BUROS
CENTER FOR TESTING

2014 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.

Consultants to the Florida Department of Education

May, 2014

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

Introduction

For the past five years, the Buros Center for Testing (Buros) has conducted annual reviews of selected components of Florida's statewide assessment system for the Florida Department of Education (FDOE). In 2014, the annual review focused on three components: 1) the preparation and scanning of test documents, 2) the calibration and equating of test scores, and 3) the handscoring operations for the FCAT 2.0 Writing assessment. This document presents the report of the Buros review team concerning the third review component.

The review activities included a review of the *Handscoring Specifications* document and operational handscoring statistics, as well as extensive on-site monitoring during seven site visits to facilities operated by FDOE's test contractor, Pearson. One two-day 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 Austin, TX, from March 6 to 7. Two three-day site visits were made to Pearson facilities in Jacksonville, FL, for Grade 4; Kent, WA, for Grade 8; and Tucson, AZ, for Grade 10 to observe the initial scorer candidate training from March 11 to 13 and subsequently to monitor ongoing operational scoring processes during March and April. Buros also participated in weekly (February 5 to 26) and daily (March 4 to April 25) conference calls with representatives from FDOE, Florida's Test Development Center (TDC), and Pearson. The following report presents the observations and findings of the Buros review team concerning the quality and integrity of operations for the handscoring of FCAT 2.0 Writing responses.

Test Document Preparation and Scanning

A special two-day site visit was made to observe the test document preparation and scanning operations at the Pearson scanning facility in Austin, TX, on March 6 and 7, 2014. During the visit, Buros observed all major operations, including receipt, batching, and preparation of test documents; test document scanning and editing; large print test document transcription; and storage of test documents. In addition, various documents including alert logs, procedural specifications, and operations plans were reviewed, and a scan accuracy check was performed on a small sample of test documents. Overall, the document preparation and scanning operations performed by Pearson were found to be highly secure and produced accurate electronic records of the test documents. All operational requirements specified by FDOE appear to have been met with appropriate fidelity and quality.

Scoring Time Frame and Timeline

Scoring activities for FCAT 2.0 Writing were scheduled for the time period of March 4 to April 24, 2014. Activities began with a three-day training and qualifying workshop for a group of 19 scoring supervisor candidates at each scoring site. During the following week, a three-day training workshop was held for scorer candidates followed by two days of pseudoscoring and two days of qualifying rounds. (Pseudoscoring would appear to be actual scoring for all practical purposes, but the scores do not count. It is, in effect, practice scoring.) This year the scorer training and qualifying activities were very successful. Each scoring site exceeded the respective

target number of qualifying scorers, thus avoiding the need for additional training waves. We believe that this increase in numbers helped improve efficiency this year.

Live scoring activities began on March 20, 2014. Scoring of Grade 8 and Grade 10 essays was completed four and seven days ahead of schedule, respectively. Grade 4 scoring was completed on the target date of April 24, 2014. For the duration of the project, the Pearson program team in collaboration with the TDC was proactive in ensuring timely completion of the project. Pearson and TDC monitored completion rates on a daily basis. In addition, overtime was offered at all sites to highly qualified scorers and supervisors in order to make up ground when necessary. These measures allowed Pearson to stay within and even ahead of schedule.

Scorer Recruitment and Qualifying

Scorer recruitment and training for the 2014 FCAT handscoring were very successful. All scoring sites were able to meet or exceed the target number of scorer candidates invited to the training. These scorer candidates were pre-screened by Pearson to ensure they met the educational requirements for scoring (Each scorer must have a degree in a field related to writing, e.g., English, Journalism). In addition, TDC representatives verified the educational qualifications of a random sample of approximately 20% of scorers.

Although the actual number of candidates who started the training was somewhat lower than expected due to “no-shows,” it had no impact on the overall qualification rates of scorer candidates, which were well above expectations for all three grade levels. For Grade 4 scoring, 154 candidates qualified for scoring, which is about 9% over the target number of 141 qualifying

scorers. Similarly, 154 candidates qualified for Grade 8 scoring, exceeding the target of 146 by more than 5%. For Grade 10 scoring, 159 candidates qualified, exceeding the target of 134 by more than 18%. The high success rate for the 2014 scoring of the FCAT 2.0 Writing is commendable. Based on individual conversations with TDC and Pearson representatives and information shared during the daily conference calls, we gained the impression that Pearson was able to recruit a large number of experienced scorers for this year's scoring. If this perception is correct, it may have been a contributing factor to the high qualification rates and may also have boded well for the quality of scoring overall. Having experienced and seemingly high-quality scorers also means that these individuals know what is involved in scoring and are less likely to drop out during the process.

In addition to scorer candidates who met the regular qualification standards for scoring, a small number of scorer candidates were retained who met slightly lower provisional scoring standards. Under regular qualification standards, candidates must obtain a 70% agreement rate across two sets of qualifying essays and no more than one score disagreement that is non-adjacent. Under provisional qualification standards, candidates must obtain at least a 65% agreement rate with only one non-adjacent score or a 70% agreement rate with no more than two non-adjacent scores. Although the number of regularly qualifying scorers exceeded the target number at each site, a decision was made to allow provisional scorers to continue on the project. These scorers were carefully monitored and had to maintain the same scoring quality standards for live scoring.

We understand the practical benefit of retaining provisional scorers on a scoring project, in particular in situations when the overall scoring schedule might be in jeopardy as has been the case in previous years. Furthermore, because provisional scorers must maintain the same quality

scoring metrics as those who qualify under regular standards, there is no particular concern that those scorers might negatively affect the accuracy of scores. From the perspective of the results, the retention of provisionally qualifying scorers did not jeopardize overall scoring quality and we understand that this result in terms of the quality metrics is consistent with past practice. The handscoring project requirements did set specific qualifying standards for scorers at the outset of the project to ensure that qualifying decisions are not arbitrarily made but also accepted some provisional scorers whose scoring accuracy appears to have been as accurate as those fully meeting the standards. Provisional and regular qualifying standards are high and meet industry standards.

Scorer Training

The training process followed the same general protocol that had been used successfully in previous years and was the same at all three scoring sites. The lead scoring directors for each grade level conducted the majority of the training beginning with a morning orientation on the first day followed by a thorough introduction to the scoring materials and scoring project. The majority of training time was used for extensive practice scoring, which required approximately two-and-a-half days. Representatives of the Florida Department of Education were able to emphasize certain points and underscore the importance of the process and the need for quality scoring. Their roles were important ones.

The morning orientation provided a general context to the scoring project and covered various logistical and security aspects of the training. Before details of the scoring project were discussed, scorers were required to sign a non-disclosure agreement form and were instructed on

the various security procedures implemented to keep test materials secure. Candidates received information about the quality management plan that specifies the qualification requirements to become eligible to score FCAT 2.0 Writing and to maintain eligibility. The training then focused on the details of the scoring task beginning with an overview of the various forms of reader bias, which alerted scorer candidates to their own potential for biased perceptions of student writing. Scoring directors introduced the grade-specific writing prompt along with a discussion of the range of allowable interpretations and the holistic scoring method used to score FCAT 2.0 Writing. Considerable time was spent explaining the six-point scoring rubric used to assess the student essays with detailed explanations given on the four rubric elements that define each score point. A helpful discussion of strategies for successful scoring concluded this segment of the training.

After the introduction to the scoring project and scoring task, scorer candidates were then introduced to the grade-specific set of anchor papers that are used to exemplify the six score points of the scoring rubric. These anchor papers serve as the primary guide for correctly applying the scoring rubric to the writing responses and stay with scorers for the duration of the project. The sets typically comprised a total of 18 student responses, three for each of the six score points. At each score point, anchor papers were ordered from low to high in order to illustrate a progression of writing performance. For example, the three anchor papers for a score of "4" represented a low 4, a middle 4, and a high 4. The Grade 4 set included only one example of a score point 6 at the beginning of scorer training. However, by the time live scoring began, two additional examples had been identified and were added to the anchor paper set. Overall, we found that the anchor paper sets were carefully selected and exhibited a clear and logical progression of writing performance.

To provide additional guidance to scorers, each anchor paper was given a set of annotations that explained the score point rationale in terms of the four rubric elements. During the training, scoring directors first read each individual anchor paper aloud and then discussed the annotations providing further commentary and clarifications. Once the entire anchor paper set was introduced, scorer candidates were able to ask questions or provide commentary before the training moved to practice scoring.

To give candidates extensive opportunities to learn and practice the correct application of the scoring rubric, five sets of practice papers had been assembled. The first two sets comprised five practice papers that covered a limited score point range. The first set focused on score points 1 through 3, and the second set focused on score points 4 through 6. Before attempting each practice paper set, scorer candidates were informed about the relevant score point range for each set. The last three practice paper sets comprised 10, 15, and 20 papers covering the entire score point range. Papers were selected to approximate historical score point distributions with a larger number of papers representing score points 3, 4, and 5 within a set. While sets 3 and 5 included mostly typical student essays, practice paper set 4 could also include a few examples of less typical student work.

For each practice round, scorer candidates were given sufficient time to complete the set. Practice rounds were generally very orderly and quiet. Candidates who finished early were asked to leave the room and to return at an agreed time. Scorer candidates marked their scores on a score sheet that was collected by their assigned scoring supervisor, who returned the sheet with the scorer's percent of agreement (perfect and perfect plus adjacent) achieved for the practice set. Scores were also entered into a spreadsheet that allowed the Pearson program team to monitor scoring performance for the entire pool of candidates. This information was also very useful to

scoring directors who could target specific practice papers during the subsequent discussion of each practice paper set. During these discussions, scoring directors used prepared notes with discussion points similar to annotations to guide the review of the paper. Scoring candidates were encouraged to ask questions that scoring directors generally answered willingly and appropriately. For all discussions, candidates were asked to refrain from questioning the validity of a score point or to engage in hypothetical discussions. (The scores assigned to the anchor papers were previously agreed on during a Rangefinding process involving Florida language arts supervisors and grade-level appropriate Florida classroom teachers.) Those reminders helped to keep the discussions productive and focused on the scoring task. In general, scoring candidates seemed motivated and engaged. They remained attentive and asked appropriate questions. Many of them marked up their materials liberally and kept notes. It was clear to us that the majority of candidates were eager to succeed at this task.

After the completion of the three-day scorer training workshop, candidates were given additional time to practice scoring before they moved on to the qualifying process. This modification to the training process was introduced last year and appears to be quite effective in allowing scorers to become accustomed to the scoring task and to the computer environment in which scoring is carried out. Candidates are first introduced to the ePEN scoring system through several training modules aimed at familiarizing candidates with the environment and the various tools and quality control mechanisms available therein. To continue practice scoring, which at this point is called pseudoscoring, candidates score actual student essays from the 2014 operational test administration. However, these scores are not recorded. As noted previously, the scoring process mimics actual live scoring in that the same scorer calibration and quality

monitoring tools are in use. Candidates spent a total of two days of pseudoscoring before they moved on to the qualifying rounds, which were also administered in ePEN.

Overall, the training process was well organized and competently carried out. The high qualification rates for all three grade-level scorings speak to the high degree of success of the training. We know that English Language Learners (ELLs) make up a significant proportion of the students in the Florida schools. We know and agree with FDOE's policy to grade such papers like all others; we believe that such an approach is consistent with national standards and good practice. We hope that as FDOE transitions to a new contractor, that they maintain or even enhance this concern in the training of scorers as not all may enter training aware of the diversity of ELLs. Given the high percentage of ELL students in Florida, we believe that thorough discussion of their scoring is important in training and throughout the scoring process. In general, we believe that FDOE has addressed this concern but wish to make sure it is carried forward to a new vendor.

It is a great positive that FDOE has developed a program for identifying troubled children as part of the handscoring process. It would be interesting to collect and organize such data as part of the summary data as this step is a significant step toward quality in Florida, even though it does not address the writing ability of students directly. The collection and analysis of such data might permit the State to improve its ability to identify such students, such as by changes to its training materials, in this regard over time.

In several instances during training and operational scoring, we have heard questions from scorers about the common differences between score point levels that would help them differentiate adjacent scores. The response that we have heard to such questions is that the

trainers are instructing scorers on what each of the different scores are, and that the scorers should learn to identify 1, 2, 3 and so on. We understand that this response is consistent with how FDOE has defined holistic scoring. Such a listing of differentiations is not determinative but would simply aid the scoring judgment process. While we understand that FDOE has traditionally used "focused holistic scoring," we have also seen situations where holistic scoring is used with example grounds for differentiating papers in adjacent score points. We believe that the concept learning literature demonstrates clearly that teaching people to differentiate differences among different concepts is quite helpful to their learning the various concepts. We believe that the range-finding panel could, perhaps given some extra time, identify common differences, either qualitative or quantitative, that would help scorers differentiate adjacent score points. We do not believe that adding such information changes the scoring process away from being holistic scoring. The goal of this recommendation is simply to improve the already high accuracy of essay scoring. Moreover, as we understand that FDOE is moving to analytic scoring in the future, we hope that our mentioning this minor paradigm shift at this time permits FDOE to explore the options to continue their record of highly reliable scoring. Scoring essays is hard work and doing so reliably is even more so. Whatever aids the scoring vendor or the State can provide the scorers would be welcomed, we are sure.

Operational Scoring Process

All three scoring sites met or exceeded the project-wide quality standard for interrater reliability (IRR) and validity agreement. The target for interrater reliability, which is defined as the percent of agreement between the first and second score assigned to each student essay was

60% agreement for Grades 4 and 8 and 55% agreement for Grade 10. Validity agreement is defined as the percent of agreement between the score assigned by individual scorers and the pre-assigned validity score that represents the paper's correct score. The validity target was 70% agreement for all three grade levels. Both Grade 4 and Grade 8 scoring met the IRR target of 60% and exceeded the requirement for validity agreement with 77% and 75%, respectively. Scoring for Grade 10 was similarly successful exceeding both the IRR target with 58% agreement and the validity agreement target with 79%. Both quality metrics were very stable for the duration of the scoring project suggesting that high-quality scoring was sustained throughout.

In addition to scoring quality metrics applied to the entire project, individual scorers must also maintain minimum scoring quality standards to remain on the 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 match throughout. Scorers who fall below these standards are issued a warning and receive remediation training. If scoring accuracy does not improve after an additional 10 validity papers are scored, scorers receive a 10-paper scorer exception set, which they must pass with a 70% exact and 100% exact or adjacent score agreement. Scorers who do not pass the scorer exception set are released from the project, and all their previously assigned scores are reset. This measure ensures that all student essays are scored by qualified scorers and is a strong statement in terms of the importance of accurate scoring.

Implementation of the validity agreement rate as a quality control tool requires that scorers read one validity paper out of approximately 10 student essays. This insertion rate into the scoring queue necessitates that the pool of validity papers is continually replenished and that those papers that had been seen by scorers no more than twice are appropriately retired. Both FDOE representatives and Pearson staff were very diligent in maintaining the validity paper

pool, especially in the beginning of the project when the status of the validity pool was addressed during daily conference calls.

Pearson uses a number of additional quality monitoring and training strategies to ensure that scorers use the scoring rubric accurately. Both scoring supervisors and scoring directors have access to a variety of ePEN-generated scoring quality reports that are used to identify individual scorers and group-wide trends of scoring performance issues. Scoring supervisors monitor their scorers' interrater reliability (exact or adjacent score agreement) and backread at least 5% of scored essays, and on average among the three sites, approximately 16% of responses were backread. To ensure that scoring supervisors maintain scoring quality, they are required to score student essays for a minimum of one hour each day to allow computation of quality metrics on their scoring performance. Scoring directors backread the work of scoring supervisors and monitor their quality metrics as well as project-wide quality metrics. In particular, they are tasked with identifying specific calibration needs and administer calibrations on a daily basis to the total pool of scorers and to smaller groups of targeted scorers, if needed. Calibrations are administered via ePEN or on paper, usually twice each day. Each day, scoring directors or scoring supervisors conducted anchor reviews in the morning to ensure that scorers remain focused on the anchor paper set when scoring student essays.

The measures implemented by Pearson to monitor and maintain scoring quality are extensive, and we commend Pearson and especially the scoring directors for their substantial efforts to ensure that overall scoring quality is maintained. We found these measures to be very effective and a major contributing factor to the high scoring accuracy achieved for each grade level scoring.

Security of Scoring Materials and Information

Multiple security measures were implemented at the three scoring sites as well as at the two off-site scorer training locations. All Pearson scoring personnel, including scorer candidates, scoring supervisors, scoring directors, and project team members, as well as visitors and FDOE representatives, were required to wear ID badges while on the premises. Sign-in and sign-out procedures further ensured that access to the sites was controlled and only authorized individuals were allowed to enter. Pearson site staff constantly monitored entrance ways and exits. Overall, the facilities were quite secure.

Scoring materials such as anchor paper sets and annotations remained at the site and were stored in a locked room overnight. Staff and visitors with access to scoring materials were required to sign nondisclosure agreements and were not allowed to take any personal notes about the project home. In one instance, a scorer candidate training for Grade 10 scoring was released from the project after it was found out that he took his personal notes home. The case demonstrates that the security procedures were appropriately enforced. Overall, we found the security measures to be comprehensive and effective in keeping scoring materials and scoring information secure.

We note that access to the ePEN system was password-protected. However, we cannot generally comment on the security of the ePEN system as that exceeds the scope of this review.

FCAT 2.0 Writing Scores

The 2014 administration of the FCAT 2.0 Writing assessment saw some improvement in student performance for Grades 8 and 10 but also a decline in performance for Grade 4. In Grades 8 and 10, the percentage of students scoring at or above 3.5 both increased by 2%. In Grade 4, the percentage of students at or above 3.5 decreased by 4%.

In general the observed year-to-year changes in student performance in Writing are well within the range of score fluctuations that can be expected of assessments that are human scored. Given the large improvement in Grade 4 Writing in 2013, which was substantially larger than in the other two grades, the decline in Grade 4 performance in 2014 could also be viewed as a normal correction to last year's trend. Based on our observations and review of the 2014 scoring operations, we found no indication of improper scoring procedures. Furthermore, there were no changes to the way the test was administered this year to which the decline in student performance might be attributed. As with previously observed score fluctuations, Buros considers those changes to be an expected outcome of assessments that rely on human scorers. Moreover, there is simply no way to "equate" either essay prompts or the essays themselves in a manner that two multiple-choice exams can be equated for difficulty and scoring. Given the judgment involved in the assignment of scores, small variations will occur, and this year it appears that they may have.

Procedures for Exceptional Papers

The FCAT 2.0 Writing assessment offers several accommodations to students with disabilities, including accommodations that allow students to write their responses in non-standard formats such as large print. Transcription and conversion processes are in place that permit the reading and scoring of these special response documents through regular scoring operations. These efforts reflect a strong commitment to integrating students with disabilities into Florida's statewide assessment system, which is consistent with the best intents of No Child Left Behind.

Each year, the FCAT 2.0 Writing assessment elicits student responses that contain statements suggesting the student may be in danger of child abuse or neglect. Special procedures have been developed whereby a student essay that has been identified by a scorer for its troubling content can be alerted to FDOE representatives. Buros considers the implementation of the troubled child alert system an important measure that helps to safeguard Florida students. We commend the state for having the alert procedures in place.

Conclusions and Recommendations

The 2014 scoring of the FCAT 2.0 Writing marks the final year in the partnership between Pearson and the Florida Department of Education. During the past five years in which Buros served as an independent, outside observer of the scoring operations for this assessment, we have found this partnership to be very successful, and 2014 is no exception in this regard. We continue to be impressed by the dedication and commitment of the staff from the Florida

Department of Education, Florida's Test Development Center, and Pearson to ensure that each student's writing ability is accurately and fairly assessed. We wish to commend the Pearson employees especially this year. The scoring process occurred after it had been announced that another vendor would be providing assessment services for the Florida Department of Education in future years. Yet, we observed no difference in the Pearson employees, who continued their work with professionalism and dedication that is the hallmark of a successful and quality-oriented company.

Scoring operations in 2014 were very successful in terms of recruiting a highly capable pool of scorers who maintained a high standard of scoring quality. Overall we found that the scoring processes were effectively organized and adhered to the best practices in the field. While we believe that the State of Florida and its vendor, Pearson, have engaged in a highly professional essay testing and scoring process, in the following, we wish to offer a few minor recommendations and suggestions that may be helpful in guiding future handscoring projects.

Recommendations:

1. Continue to have Florida Department of Education officials work as part of the vendor teams during the supervisor and scorer training. Their input contextualizes the importance of the process and makes the work more than employment.
2. Continue the incorporation of an extended pseudoscoring period prior to the administration of qualifying rounds. The additional time spent in pseudoscoring appears to improve scoring performance of scorer candidates and may contribute to higher scorer qualification rates.

3. Given that FDOE is moving to a new approach to scoring essays with its new vendor, it should explore developing guidelines for differentiating adjacent score points so that the common differences, for example, between 3 and 4 and so on are explicated to the extent possible by essay scoring experts, such as those involved in the range-finding process currently.