers expressed concern about the time pressures resulting in reviewing these essays, all scorers addressing this topic believed that annotated

essays provided valuable information not otherwise available from anchor items and their (sometimes) vague memory of previous calibration training.

Conclusion

Overall, the results of the two observation periods and specific reviews of training documents suggest that the Auburn PSC followed with fidelity the *Handscoring Specifications* for Grade 8. Pearson's ePEN software system effectively managed multiple scoring tasks (e.g., validity, IRR, calibration, qualification) with no observed compromises to either the timeliness or integrity. The procedural changes made to this year's scoring process (documented in the general observations section) appear to have been effective at retaining greater numbers of qualified scorers and streamlining the overall scoring process.

Recommendations

Based on the observations of the Grade 8 FCAT scoring procedures during the current year, the following recommendations are made:

1) Identify personnel practices and policies at the Auburn PSC that allowed the successful recruitment of the large number of skilled candidate scorers during this year's FCAT scoring;

2) Continue this year's practice that postponed scorer qualification until after pseudoscoring had been completed in order to allow scorers sufficient experience with ePEN and the holistic scoring process;

3) Instead of using small group instruction for scorers at-risk of being discharged, continue this year's practice of remedial training administered individually through the ePEN system and through targeted calibrations;

4) Maximize the numbers of annotated essays and ease of access to the ePEN reference library (especially during the early scoring stages) for scorers seeking help with difficult score point discriminations.

FCAT 2.0 Observations of Operational Scoring

Grade 10 Writing, Tucson, AZ

Kurt F. Geisinger

April 17-19, 2013

Introduction

I observed the scoring of the FCAT 10th Grade Writing Test from April 17-19 in Tucson, AZ, Wednesday through Friday. In addition to making frequent observations in as unobtrusive a fashion as possible, I also interviewed several participants briefly and discussed the scoring process informally with many others. The Pearson staff on hand included three scoring directors (Alex Atrubin, Milton Eichacker, and Anita Cook) and the Site Manager, Martina Teran. In addition on April 17, there were some 95 of the 102 scorers and all 12 supervisors in place. After their second wave of training, 125 scorers met standards to be scorers. On the second day of this visit, one scorer had not shown up for several days and was eliminated; therefore, from April 18th on, the number of possible scorers was 101.

I had not seen this scoring site before, but had seen three other Pearson scoring centers. This one is without a doubt the best and most conducive for scoring. The facility is spacious and not crowded as some others have been. Scorers have room to spread out if they wish to do so or to be nested with colleagues if that is their preference. I was struck that the large, bright monitors are excellent for scoring essays and are certainly to be preferred over laptops. I observed a number of essays and they were generally quite easy to read, and that makes the scoring process go easier. My observations, described again below, also indicated that the scorers were taking their work very seriously and appeared quiet and focused. Scorers were also able to increase the size of the handwriting, making the process easier when handwriting or wording is somewhat difficult.

The facility is obviously an unusual one; it is the top floor of a building that used to be a furniture store. However, that means that the ceilings are high, there is a large atrium-like skylight across much of the ceiling, and two very pleasant places that can be used by the scorers during breaks and for lunch. The walk from the front door and rest rooms to the scoring area is perhaps 50-75 yards, which gives people a nice chance to stretch their legs a few times each day, although some scorers complained about the distance a little. The space is also quite quiet and conducive for the type of work required of the scorers. There is a discount movie theatre at the opposite end of the fourth floor of the building, and I could not hear any sounds at all coming from that facility. (Indeed, it is at the opposite end of the building, perhaps 100 yards and some walls away.)

I understand that the 95 scorers present were more than 90% of the full complement of scorers. I was informed that they are ahead of their schedule and have only permitted 2 days of overtime to the scorers thus far. Supervisors sometimes worked overtime resolving nonadjacent scorings (resolution scoring) while the scorers were not there and on the 18th during the conference call, it was asked that the supervisors come in for this purpose. At the conclusion of the daily training/calibration on the 17th, one individual asked a question about the possibility of overtime the upcoming weekend.

Security

The security at this facility is quite remarkable. All employees including scorers wear picture id cards at all times. There is a manned desk immediately by the door to the scoring space. During normal working hours, one cannot get into the space without passing through that doorway and I understand that the other doors to the building are locked at all times. Visitors must sign in and be expected to gain entry. In addition, the site manager, Ms. Teran, showed me the two wall-mounted closed circuit cameras that are aimed at the lobby (at the entry area outside the scoring area), at the door of the scoring area, and in other places. There are also motion detectors. Although there are other doors to the building as a whole (I was told this but did not walk through the closed section of the building to count them or ascertain that they were locked), I was informed that Pearson keeps the doors locked and has the doors rekeyed each year. It would appear to be a very secure space.

Materials and Arrangements

Each scorer works behind what is or amounts to a terminal that has a large Dell computer screen. They work at these work stations that are lined up in fours or fives using the ePEN system to score the 10th grade essays. In some instances, all the workstations—which are about the size of a small rectangular kitchen table—in a row are used, in other cases, they are empty. There is also a whole room of terminals that is not being used presently, but it was used during training. That they have this extra space would be ideal if more scorers had qualified to serve. I did not know why the spacing of scorers throughout the room was so uneven, whether their desks were where they were first during training, or whether they simply sat in teams in a manner they preferred. The issue is probably of little consequence. One of the scoring directors told me that because of the loud speakers in the front of the room, scorers do not wish to sit too close to the front. The scorers sit with their teams and near their supervisors.

Each scorer has a keyboard and the screen and they can enter their scores using the keyboard after they read the essays on the screen. One advantage of these screens that I noticed is that the scorers can change the size of the handwritten essays they are scoring to be more

certain of what students are writing, or to make the essay more legible. The scorers are spaced very reasonably in a bright and comfortable room. They have scrap paper available to make notes and various office supplies in close proximity, usually in front of the supervisor. Their supervisors are nestled among them, typically, it appears at the end of the rows of tables. The screens are 23 inches and they are large and quite bright. Scorers read essays and input the scores that they assign to the essay using the keyboard. It is an efficient process, made more so by these excellent monitors and terminal combinations.

The scorers all have notebooks with the anchor paper sets available. Certainly, the vast majority of the scorers had their books open in front of them during the whole day. They refer to the anchors often. Watching them, one can see a scorer read the essay, then refer to the anchors, turn pages back and forth in the anchor book, and then look back to the essay. This iterative process continues until they assign a score. I assume it happened even more often earlier in the scoring process as these scorers are now quite experienced on this rubric and essay prompt.

Each day began with some instruction and reminders related to the rubric. The staff decided what score point may need the most attention on a given day and the next day it was discussed. After this instruction, the scorers evaluated one or more calibration papers that have previously been identified with what is believed to be that essay's "true score." The percentages of scorers who get the correct score for the question are determined and announced at the afternoon conference call. It is also determined and shared whether those who answer incorrectly have given a higher or lower score. Such feedback helps the Pearson team to plan future instruction. The first day I attended (April 17), the instruction took perhaps 20 minutes. On my second day, the instruction went for just over 30 minutes and was quite informative and even lively. It began with some reminders of scoring procedures, and subsequently, an essay was read aloud by the lead Scoring Director Alexandra Atrubin, who was the instructor that day. All of the scorers could also see the essay on their own screens. The paper that day was a "low 5.".The scorers all reviewed it and entered their scores. A show of hands indicated that the vast majority scored the essay a 5, a very few a 6, and about the same number a 4. That type of scoring was indicative of reliable and valid scoring. Ms. Atrubin continued after this to discuss any issues the scorers had; she justified the score of 5 and then talked about the distinctions between essays that were 3s and 4s, the so-called 3-4 line.

The scorers work from 8:00 am through 4:30 pm. They receive one half hour for lunch and two 15-minute breaks, one in the morning and one in the afternoon. Brief conversations with about a dozen scorers indicated to me that many found the job challenging, all found it important, and they worked diligently. Some said that they experienced eye fatigue near the end of the day. I am sure that the monitors reduce this to a great extent.

Throughout the day, scorers occasionally ask their supervisors for advice on specific essays. The supervisors spend a considerable amount of time back-reading the essays scored by

the individuals in their scoring team. They then provide suggestions and guidance to their scorers.

Results

The quality control and scorer monitoring procedures included *validity papers, calibration papers,* and *backreading*. Validity papers represent prescored and preselected essays that were agreed upon by Pearson and the FDOE. These prescorings were probably based on essays written during pretesting. Scorers must score 70% of these papers exactly as the assigned score to continue to qualify for scoring. Scorers must also demonstrate 55% interrater reliability, which is defined as scoring 55% of their essays the same as the second score assigned to these essays by other scorers. Scorings supervisors also backread and scored essays scored by their team members, focusing particularly on members whose validity or interrater reliability statistics suggest they are falling off the rubric. At least 5% of a scorer's essays are backread, but Supervisors are likely to read more for scorers for whom they are most concerned. Calibration papers focus on specific "lines" in the score rubric and are chosen to coach scorers for specific problem areas noted during recent scoring (e.g., the previous day). Thus, there are multiple types of quality control. For example, each scorer is evaluated regularly in terms of their individual IRR rates and validity scores. In addition, the supervisors regularly review the work of each scorer and provide feedback to them in this regard. In addition, of course, there is regular training to make certain that the scorers continue to understand and use the rubric as it is.

During the daily phone call, it was reported that the IRR rate here was running at a cumulative 56% accuracy and had been operating at 57% and 59% on the 16th and the first part of the 17th, respectively. Their rate on the 17th and the first part of the 18th was 59% and 59%, respectively. Their cumulative on the 18th was 56%. They had a validity index of 81% for the 16th and the first part of the 17th% also achieved 81%. Their validity index on the 17th and the first part of the 18th were 81% and 83%, with a cumulative validity of 77%. Moreover, it was reported that they are ahead of schedule and have operated with only 2 days of overtime for the scorers to date. Supervisors may have some additional opportunity for overtime, and all scorers who achieve a certain degree of accuracy in scoring are permitted either to come in an hour early to work or to stay an additional hour late at the end of the work day.

Recommendations

The scoring of the 10th grade FCAT 2.0 Writing Test in Tucson is functioning smoothly from both operational and psychometric grounds. There are few recommendations to make. Nevertheless, two are made below.

1. I recommend that the sites that are still using laptops switch to the full-size screens as soon as practicable or possible. Having seen both, this system is far superior.
2. During each morning's instruction, the scorers read an essay that has already been given a "true score", and they are asked to score it. On the 18th, after they had indeed scored it (or at least after they reported that they had scored it), the instructor told them what the score was and then asked for a show of hands to see who had assigned it that score. Two recommendations flow from this procedure.
    a. First, I think it would be better if the instructor asked them who scored a 1, a 2, a 3, and so on. This strategy would probably yield a more accurate assessment of how scorers are performing their duties.
    b. It appears that a scorer could enter his or her response after the score was announced. If so, these overall percentages were not entirely accurate.