BUROS
Center for Testing

Report on the Scoring of the FCAT Writing Assessment

Kurt F. Geisinger, Ph.D., Anja Römhild, M.A., and Tzu-Yun Chin, Ph.D.

Buros Center for Testing

Consultants to the Florida Department of Education

May 2012

With questions or comments, please contact:
Kurt F. Geisinger, Ph.D.
Kgeisinger2@unl.edu
(402) 472-6203

Report on the Scoring of the FCAT Writing Assessment

Kurt F. Geisinger, Ph.D., Anja Römhild, M.A., and Tzu-Yun Chin, Ph.D.

Buros Center for Testing

Consultants to the Florida Department of Education

May 2012

As part of the Buros Center for Testing's analysis of various aspects of the 2012 Florida Comprehensive Assessment Tests (FCAT) for the Florida Department of Education, members of the Buros audit team conducted a comprehensive review of the FCAT Writing handscoring process. The evaluation included a review of *Handscoring Specifications* documents and operational handscoring statistics, a two plus-day site visit to monitor the scanning and preparation of FCAT Writing test documents, and two three-day visits to each of the three FCAT Writing scoring sites to observe the initial training of scorer candidates from March 12 to 14, and subsequently to monitor ongoing scoring operations. These site visits are summarized in Appendix A and B of this report. Our staff and subcontractors also participated in daily calls between State Department of Education officials and leaders at Pearson, the State's scoring contractor. This report summarizes our impressions of the quality of the scoring of FCAT Writing responses.

The organization of this report largely follows that of the visits. We discuss the time frame and timeline, the training of scorers, the supervisors, the scoring

itself, retraining of supervisors and scorers, standards for reliability, standards for validity, and ongoing monitoring of the scoring process. Finally, we provide some comments about this year's student performance and district accountability measures (school grade calculations), as compared to the past year's scores.

Scoring Time Frame and Timeline. The time frame for the FCAT Writing scoring was documented in various contracts and notes. The work began in the scoring centers on March 5, 2012 and was completed May 2, 2012, approximately a week and a half after the scheduled completion date of April 20[1]. The delay in completing the FCAT Writing scoring primarily occurred with the Grade 4 and Grade 8 scoring. The Mesa, AZ, site, which scored Grade 10 essays, finished within the agreed-upon schedule. The Jacksonville, FL, site finished scoring the Grade 4 essays on April 24, and the Auburn, WA, site finished scoring the Grade 8 essays on May 2.

We believe that the original time frame, though ambitious, is doable when one is working with an experienced team of scorers of written examination essays. Pearson experienced some difficulties recruiting an adequate number of excellent and experienced essay scorers, which extended the duration of the project. Pearson, in conjunction with Florida Department of Education professionals, responded swiftly when the difficulty arose, holding additional training sessions and offering overtime to qualified scorers and supervisors to meet the demands. It might be

[1] The completion date was originally set for April 18 but was extended by two days when it became apparent that an insufficient number of scorers had passed the scoring qualification criteria during the initial scorer training.

noted that all candidates to become scorers did meet stringent scoring criteria in order to be hired and needed to continue to meet exacting standards to remain on the project. At this point we can only speculate as to why fewer experienced scorers chose to score essays this year. We believe that the improving economy may have resulted in some previous years' scorers obtaining other positions. Additionally, a good proportion of the scorers are retired teachers, who may have decided to forgo the work this year.

Scorer Recruitment and Training. All scoring centers experienced an initial shortfall of scorers able to meet the qualification criteria at the end of the first wave of training, which was held one week prior to the start of operational scoring. The sites had recruited 134% (Jacksonville, FL), 135% (Auburn, WA), and 155% (Mesa, AZ) of the planned number of qualifying scorers needed for the project. In years past, these numbers would generally have been adequate to meet scoring needs. Despite recruiting well above the target, only 87% of the needed target number qualified for Grade 4 scoring, 64% for Grade 8 scoring, and 76% for Grade 10 scoring. In response to the lower than anticipated qualification rates, additional training waves were conducted as operational scoring began.

For Grade 4 and Grade 10 scoring, Pearson held two additional training waves during the first and second week of the scoring window. For Grade 8 scoring, a third additional wave was needed, which was held during the third scoring week. Both, the Mesa (Grade 10) and Auburn (Grade 8) sites invited back candidates who had participated in the first training wave and who narrowly missed the

qualification criteria. These repeat candidates made up all (48 of 48 in Mesa) or a substantial portion (70 of 92 in Auburn) of the second wave candidates. They were given the same training and qualifying materials as the first training wave with a few modifications (i.e., some papers were exchanged or presented in different order).

The reuse of training materials with repeat candidates is something to reconsider in future years, at least when making the determinations that a candidate is qualified to be a scorer. The repeat candidates may be more likely to meet the qualification standards by remembering a correct score rather than by assigning that score based on their own judgment; indeed, for at least some of the training and test essays, explanations as to why they have been scored as they were are provided to the scorer candidates. We do understand that in some cases, some essays were replaced in a qualification set from one wave to the next and that the essays were re-ordered. However, a scorer might, for example, remember how they scored an essay previously and what the correct score for a given essay was.

Members of the Buros audit team attended the initial scorer training at each of the three scoring sites and found them to be professionally conducted. Appendix A includes site visit reports summarizing our observations of these scorer training sessions. Initial training, whether one has previous Pearson scoring experience or not[2], occurs over a three-day period for both scoring supervisors and scorers, with supervisors trained and selected first. After this training, supervisors and scorers must pass a qualifying test with rigorous standards for scoring accuracy. The

[2] One observer was informed that over one-half of the scorers had previous Pearson essay scoring experience.

qualifying test involves having the supervisor and scorer candidates score carefully selected samples of essays that were previously scored by a panel of experienced, expert scorers as part of the rangefinding work. The scores assigned by the potential supervisors and scorers are compared to the scores provided by the expert panel. Supervisors must take three qualifying sets of essays and achieve an exact agreement of 75% on two sets, with none of the three sets below 60% exact agreement. The scorers, as opposed to the supervisors, need only reach an average of 70% agreement across two sets of essays, which we believe is a demanding standard. They also must have 95% agreement within one point of the intended score, which means that to qualify as a scorer, only one of 40 essays across the two sets of essays can differ from the expert panel by more than one score point. These criteria assure that the scorers are able to score essays accurately.

Every day during the actual scoring, the leaders at each site provided focused training on specific scores (e.g., scores in the middle of the distribution or scores that are low within their score band). It is likely that such continuous training kept the rubric centrally in the minds of the scorers. Finally, the Pearson staff reviewed the accuracy of each and every supervisor and scorer on a daily basis and shared summary data with Florida DOE staff and their consultants on a daily basis. If an individual scorer's statistics did not meet specific criteria for accuracy, he or she had to participate in retraining measures. If scoring accuracy did not improve, scorers were removed from the pool and the scores they had assigned up to that point were removed as well; these essays were then rescored. In some cases,

scores were removed if a scorer left the project for a different reason but his or her low scoring quality indicated that release from the project was imminent.

Overall, the approach to accumulating scores for the essays was a careful one. In general, we believe that this approach to the holistic scoring of essays was professionally done and was in keeping with the best practices of the profession and of the assessment of writing in educational systems. We commend both the Florida Department of Education and their counterparts at Pearson for their professionalism and exactitude in processing these essays. We believe their review of scoring on a daily basis meets the highest procedural requirements of our field.

Supervisors. As noted previously, supervisors met higher standards to obtain their positions. They are expected to work with the scorers, especially those having some difficulties. Each supervisor oversaw approximately 10 to 15 scorers, which we believe is a reasonable supervisory ratio. Supervisors also "back-read" approximately 5% of the essays scored by those that they supervise. This process is intended both to check the scorers' accuracy and to provide the scorers with guidance when needed. We believe the back-readings in addition to the statistical indicators of scoring quality provide an effective monitoring system and help facilitate accurate scoring of FCAT essays.

The Work of the Scorers/Standards of Reliability and Validity. Approximately one in every seven essays that a scorer reads is a validity paper rather than an essay for operational scoring. Validity papers are essays that have

been pre-scored by scoring directors and approved by Florida Department of Education representatives. They are embedded into the operational scoring of responses to check that scorers continue to work in accordance with the scoring rubric. Scorers are blind to whether any given essay is operational or a validity paper.

Scores from validity papers are used to compute the validity agreement rate, the percentage of perfect agreement with the pre-assigned validity score. Scorers are expected to maintain a minimum agreement rate of 70%. If they fall below the 70% standard, more intensive monitoring and retraining measures are initiated. If a scorer falls below 60% validity agreement, the scorer is in jeopardy of being released from the project unless scoring performance improves over the next 10 validity papers or the scorer passes a 10-paper calibration set achieving at least 70% agreement with no non-adjacent scores. As noted previously, should a scorer be released from his or her employment for failing to maintain scoring quality standards, the scores that they had assigned heretofore are also removed from the scoring database and these essays are re-scored by other scorers.

In addition to the validity agreement rate as a standard of scoring quality, a measure of inter-rater reliability (IRR) is also computed. Because each essay is scored by at least two raters, the inter-rater reliability can be computed as the percentage of exact agreement between a paper's first and second score. The standard for inter-rater reliability is defined for the entire project, not for the performance of individual scorers. However, scorers with low inter-rater reliability

were targeted for increased monitoring and retraining. For the scoring of Grades 4 and 8, the project's inter-rater reliability target was 60%; for Grade 10 the target was 55%. The Grade 10 target was lower because high school student writing is more complex, and 55% IRR has been the historical trend for a number of years. These target values are comparable to inter-rater reliability rates reported for similar assessments, for example NAEP writing prompts. As such, these reliability estimates are reasonable. Table 1 below provides the inter-rater reliability rates for 2012 by grade.

The inter-rater reliability goal proved more challenging for the Grade 8 scorers than for those scoring Grades 4 and 10. In our opinion, the more difficult persuasive writing mode and complexity of the prompt in combination with the more stringent inter-rater reliability goal likely contributed to the difficulties experienced by the Auburn scoring site.

Half way through the scoring window, the Pearson team, at the request of the Florida Department of Education, initiated several measures aimed at improving the inter-rater reliability at all sites, but especially in Auburn. In addition to continued targeted monitoring of poorly performing scorers, additional calibration sessions and trainings were implemented. A training form called paired scoring (one-on-one scoring with immediate supervisor feedback) was initiated and was reportedly most helpful in improving scorer performance. Despite these efforts, the inter-rater reliability for Grade 8 scoring remained below its target for much of the project's duration. However, through the committed efforts of FDOE and Pearson,

the rescoring of papers resulted in the improvement of the agreement rate, which met the target value.

The use of the exact agreement rate of scores from two or more raters is a widely used and accepted indicator of inter-rater reliability, although it is not the only statistic available for that purpose. In our view, the exact agreement rate provides an incomplete assessment of reliability for the FCAT Writing assessment because it discounts the occurrence of adjacent scores, even though valid operational scores are produced from them through averaging. A supplementary indicator of scoring quality may be the agreement rate of perfect plus adjacent score points, for which explicit project-wide targets and expectations for scorer performance could be established (the current scorer disqualification criteria only focus on validity agreement rate). The perfect agreement index would need to be used alongside this latter index, as the use of perfect plus adjacent score agreement alone might permit a scorer to continue scoring essays even if he or she continually evaluated essays as one point too high or too low. In addition, as a project-wide standard for reliability, the Florida Department of Education and Pearson may wish to consider a standard reliability index, the intraclass correlation, which is a measure of the consistency among raters (Shrout & Fleiss, 1979)[3]. We believe that it is well suited to evaluate the reliability of the operational essay score that results from averaging scorer ratings, and could serve as a complementary index of

[3] The intraclass correlation for this case will be ICC(1,2) using Shrout and Fleiss' notation system (1979). Note, we do not recommend using the ICC(1,2) to evaluate individual scorers as the coefficient is dependent on the score distribution of essays.

reliability next to the currently used inter-rater reliability agreement rate. The intraclass correlation coefficient is one of the primary indices used in education and psychology to indicate inter-rater agreement. It provides an index that is directly comparable to a traditional reliability coefficient, such as coefficient alpha that is used for multiple-choice tests. Moreover, one can easily compute the reliability of a single scoring and that of the combination of two raters, so that one can see the improved reliability. It is not possible to provide the reliability of the combination of two scorers using the agreement percentages that the Florida DOE and Pearson have been utilizing. The index also takes into account when different scores for the same essay are more or less similar. (That is, a disagreement when one scorer has assigned a "3" and the other a "4" is less severe than when one assigns a "1" and the other a "5.") The standard IRR that has been used treats them both as equivalent. Finally, the intraclass correlation coefficient is preferred when base rates are high. Base rates are the percentage agreement due simply to chance. Given the frequency distribution of scores (which can be estimated using data from Table 2), it can be seen that the modal score is a "3." If scorers simply assigned every score a "3", the IRR would be 100% agreement, even though the validity of these assigned scores would be much lower. In this hypothetical case, the intraclass correlation coefficient would be 0.00. Please note that we are simply suggesting that including the intraclass correlations as an index of reliability would be a valuable addition because it is more directly comparable to a traditional reliability coefficient without encountering the problem of base rates present in agreement percentages. (We note

that sometimes people also use kappa coefficients to avoid the base rate concern, and while kappa coefficients have many advantages, they are not directly comparable to coefficient alpha, which is itself an intraclass correlation coefficient.)

Results of Scoring Performance. Table 1 below provides the targeted and achieved agreement rates for validity and inter-rater reliability for each grade level scoring.

Table 1

Inter-rater Reliability (IRR) and Validity Rates

	Grade 4	Grade 8	Grade 10
# papers scored	193,284	195,181	189,281
max # scorers on roster	156	227	236
max # supervisors on roster	18	24	21
project IRR (target IRR)	60% (60%)	60% (60%)	56% (55%)
project validity (target val.)	79% (70%)	75% (70%)	79% (70%)

It can be seen that the target inter-rater reliabilities (IRRs) were all met. The validity agreement rates were all exceeded by meaningful amounts. This latter value has a highly meaningful impact to the lowering of scores, as noted in the following section of this report. That between 75% and 79% of essay scores provided

by scorers matched those of the expert scorers indicates that the rubrics were followed in a highly consistent manner. The new and more rigorous scoring criteria were applied accurately and reflected the scores given to validity essays by the expert scorers.

2012 FCAT Writing Student Performance. Responses to a writing prompt are written by virtually all Florida students in fourth, eighth and tenth grades as part of the FCAT. All students in a grade level respond to the same essay prompt representing a particular writing mode or purpose (i.e., narrative, expository, or persuasive). The 2012 FCAT Writing asked fourth graders to respond to a narrative prompt, and eighth and tenth graders responded to a persuasive prompt. Table 2 below provides a listing of the numbers of students at each score point across the three grades. The average scores for fourth, eighth, and tenth grades were 3.25, 3.28, and 3.42, respectively, as can also be seen in Table 3 below.

Table 2

Percent Students at Score Point

Score point	Grade 4	Grade 8	Grade 10
0	1.03	0.77	0.53
1	1.41	1.7	1.25
1.5	1.4	1.51	1.32
2	5.91	8.48	4.58
2.5	9.47	10.14	8.28

3	32.48	25.44	24.09
3.5	21.02	18.96	22.31
4	19.25	23.11	25.92
4.5	5.29	6.23	7.8
5	1.9	2.63	2.84
5.5	0.62	0.74	0.84
6	0.22	0.29	0.24

Why were scores lower this year? Several changes were made to the scoring criteria for the 2012 FCAT Writing. Specifically, more stringent scoring criteria were introduced that expanded expectations concerning the correct use of English conventions and the quality of supporting detail. These criterion changes were built into the selection of anchor essay papers by the Rangefinding Committee. These changes made the FCAT Writing a more difficult test for students across all three grade levels. With the previous school accountability measure of the percentage of students scoring 4 and above still in place, it was more difficult for students to reach this criterion. It should be clear that achieving such a status is not "set in stone" and can change over time. Another significant change in 2012 concerns the re-introduction of double-scoring of essays[4] and the use of score averages as operational scores. This practice resulted in the possibility of half-point scores on the operational score scale. A half-point score occurs, for instance, when

[4] Double scoring of essays was practiced prior to 2010, but had been replaced with single rater scoring during 2010 and 2011. We appreciate that the State of Florida made the decision to have all essays scored by two trained scorers in 2012. Buros recommended that this practice be reinstituted last year, and Florida saw fit to do so.

one scorer assigns an essay with a "3" and a second scorer, blind to the first scoring, assigns a "4". In such an instance, the essay receives a score of "3.5". Students who received half-point scores this year would have been assigned to one of the two adjacent integers if a single scorer scoring protocol were used. As a result the change of scoring protocol may have had an impact on some students' meeting the proficiency standard.

What prompted the dramatic drop in students' proficiency rates this year? There is ultimately no way to answer this question with factual accuracy, and hypotheses are educated speculations. A few of the possible reasonings follow.

1. The essay prompts were simply more difficult.

2. The more rigorous scoring criteria implemented by the Department of Education in the attempt to increase standards affected the scores so that lower scores resulted.

3. Because the scoring standards were increased in 2012, the essays selected to be the anchor, training, and validity papers were assigned scores that were somewhat lower than what they would have been under the previous writing standards, and the operational scoring of the student essays duplicated this more stringent essay scoring.

4. The actual writing of students declined from one year to the next.

Of the hypotheses above, only the first and last ones can be largely discounted. Addressing the fourth, it is simply improbable that writing instruction and student quality across the state would have declined to such an extent across a single year. Without a catastrophe having occurred to the State and its educational system, such an explanation would defy logic and experience. Similarly, the essay prompts had been pretested during earlier testing years and had been found to generate score profiles similar to previous prompts. Thus, the first hypothesis can also be largely discounted.

If the current report writers had to make educated guesses as to why the decline in scores occurred, it would be a scenario like the following. The State Department of Education wished to implement higher writing standards, and they did implement these through raised expectations in the scoring criteria. Those new scoring criteria were reflected in the way that the range finding scoring team evaluated the essays used as validity and anchor essays as well as those used in the training of the scorers. The training of scorers was effective and therefore the scorers placed into practice these higher standards through the scores that they assigned to individual essays. Thus, the second and third hypotheses above are related. The state's more rigorous criteria were implemented and were reflected in the selection of anchor papers. The scorers graded reliably and validly using the above criteria and scores fell, even though it would appear that student writing performance was as good as it was the year before.

When a more rigorous scoring system is implemented, it makes year-to-year comparisons difficult under the best of circumstances. Let us provide an analogy to this situation. Imagine that the train system for a given city has been studied for on-time performance for years. They have identified trains as either on-time or late for 10 years using the criterion that if a train reaches its destination within 5 minutes of the scheduled time, it was seen as on-time. If it arrives after that time, it was identified as late. A new station master comes into office and wants to raise efficiency. He or she announces that forthwith trains will be considered on time if they are within 2 minutes of the scheduled time, rather than 5 minutes. If the percentage of trains identified as late increases, does that mean that the trains are running later on average? Well, it might be that, but it is impossible to tell from that limited information. That is the situation in which Florida finds itself. Only time will tell if instruction improves student performance across the state, but there is evidence in the literature that higher standards do lead to higher performance.

What should standards for the new and more rigorous school accountability measure for the FCAT Writing Test be? The answer to this question is a policy one rather than a psychometric one, and therefore, Buros will not answer it. However, we provide the data below in Table 3 indicating the percentage of students identified as meeting the designation of Proficiency given the cut scores of 3, 3.5 and 4.

Table 3

Proficiency Rates of Students (with descriptive score information)

Grade	cut score	% not reaching proficiency	% reaching proficiency	Mean
4	4	72.72	27.28	3.25
	3.5	51.70	48.30	
	3.0	15.97	80.78	
8	4	67.01	32.99	3.28
	3.5	48.05	51.95	
	3.0	22.61	77.39	
10	4.0	62.37	37.63	3.42
	3.5	40.06	59.94	
	3.0	19.22	84.03	

What can be seen is that the use of 3.0 as the proficiency criterion score maximally keeps the classification rates similar to what they were in previous years. This approach is a type of equipercentile equating. The scoring system for the FCAT Writing Test, as it is for virtually all holistically scored performance-based writing tests, is less precise than other kinds of testing, such a multiple-choice testing as it is only seven whole number score points and six additional half-score points. It is, of course, however, a direct measure of writing.

Conclusions and Recommendations

In general, in the opinion of the Buros Center for Testing, which has evaluated essay scoring for the Florida Department of Education as performed by Pearson for the past three years, we believe that this partnership is working well, that providing valid scores of writing ability is the number one concern of the process[5], that neither politics nor pressures from the client are involved in any way, and that the work is uniformly professionally performed. In short, it is our opinion that the program meets the standards of best practices in state-wide testing of writing. We acknowledge, too, that many states do not assess writing, as they are not required to do so as they are with English Language Arts, Mathematics, and Science under the No Child Left Behind and Elementary and Secondary Education Acts. Writing is certainly one of the most important educational skills, and by assessing it, Florida is identifying writing as a critical skill to be learned by students in the State of Florida schools.

One must also accept, however, that scoring essays is not as exact a process as scoring some other types of tests. Ultimately, professional evaluators of writing set the scale by identifying essays that they believe embody 1s, 2s, 3s, and so on using the entire score scale. However, the questions to which these responses are written differ year by year. When the essays are selected, if any differences year-to-year occur, then the averages might well be affected. Therefore, to the extent that

[5] Comments to this effect are made by officials of the Florida Department of Education throughout the daily phone calls. These officials obviously want scores to be rendered in a timely way, but they emphasize consistently that the provision of accurate scores is the most important goal.

the rangefinding panel is not exact in assigning example essays to score points, differences will appear. Moreover, during the 2011-2012 academic year, the Florida Department of Education called for a more stringent scoring of the essays. Therefore, one should give substantially less emphasis to year-to-year fluctuations on writing tests such as the FCAT as opposed to more traditional multiple-choice measures that can be equated year to year and when the scoring criteria for the essays have not been changed.

Pearson faced some additional complexity in acquiring enough scorers and having them qualify this year. It is unclear why these issues occurred. Perhaps the job market sufficiently turned around and quality people were generally less interested in this work. Perhaps some of the retirees who have engaged in this work regularly have reached an age where they are less interested or less able to succeed at the work. Such points are only speculation, of course. Nevertheless, it appears that Pearson was able to attract a number of quality scorers to meet their standards of scoring as they have done in previous years. This matter led to some delays in the fulfillment of scoring on schedule. Please note that Pearson responded swiftly when this issue arose and engaged in vigorous efforts to attract and train scorers even after the initial scoring had begun.

We raise a question on the re-training of those potential scorers who fail to achieve satisfactory qualifying criteria at the end of training. Many of these individuals go through training a second time, something we find to be totally appropriate. However, we also believe that many of these individuals see the same

or similar training materials and score the same set of essays on their second qualification attempt. We do understand that the essays that are used to qualify the scorers, known as qualification sets, were generally the same from one training session to the next, although very appropriately, we understand that the trainers re-ordered the sets of essays and re-ordered the responses within sets for these repeat scorers. We believe that while such an approach makes sense operationally, we question whether some of these individuals may remember the scores assigned to certain essays. While re-using such materials for training is certainly not a concern, it may well be for the assessment of scorer accuracy. We encourage Florida and Pearson to consider developing additional sets of criterion essays so that those individuals undergoing training a second time would be examined on essays that differ from those they saw the first time. We understand that there are cost implications to this recommendation, but they are relatively minor compared to other possible concerns that a program such as this one might face. Operationally, this may mean that all scorers-in-training during second or third waves of training would receive this second (or third) qualification set of essays.

Nevertheless, we heartily acknowledge that Florida is effectively assessing writing and commend them for this effort. We know that to a large extent, teachers teach what tests test. Few academic areas are more important than writing. Such tests are expensive, but if a state desires its students to learn and be taught how to write, assessing it indicates its importance. We also recognize that the State has listened to our recommendations in past years, we hope fruitfully. For example,

this year, the State of Florida had every essay read by two scorers; in the past, only 20% of essays were scored twice.

The State of Florida has set and Pearson applied rigorous standards for the scoring quality of these essays in terms of both inter-rater reliability and validity. These standards have been met this year, even in the face of the concerns over recruiting enough qualified scorers. We might suggest a second way of gauging the reliability of scorers. We suggest that in addition to the percentage of exact scorings that is currently carefully analyzed every day during the scoring process and kept for future reference, that the vendors provide the intraclass correlation that is used often to evaluate the reliability of scoring. We would not recommend that this index be used to evaluate the qualifications of scorers. Rather, it has the benefit of telling the extent to which scorers agree generally as opposed to agree exactly.

Writing is an inherently subjective process, both in writing itself and in scoring the products of student writing. We believe that Florida has achieved considerable success in this direct assessment of the writing of virtually all of its students in fourth, eighth, and tenth grades. We believe that Pearson has implemented the writing assessment portion of the FCAT quite effectively and meeting what we understand to be the specifications set by the Florida Department of Education. Florida implemented more rigorous standards for the scoring of student writing this year and it certainly appears that these expanded standards were reflected in the manner in which scoring occurred. In evaluating this assessment process, we have attempted to determine whether the standards of the

testing profession as reflected in the **Standards for Educational and Psychological Testing** (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) have been followed and we believe that they have. This testing program meets the standards of our profession.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Shrout, P.E. & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *2*, 420-428.

Appendix A:

Reports on site visits to observe scorer training in:

Jacksonville, Florida (J. Randall)

Auburn, Washington (R. Spies)

Mesa, Arizona (A. Römhild)

Observations about Scorer Training

Jennifer Randall observed the 4th grade essay scorer training from Monday 3/12/2012 to Wednesday 3/14/2012 in Jacksonville, Florida. She gathered information to evaluate the effectiveness of the scorer training from several sources: (a) observations of scorers (trainees), supervisors, and trainers; (b) the training manual (including anchor papers), (c) training materials (including practice and qualifying essays); and (d) discussions with Rob Sights and Steve Ash. The following is a summary of her observations.

Security

All scorers, supervisors, and trainers were met at the front door and required to check-in with a representative. Identification badges were required to prevent unauthorized entry between the training hours (8:00 am – 4:00 pm). All scorers were also required to sign confidentiality forms before training commenced. Scorers were also reminded not to talk about the project outside of the scoring center to anyone; and to respond to questions about their activities by saying "I am scoring an assessment" with no mention of the state or the name of the test. This requirement was emphasized particularly due to the scoring occurring in the state of the assessment (which is apparently unusual).

Materials

1) The fourth grade essay was the result of a writing test and was scored on a 1-6 score scale basis, where scores must be assigned as whole numbers.

2) Notebooks were provided as part of the training as well as practice in actual scoring by the contractor when training scorers. These notebooks include well-written descriptions of the six ordinally[6] organized rubric scores as well as anchor papers. In addition, the notebooks include descriptions of possible sources of scorer bias, and a description of the writing prompt.

3) 18 anchor papers are provided in the notebooks, three for each rubric point.
    - Of the three anchor papers provided for each rubric point, one of the anchors represents a lower level of performance within that particular

[6] Ordinally simply means that these numbers are rank ordered. From a measurement perspective, we can say neither that the difference between scores 3 and 4 is equivalent to the difference between scores of 5 and 6 nor that a score of 4 is twice as proficient as a score of 2. Such statement would require a substantially higher level of psychometric research.

scale point, one in the middle of the distribution of essays receiving that score, and one at the higher end. For example, for the score of "4," there are three anchor papers, one relatively weak for a scoring of four, one average response for a four, and one high essay.

4) The notebooks that were provided to candidates during the scoring and scoring supervisor training hence provide the basis for all scoring. The rubric is ultimately the basis for this scoring, although in training it is suggested to the scorers to compare student-written essays more to the anchor papers than the rubric per se.

Training Procedures

The training began with a description of the test and the context in which it was given. It proceeded sequentially to the rubric, a review of the bias handout, the above-described anchor papers, several highly structured rounds of practice with feedback, and finally to qualifying rounds.

Scorers were encouraged by Steve Ash, Florida DOE, to give great effort (as all kids deserve their best effort) and to listen closely to the three trainers – Mary, Janice, and Nancy. Moreover, the scorers were asked to listen actively and respect the opinions of others. Finally, scorers were told not to get emotionally attached to a score: "The score will not change no matter what you say; getting upset may affect how you score the remaining of the week and jeopardize your qualification."

Scorers (particularly scorers from previous years) were informed of changes on the scoring rubric. Specifically, they were told of the increased attention to conventions and support (a focus on standard English rules). They were reminded that although the expectations had changed (increased), scoring would remain holistic (not analytic) in nature. To that end, scorers were told that they were *scoring* essays and not *grading* them. As such, they were asked to focus on what was right about the essays as opposed to what was wrong or missing. Indeed, scorers were reminded repeatedly throughout the training process that they were scoring the essays holistically. Scorers were also told that they were expected to eventually score approximately 20 essays per hour, and that most score between 17 and 35 essays per hour. Scorers were encouraged not to score too rapidly (e.g., 100 essays/hour) and to read each essay carefully.

There were five rounds of practice scoring. The first round included a training pack of six papers identified as lower range essays (scores 1 – 3). The second round

included a training pack of six papers identified as upper range (scores 4-6). Both the first and second training rounds were completed on the first day of training. The third practice round included the mid-range practice set of six papers. Practice rounds four and five included essays across the full range of the scoring rubric.

After each practice round, feedback was provided to the scorers; and scorers were urged to ask questions. Scorers were encouraged to ask questions that would help them understand why a particular score point was assigned, and not questions that sought reasons for why another score point was not assigned. At all times, trainers answered questions patiently explaining score assignments. Scorers were also reminded repeatedly by the trainers that the practice sets were selected specifically due to the difficult/possibly controversial issues present in the essays. They were encouraged to remain positive during the practice/training session as they qualifying papers would look more like the anchor sets (i.e. more typical in nature).

Common Scorer Issues/Concerns

- The proportion of the essay actually devoted to writing about the camel[7]
- Are the four parts of the holistic scale (conventions, focus, organization, support) weighed the same?
- Can ideas be a fantasy or made up?
- Not allowing one or two phrases to elevate or detract from an essay
- Use of dialogue must further the story along (no matter how beautifully written, it is just fluff if it does not move the story along)

The trainers responded to all of the scorer issues/questions with considerable detail and patience. Specifically, they noted

- The camel did not have to be the center, or focus, of the story. Students simply had to mention the camel, or camel ride, at some point in the story to indicate that they had, in fact, read the prompt. Scorers were instructed to focus on the conventions, focus, organization and support of each essay as outlined in the rubric and illustrated by the anchor set responses. The trainers explained that the prompts were simply meant to serve as a tool (or jumping off point) to get students writing – as opposed to asking them more broadly to "write a narrative about anything."

[7] Camel was a component of the essay prompt.

- Because student essays were to be holistically scored, the trainers told scorers to consider all four parts (conventions, focus, organization, & support) of each essay together as equally as possible.

- Scorers were told that student essays could be completely fantasy in nature (e.g. flying camel rides, talking camels, camels with polka dots), and that students should not be penalized, in terms of their final scores, for fantastical stories. In fact, some higher score point (4, 5, 6) anchor papers involved fantastical stories.

- Trainers encouraged scorers neither to allow one poorly written sentence to adversely affect a student's score nor to allow one or two beautifully written sentences to improve a student's score significantly. Again, scorers were repeatedly instructed to score each essay holistically taking into account the entire essay from beginning to end.

- Scorers were reminded repeatedly not to become distracted by well written dialogue in the narrative essays if that dialogue did not further the story along. The trainers provided examples of dialogue that failed to move the story along through time, so that scorers would be more likely to recognize such a scenario when scoring student essays.

Scorer Qualifications

- Must average 70% agreement on two of the three qualifying sets
- May have one non-adjacent score across all three qualifying sets
- Agreement must not fall below 60% on any one qualifying set

Comments

Given the observations of the trainers, supervisors, and scorers (trainees) as well as the conversations with Pearson and Florida DOE personnel, it is clear that the training of the 4th grade essay scorers was conducted efficiently, successfully, and appropriately.

Site Visit Report of FCAT 8th Grade Writing Assessment - #1

Robert Spies, Ph.D.

Buros Center for Testing

The Puyallup Fairgrounds was the original site selected for FCAT 8th grade scorer training. At the prearranged time of 8:30 a.m. on Monday, March 12, 2012, site manager Bonnie Bizzell and Assistant Manager Nicole Nelson provided the required badge needed to gain access to the building site. They also supplied a binder of training materials used by candidate scorers. After signing non-disclosure and confidentially forms, direct observations of formal scorer training began from the back of the room where all instruction could be observed.

For the next three days, the training provided to candidate scorers was observed. This training followed the formal processes outlined in the Florida Comprehensive Assessment Test FCAT Writing 2012 Handscoring Specifications. This report summarizes the direct observations, conversations with FDOE and Pearson representatives, and daily telephone conferences conducted during this time period.

Logistics and organization

The Puyallup (Washington) Fairgrounds was selected as the original site for the FCAT candidate training due to limitations in the size of the Auburn PSC facility. The Puyallup site provided sufficient space for the expected 250 candidates. Candidate scorer instruction and qualification was to proceed through Thursday, March 15, 2012 at the Puyallup Fairground. On March 16, instruction was to shift to the Auburn Scoring Center for those individuals who had satisfied previously established scorer qualification standards.

Although the ceilings were high at the Puyallup Fairgrounds, the room had sufficient sound amplification for the Scoring Directors to be easily understood throughout the room. When questions were asked, either a microphone was brought to the candidate or the Scoring Director repeated the question. Instructional materials were provided to candidates in a large ringed binder.

Presentations of instructional materials were projected on a large screen at the front of the room and were also available in paper form. As the materials for practice sets and qualification sets were made available to candidates, supervisors would pick up the materials from a secure room and distribute them to all participants. At the end of the day, all training materials were collected by supervisors and held in a secure location. Overall security was managed with Photo

ID badges, with access points monitored by supervisors and staff of the Auburn Training Center.

Stand-up Training

Robert Heinzman and Susan Blake (Auburn Scoring Directors) were the primary instructors for this initial group of candidate scorers. During this initial training, Helen Devitto (lead Scoring Director) was unavailable due to a family emergency. Her responsibilities were shifted to Kenna Cagle (Pearson Content Specialist) for the period of this training.

Instruction of candidate scorers began quickly from distributed training manuals allowing candidates to read the specific persuasive writing prompt (i.e., the writing situation and the directions for writing) received by each of Florida's 8th grade students in their writing assessment. The lifecycle of the prompt, the general process followed during rangefinding meetings, the importance of accurate scoring, and the basic concepts of holistic scoring were also outlined to candidates. Training directors reviewed the general guidelines and conditions of candidate employment (e.g., attendance, security, confidentiality) along with the procedures that would be followed to qualify each candidate for FCAT scoring.

Candidates received an explanation of six allowable interpretations of essays that might serve as tools for their scoring decisions. Potential scorers were instructed in seven strategies to avoid bias (e.g., considering length of essay response, essay neatness) with their ratings. The holistic method of scoring used by FCAT was explained in greater detail, with both verbal presentations and text illustrations for each of six score points. Three anchor papers were used to illustrate these six score points with papers illustrating one low score point, one medium score point, and one high score point within the overall range of scores. One exception to this process was score point six, which had only two anchor papers. After the rubric was restated for candidates, the overall expectations for FCAT scorers were again reviewed.

After anchor point training, candidate scorers reviewed and scored Practice Set 1 (six essays restricted to score points 1, 2, and 3), Practice Set 2 (six essays restricted to score points 4, 5, and 6), Practice Set 3 (ten essays restricted to score points 3, 4, and 5), Practice Set 4 (fifteen essays of all score ranges), and Practice Set 5 (thirteen essays of all score ranges). The total practice responses (50) were very close to the 52 responses specified in the Handscoring Specifications, and the score point range adequately approximated the required distribution focusing on the middle score

point range of 3, 4, and 5. Each practice set appeared to increase in complexity and in the cognitive load requirements needed for holistic scoring.

As these intellectual demands increased with each practice set, the tension in the room intensified. Candidate scorers who missed the "correct" practice scoring option sometimes requested Mr. Heinzman to readdress the practice set scores and justify the preferred option. During these question and answer sessions, Mr. Heinzman remained both calm and consistent in his responses. He rearticulated the FCAT scorers' goals (accurate scoring), referred to the anchor sets, cautioned candidates to consider the four writing elements, and reviewed when necessary the rationale for the specific scoring.

On Wednesday afternoon (the last day of this observation), Qualification Set #1 was given to candidate scorers. On Thursday, the two remaining qualification sets were administered. All qualified candidates were moved on Friday to the Auburn Scoring Center where instruction was to begin using the ePEN scoring system and the pseudoscoring of essays.

Supervisor and Scorer Qualification Procedures

The specific flow to be followed in this training was specified in the Operational Scoring Quality Management Plan (see Appendix B of the Handscoring Guide). This process was consistently adhered to during the observational period of March 12, 2012 through March 14, 2012. Supervisor qualifications were initiated one week before this observational period. To qualify as a supervisor, at least 75% perfect agreement was needed on two sets with no scores lower than 60% on any sets and with only one non-adjacent score across all qualification rounds. In contrast to the previous year, fewer supervisors qualified than expected (62% or 13 of 21 supervisor candidates) rather than 19 of 21 supervisor candidates anticipated in the Handscoring Guide. To compensate for the lack of supervisors, procedures allowed qualifying candidate scorers to receive field promotions to supervisor in order to meet the required supervisor numbers after the first qualification round was completed.

In contrast to supervisors, FCAT scorers were required on the first two sets to average 70% across the two qualifying sets with no set scores lower than 60% and with 100% adjacent agreement. If the third set was needed to establish eligibility, the acceptance criteria changed to averaging 70% on two of the sets with no sets below 60% and not more than one non-adjacent score between all three sets of essay responses.

It was soon obvious, however, that candidates would not approximate the qualification success rate from last year to become FCAT scorers. In the previous year, 200 of the original 252 scorers (almost 80%) qualified as FCAT scorers. It was anticipated (in the 2012 Handscoring Specifications) that the budgeted numbers for 259 candidate scores would provide for approximately 181 qualified scorers or a success rate of 70%. This projection was a reasonable estimation given the results of the previous year.

By the early stages of candidate training and in consideration of the experience with supervisor qualification rates, preparations had begun to recruit additional waves of candidate scorers. Based on daily conference calls between FDOE and Pearson during this period, concerns were expressed that too few candidates would qualify to score current year FCAT writing essays. FDOE requested frequent updates on the process of recruiting additional rounds of candidate scorers. Similar scorer qualification problems were also noted with 10th grade FCAT scoring process.

When the success rate for the first wave of candidates at the Auburn Scoring Center was compiled, less than 50% of first day recruits (116 out of the original 249) qualified as FCAT scorers. When too few candidates qualified on the basis of their scores, supplemental rules were implemented to allow some of the first wave candidates who were close to qualifying to retrain for the second wave or to spend additional time in the pseudoscoring process.

Comments on FCAT Scorer Candidate Training

Six-point holistic scoring of written essays that considers the integration of four writing elements (focus, organization, support, and conventions) is a complex task involving difficult judgments for individual scorers. Training scorers for this role requires advanced planning, well-developed teaching materials, consistency and persistence. The materials and training observed during the observation period at the Puyallup Fairgrounds met these standards. The Auburn Training Center and the Florida Department of Education closely followed the procedures outlined in the 2012 FCAT *Handscoring Specifications*. Although the first wave of candidate training produced fewer qualified scorers than expected, the integrity of the training process was maintained during the observation period.

Site Visit Report on Scorer Candidate Training for FCAT 10th grade Writing

Anja Römhild

Buros Center for Testing

From March 12 to March 14, 2013, I observed the Scorer Candidate Training for the 2012 FCAT 10th grade Writing Test at the Pearson Scoring Center in Mesa, Arizona. This report summarizes observations and evaluation of the stand-up scorer candidate training and the four online training modules on the Pearson ePEN scoring system. Information in this report is based on observations of the stand-up and online training, conversations with the Florida DOE representative Sally Rhodes and members of the Pearson project team including the Senior Project Manager Wendi Winkie and the Pearson scoring directors Alexandra Atrubin, Milton Eichacker, and Anita Cook, and from examination of the scoring specifications manual and statistical summaries of scorer candidate progress during training.

Observations of stand-up training

Logistics and Organization

For the scoring of the FCAT 10th grade Writing assessment, Pearson planned to recruit 193 scorers. A substantially larger number of scorer candidates had been invited to the training to account for an expected 30% of candidates to drop from the project due to failure to qualify. Of the 319 confirmed recruits 300 were present on the first day of scoring[8]. Over the course of the training, an additional 17 candidates dropped out for various personal reasons. At the end of the stand-up training, 283 scorers participated in the qualifying rounds. A total of 146 candidates met the scoring quality criteria and moved on to pseudoscoring. Due to the shortfall of qualifying scorers, a second and third wave of training was quickly scheduled for the following two weeks.

The training of the entire candidate group took place in a very large room that was equipped with microphones, speakers, and a presentation screen. Candidates were

[8] A local bus strike prevented some candidates from coming to the site.

seated at desks with sufficient room to spread out the training materials. Despite the size of the room, acoustics were good. When needed, candidate questions were repeated through microphone.

For the initial orientation on the morning of the first day of training, a PowerPoint presentation had been prepared which was projected onto a large screen. Visibility of the presentation screen was limited or obstructed for candidates seated in corners and on the sides. These candidates needed to use the presentation hand-out to follow along; the small print of the PowerPoint hand-out may have made that task difficult for some.

The presentation materials were distributed in a binder and included the PowerPoint slides, the scorer quality management plan, an information sheet about scorer bias, the writing prompt with allowable interpretations, an information sheet about the holistic scoring method, the scoring rubric, additional information about the rubric element *support*, the anchor paper set with annotations. Additional packages containing the practice papers and qualifying papers were handed out separately throughout the training. At the end of each training day, all training materials were collected by the scoring supervisors and kept locked at the site.

Site security measures included the wearing of photo-id badges by all scorer candidates upon entry of the building and during training. Scorer candidates also signed a confidentiality agreement at the beginning of training and were instructed not to discuss the project, the assessment, or individual student essays outside of the scoring center.

Stand-up Training

During the morning orientation of the first day, the Pearson site manager and a human resources representative addressed employment and logistical issues; the senior project manager explained the scorer qualification criteria; the lead scoring director discussed the notion of scorer bias and introduced the item prompt, the holistic scoring method, and the six-point scoring rubric. The remainder of the training consisted of a thorough review and discussion of the anchor paper set and of five rounds of practice paper scoring. Review of the anchor paper sets was repeated the mornings of the second and third day of training. At the conclusion of the practice paper review on the third day, scorer candidates participated in two qualifying rounds scoring two sets of 20 qualification papers. Those who did not immediately qualify or disqualify went on to take a third qualification set.

The training and qualification sets were compiled from papers reviewed during rangefinding activities. The anchor set was comprised of 18 papers in total, three per score point representing low, mid, and high responses within each score point. Each anchor paper included annotations which were organized around the four major elements of the scoring rubric: focus, organization, support, and conventions.

The practice paper sets were assembled to target specific score point ranges. The first two sets consisted of six papers representing low score points (1, 2, and 3) and high score points (4, 5, and 6). The third set included 10 papers from the middle section of the score scale (3, 4, and 5). The fourth and fifth sets included 15 papers from the full range of score points and approximated the historical score point frequency distributions.

The review and discussion of the anchor paper and practice paper sets were mostly led by the lead scoring director with occasional assistance from the associate scoring directors. For the initial anchor paper review, the lead scoring director read out loud each anchor paper and its corresponding annotations. Scorer candidates followed along in their own materials and asked questions. As training went on, anchor papers were repeatedly referenced in the discussion of practice papers.

For practice scoring, candidates independently reviewed papers and assigned scores. Supervisors collected the score sheets for their team and returned them with actual scores for each paper and the percent agreement achieved. This information allowed scorer candidates to gauge how well they did in reference to the scoring qualification standards. The subsequent review and discussion of the practice paper sets focused mostly on the more challenging papers which had produced higher rates of disagreement.

Throughout training, papers were discussed in terms of the four focus elements of the scoring rubric and with reference to one or more papers from the anchor set. Scorers were instructed to evaluate each paper as a whole and to consider the four writing elements of the rubric in integration. It was explained that no single error or characteristic should limit a paper to a particular score point. In addition, it was repeatedly emphasized that scorers should try to compare a paper to one from the anchor set.

With respect to the rubric element support, an additional hand-out was provided with the materials which introduced specific labels to describe different levels of quality in the use of supporting details (bare, extended, layered, elaborated). These labels were frequently used by the scoring directors to discuss and explain a

particular paper's score, and provided useful common language with which scorer candidates could discuss papers and ask questions.

Occasionally, scorer candidates questioned the accuracy of a particular paper's score. Scoring directors usually responded by asking candidates to refrain from questioning a particular score and to ask instead why a particular score point was given. Scorer candidates were also asked not to invoke hypothetical papers in discussing an actual paper or score point.

Some scorer candidates seemed to have difficulty with the holistic nature of the scoring task and asked for more precise decision rules or criteria for assigning a particular score or asked how the individual rubric elements should be weighted. Other questions were requests to clarify or "quantify" certain attributions given as rationale for an assigned score point. Examples of these attributions are: "ample support", "tight control", "uneven development", or "good language". For the most part, questions were handled with great patience, empathy, and, most importantly, provided the needed clarification.

Scorer qualification standards

The following information was gathered from the *Handscoring Specifications*. Scorer candidates qualify outright by reaching an average of 70% perfect agreement across two qualifying sets with a minimum of 60% perfect agreement per set. In addition, scorer candidates may not have non-adjacent score disagreements. Candidates who did not qualify with the first two qualifying sets take a third set. The same qualifying criteria apply with the exception that scorer candidates may have one non-adjacent score disagreement across the three sets.

After the initial qualifying rounds, an insufficient number of scorers were able to qualify. A decision was made to offer scorer candidates meeting probationary qualification standards to stay on the project with additional training provided to them.

Probationary scorers are monitored more rigorously by supervisors and spend an additional day for a total of three in pseudo-scoring. Scorers who qualified outright spend two days in pseudoscoring.

Online training modules

Before scorers begin the pseudo-scoring on Pearson's ePEN system, they are required to complete four online training modules intended to familiarize scorers with the functionality of the ePEN scoring system, and to provide additional background information on Pearson and the work as a scorer. Scorers complete the modules in sequence. A training flow menu keeps scorers apprised of their progress.

The first module *Scoring for Pearson* introduces the concept of scoring, gives an overview of the tools and support personnel available to scorers, and explains professional expectations and rules set for Pearson employees.

The next two modules *Pearson Scoring System Part 1* and *Part 2* introduce the ePEN system and its various tools and navigation features. Before completing these modules, scorers are required to agree to the ePEN terms and conditions of use. The modules explains how to use the different tools on the screen including how to access a reference library with anchor papers; how to work in the student response window with its zoom, drag, and fit-to-screen functions; how to view daily and cumulative scorer performance statistics; how to submit an essay score using the graphical scoring grid; and how to access the scoring rubric, annotations for a particular paper, and the writing prompt. The modules also give instructions on test security, on how to send essay responses for review, and how to use the validity review feature. The validity review allows scorers to view annotations of a validity paper, whose score was missed by the scorer.

The last module *Before You Score FCAT* provides specific information and instructions on what actions to take regarding irregular essay responses including troubled child alerts, testing irregularities, condition codes, poor image quality, content or scoring questions, and blank responses.

The online training modules were easy to follow and navigate. In some instances, they addressed important information that had not been covered as fully in the stand-up training. The *Pearson scoring system* modules which introduce the tools and functionality of the ePEN scoring interface were well designed providing opportunities to try out and become familiar with the many tools and features.

Comments on Scorer Candidate Training

Given that training and calibrating individuals to score essay responses accurately on a holistic scoring rubric is an inherently difficult task, the outcome of the

observed training appeared to have been a success. The issues encountered by the scorer candidates seem to be germane to the nature of the scoring task rather than the quality of the training. In my estimation, the use of extensive practice scoring with subsequent review and discussion of student responses is well-suited to accomplish the training task.

Appendix B:

Reports on site visits to observe operational scoring activities in:

Jacksonville, Florida (J. Randall)

Auburn, Washington (R. Spies)

Mesa, Arizona (A. Römhild)

REPORT OF OBSERVATIONS

Dr. Jennifer Randall observed the 4th grade essay scoring from Monday 4/9/2012 through Wednesday 4/11/2012 in Jacksonville, Florida. She gathered information to evaluate the effectiveness of the scoring from interviews with Robert Owen (site aide), Dana Stimpert (site manager), the trainers/score directors (Mary McIntyre, Nancy Shafter, and Janice Tarleton), as well as supervisors and scorers, observations of scorers and supervisors during scoring, and the scoring specifications manual. The following is a summary of her observations.

Logistic Implementation

Facilities

Each day began with a morning brief in one large room at 8:00 am. At this briefing, scorers were provided with general information/ announcements and a review of the anchor sets. Actual scoring occurred in two large rooms (the morning briefing room and an additional large room behind it); each room with the capacity to hold approximately 140 scorers each (142 & 144). In each room scorers were sitting in rows at large tables in assigned seats all facing one direction. The team supervisor for the 10 to 12 scorers faced the opposite direction of the scorers. This setup allowed scorers the opportunity to get the supervisors' attention easily by simply raising their hands. Both rooms were quiet and well lit. A smaller conference room without laptop computers was used for calibration review each afternoon and/or supervisor meetings.

Each large scoring room also included easy access to one of two restrooms for male and female scorers. Two break rooms were also on site. The larger break room included three large refrigerators, six rectangular tables, a microwave, vending machines, and a coffee maker. The second, attached, break room included five additional tables. In addition, to improve morale and overall comfort, each scoring team was provided with a box of Kleenex and a bowl of hard candy refilled as necessary. Scorers were encourage to take brief walks, get a cup of coffee, or stretch as needed to prevent fatigue.

Security

All scorers and visitors are met at the front entrance by Robert Owen, the Site Aide. All authorized visitors received a temporary pass until a picture ID could be created

(approximately 1 hour). All scorers were required to wear their security tags at all times, and were not permitted to enter the building without their tags.

Materials

Each scorer worked on a Dell laptop with a sufficient screen size (approximately 15 inches) to enlarge the essay when necessary. The essays' electronic image 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 repeatedly during the scoring process and had written extensive notes, highlighted text, and annotations on the anchor essays. The scorers appeared to be concentrating on scoring the essays at all times.

Procedures/Time Frame

In addition, to prevent fatigue there was a 15-minute break in the morning and afternoon session, as well as a lunch break for 30 minutes in the middle of the day. Breaks were staggered by room to prevent overcrowding in the break room, an excessive number of cars leaving the parking simultaneously, etc.

Scorer/Supervisor Recruitment

Training

Scorer training occurred in three waves. The first, and largest, wave began Monday March 12, 2012 through Thursday March 15, 2012. Out of the original 224 trainees, 142 scorers qualified during this first wave. The second wave immediately followed pseudoscoring on Monday, March 19, 2012 and lasted three days. This wave began with 17 trainees and resulted in 12 additional qualified scorers. The second wave included brand new trainees as well as trainees who were unable to complete the first wave of training due to illness or missed days. The third wave, which began a week after the second wave and lasted three days, started with 42 trainees and resulted in 29 additional qualified scorers. As with the first wave of training, all trainees in the third wave were new (no repeats from the first or second wave of training).

Pseudoscoring began Friday March 16, 2012. Each qualified scorer was required to complete 12 hours (6 hours/day) of pseudoscoring. Moreover, provisionally qualified scorers were required to complete 18 hours of pseudoscoring. During this time, IRR

and validity statistics were tracked for each scorer. If scorers performed poorly based on this evidence, they were let go during pseudoscoring. This procedure allowed the trainers/scoring directors the opportunity to identify and remediate scorers experiencing problems accurately scoring student papers before live scoring began and to assure the process that the scorers were qualified to succeed in their work.

Scorer Experience/Eligibility Criteria

All scorers were minimally required to hold a verifiable bachelor's degree before they were hired. Scorers on the FCAT project ranged in experience from first-time scorers to the highly experienced (since 2001 when then scoring center opened in Jacksonville, Florida).

Quality Control

Reliability

Reliability was measured using an interrater reliability index (IRR). Because all papers were scored by two raters, the estimate was computed based on the exact agreement between the two scores. The standard previously set by the Florida Department of Education for the 4th grade essay FCAT was 60%.

Validity

Validity was measured using the percent of exact agreement scoring "true score" essays. A "true score" essay has been pre-scored by scoring directors and approved by Florida Department of education representatives before it is used for validity evidence. The validity criterion was based on 70% of true score essays with exact matching score. Each scorer received one validity paper for every seven essays scored.

Backreading

Supervisors are required contractually to back read at least 5% of the total student essays. In an effort to improve the quality of essay scores, supervisors focused on back reading the essays of any scorers thought to be struggling (with respect to IRR or validity). Supervisors who fell below a daily validity rating of 70% could not provide supervisory duties, including backreading.

Removal of Scorers

The following describes the Score Exception Policy used to insure quality scoring of all papers. All scorers are required to maintain a validity score of 70% or more. When an individual's validity falls below 70%, they receive a warning message alerting them to this problem. Supervisors are also notified when a scorer receives this warning. At that point, scorers are encouraged to speak with a supervisor to receive additional assistance and/or review their anchor papers. A check point system is in place that reviews the warned scorers after an additional ten essays have been scored. If their validity continues to fall below 60%, they receive a final warning indicating that they are about to receive a set of ten Target Calibration Essays. Scorers must score seven out of ten of these papers accurately in order to continue scoring (i.e. if they fail to do so they will be locked out of the scoring queue). Again, scorers are encouraged to speak to a supervisor immediately about remediation if they receive the "Target Calibration" warning. For quality assurance, all of the papers for any disqualified scorer are reset and rescored. Moreover, score directors examine the quality statistics for any scorers who voluntarily resign from scoring to insure that these scorers also met the agreed upon quality guidelines.

On Day 1 (Monday), three scorers were removed due to poor quality statistics and three scorers voluntarily withdrew from scoring (citing frustration with the scoring process and inability to holistically score). On Day 2 (Tuesday) one scorer voluntarily removed himself. On Day 3, one scorer was removed due to low quality and two scorers dropped because they found other jobs.

Observations of Ongoing Scorer Training

Daily Calibration

Morning Group Training

Each day began with a large group morning meeting/calibration. Scorers were encouraged to ask questions, and the score director emphasized, or addressed, any issues from the previous day's scoring. On the first day (Monday) of the three day audit, an extensive anchor review was conducted by Mary McIntyre, score director/trainer. This review included a discussion of essays from every scale point (both low and high ranges). Day 2 (Tuesday) began with a more focused anchor

review focusing on the mid-range three, four, and five scale points. Day 3 (Wednesday) began with a review of the 3/4 and 5/6 lines.

Afternoon Group Training

No large group calibrations were conducted in the afternoons. Instead, small group sessions were held with scorers deemed to need extra training. The focus of the small group sessions was determined by the score directors based on the quality statistics for that day.

Calibration Essays

All scorers received a set of calibration essays twice daily – morning and afternoon. These calibration sets ranged between one and three essays. Scorers were notified (via their computer screens) that they needed to complete a calibration set. Once they scored the calibration set, they were provided with the correct scores along with an annotated explanation for the scale score point. In addition to the whole room calibration sets, targeted calibration sets were also distributed to particular scorers when the score directors notice specific problems with their scores. For example, on Day 2 of the audit a targeted set of scale score three essays were sent to 27 scorers. All scorers had access to the calibration sets throughout the scoring process (in addition to the anchor sets provided during training).

Supervision

At the time of the three-day audit, 15 supervisors, 3 scoring directors, a site aid, and site manager were present on site on three days. Each supervisor worked with 10-12 scorers. Supervisors were tasked with answering scorer questions, back reading essays, providing scorers with remediation (e.g. paired scoring), and tracking scorer quality statistics. Scoring directors conducted the large group morning meetings as well as the small group meetings in the afternoon. They were also available to answer scorer and supervisor questions. In addition, scoring directors tracked quality statistics and identified/distributed calibration sets based on these statistics.

Retraining

Small Group

Scorers who were performing unsatisfactorily given the previous information regarding monitoring a scorer's performance (e.g. IRR and validity rates) were retrained in a small group setting. These small group sessions can, and have,

included between 8 and 30 scorers. On Day 1 (Monday) I observed one small group (*n*=17) training session from 1:45 pm to 2:15 pm. This training session focused on working with scorers with low IRR scorers and/or high numbers of non-adjacent scores. Participants reviewed three essays with a 3/5 score point split. The trainer/score director, Nancy, answered scorer questions and explained the issues (possible components that might be confusing to scorers) with each essay. The Day 2 (Tuesday) small group (*n*=19) focused on the 3/4 line. Scorers reviewed one high three scale point paper and one low four scale point paper. The score director pointed out the differences in the two papers (e.g. use of precise language, controlled/focused/complete story line, lack of repetition). Scorer issues that were raised during the Tuesday small group included: (a) terrible conventions, but with a complete story, (b) horrible handwriting, (c) the ideal length of a paper, (d) good stories with no punctuation, (d) failure to complete the story – likely due to running out of time – despite having a good beginning and middle. The Day 3 (Wednesday) focused on the 4/5 split. Twenty-two scorers reviewed two papers. Again, Nancy explained the components of each essay that differentiate the 4 (lack of depth, more general, but with nice language) from the 5 (more details, more satisfying story). Scorers issues that were raised included (a) the 5/6 line, (b) the level of involvement of the camel ride, (c) students stopping in the middle of a sentence, and (d) guidelines on expository – must be the same and include a story line. Each day Nancy, the scoring director, addressed all scorer issues patiently and in great detail. For example, she explained that student handwriting was not to be considered in any way when scoring essays. Furthermore, she encouraged scorers to seek help/assistance from a supervisor if they found the essay difficult to read due to poor handwriting. She also explained, repeatedly, that there is no ideal length for each essay, and that essays should be scored based on the holistic scoring rubric only, not on the number of pages. For example, a five-page essay with no mention of a camel – even if written beautifully with no punctuation errors – should not receive a high score because all essays must at least mention the camel (although the focus of the essay did not have to be the camel).

Individual

Scorers identified by a supervisor or score director as having difficulty scoring accurately received individual trainings if necessary. Specifically, these scorers would pair score (scoring each essay together) with a supervisor until the supervisor and scorer were satisfied with the latter's performance. On the third day of the audit (Wednesday), supervisors were asked to begin the morning pair scoring with the scorers with the lowest IRR values

Site Visit Report of FCAT 8th Grade Writing Assessment - #2

Robert Spies, Ph.D.

Buros Center for Testing

The following report on FCAT scoring at the Auburn Performance Scoring Center (PSC) was based on direct observation from March 27 through March 29, 2012, daily telephone conference calls, and interviews from March 12 through May 4. Among those individuals being interviewed were Nicole Nelson (Assistant Site Manager), Helen Devitto, Robert Heinzman, Susan Blake (Scoring Directors), and Kenna Cagle (Pearson Content Specialist). In addition, Rick Brennenan (Pearson Content Specialists) and Jeremy Bleil (Pearson Scoring Manager) were on-site during three of the days of direct observation. This information was subsequently confirmed (where possible) in the Handscoring Specifications manual, conference calls records, personal contacts, and system reports. However, not all reports have been fully compiled. For additional information, see the first report on the Auburn PSC during March 12 through March 14, 2012.

Observations of the live scoring process

Time frame

Scorer training originally began on March 12, 2012 with four days of training and qualification rounds. The 8th grade FCAT scoring project was completed on May 2, 2012 when the last essay was officially scored.

Based on the 2012 FCAT *Handscoring Specifications*, an estimated 198,995 essays were to be scored by a total of 181 qualified scorers and 19 supervisors. The completion date for FCAT scoring was originally estimated at April 18, 2012 but was later changed to April 20, 2012.

Based on numbers compiled from a variety of sources (including Robert Heinzman in a telephone communication on May 4), the 8th grade FCAT scoring project had approximately 312 scorers successfully complete their qualification rounds, although we note that there were never more than 227 scorers on the roster of scorers on any given day of scoring. A maximum total of 24 supervisors were present during scoring with 100% of essays receiving two scorer ratings. Because of problems with the qualification and retention of scorers, four training waves were

initiated by the Auburn Scoring Center. A total of approximately 565[9] first-day candidates received scorer training for an overall 63% scorer qualification rate.

Facilities

First wave training: The Puyallup Fair & Events Center was designated as the site for the first wave of candidate scorer training due to the facility size. A total of 249 candidate scorers attended the first day of training. All candidate scorers faced toward the front of the room during their instruction. After the initial training period ended on March 15, 2012, all future training transitioned to the Auburn PSC.

Second, third, and fourth wave training: At the Auburn PSC, a maximum of 200 candidates could be accommodated in a large rectangular area. The second and fourth wave of candidates had group sizes that could easily be accommodated in the Auburn PSC. However, with the large group of 168 first day scorers in the third wave, approximately 60% of the candidate scorers faced the middle part of the room in rows of 10-12 candidates, and the other 40% sat faced the opposite direction. The scoring director was orientated to the sides of candidate scorers. The advantage of this spacing arrangement was that most scoring candidates were within a relative close proximity to the scoring director's presentation. The primary disadvantage was that the scoring director did not directly face candidates. The suitability of the Auburn PSC was somewhat marginal in the third wave given the number of candidates. At both of the training facilities, the volume level of the sound system adequately carried the voice of scoring directors so that they easily could be heard regardless of candidate's location.

Scoring areas: When candidates transitioned to become scorers, they moved to a large area of the Auburn PSC with 15" computer monitors connected to the ePEN computerized scoring system. All 8th grade essays appeared on these monitors and could be enlarged or reduced in size where necessary to accommodate the needs of the scorers. During the time of observation, all scorers were in one large room in row configurations and with supervisors in close proximity. Supervisors were typically responsible for between 8 and 15 scorers during the observation period. When the third wave of scorers was to transition from the training room to the scoring room at the Auburn PSC, an additional area previously

[9] This number counts those candidates who repeated training as two candidates.

hidden by movable wall panels was set up to accommodate the increase in scorer numbers.

Pace, breaks, & overtime

During the two observed training sessions, the pace of training scorers could be described as proceeding at a moderate to rapid pace, depending on the questions asked and overall comfort levels of each candidate group. When candidates raised questions, the scoring directors provided additional clarification and would often widen the context of the training. Little lag time existed for scorers to become complacent.

During scorer candidate training, breaks were variable and depended on the progress of the group. However, candidates received at least one mid-morning and one mid-afternoon break of 15 minutes with an additional 30 minutes for lunch. Short individual breaks were allowed when necessary.

After qualifying to become scorers, breaks were observed on a more systematic and predictable schedule. However, as unexpected situations developed, the schedule could become flexible and scorers might break at unpredictable times. On two days of the 2nd observation period, Pearson's computer software system (ePEN) experienced problems that caused computers to shut down and the Auburn PSC to alter its normal schedule. Unplanned training was swiftly initiated but scorers ultimately had to be sent home early.

The typical day began at 8:00 a.m. and ended at 4:30 p.m. Monday through Friday. Due to the large number of scorers at the Auburn PSC, individuals were assigned specific groups that took breaks at different times to avoid too many scorers accessing the Auburn PSC facilities at any one time.

After qualifying as a scorer, the Auburn PSC allowed individuals to work an additional hour in the morning (7:00 a.m. to 8:00 a.m.) and an additional hour or two (sometimes more) in the late afternoon and evening if accurate statistics were maintained. On most weekends during FCAT scoring, both scorers and supervisors were also requested to work overtime on Saturdays and Sundays.

Security

During both of the scheduled visits, security procedures specified in the Handscoring Specifications were consistently maintained. Badges with photo identification were provided candidates and collected after an individual left the project. At both the off-site training (Puyallup Fairgrounds) and at the Auburn

Performance Scoring Center, badges were required to be displayed and were consistently checked when entering the worksite to ensure that unauthorized individuals not gain access to either the off-site location or the Auburn PSC. Phones and other media were not allowed in the scoring or training areas and were physically stored in a specific location away from the scorers' work area. During all of the days of my observation, Ginny Kortesoja (receptionist), Susan Cochran (site tech), Nicole Nelson (Assistant Site Manager) and two other temporary staff members consistently monitored the front desk to limit access to only authorized employees.

All training materials were stored in a secure room prior to the daily instructional sessions. Candidates were able to use these materials during their practice and qualification sets. However, these materials (primarily anchor, calibration, and scoring papers) were subsequently collected at the end of each day. When candidates transitioned to become qualified scorers, materials stayed on their desk and were secured by limiting access through locked doors.

Supervisor/Scorer Recruitment

Supervisor Recruitment: Supervisors from previous projects were contacted and requested to attend training to become supervisors for the 2012 FCAT Writing scoring. At the Auburn site**,** only 13 of 21 applicants (62%) qualified as supervisors. When additional need existed on site, scorers with exemplary accuracy (i.e., determined by their reliability and validity numbers) would be promoted as supervisors. Approximately half of the supervisors used in this project were promoted in this manner.

Scorer Recruitment: The Human Resources (HR) department in the Auburn PSC contacted previous scorers who met the specific guidelines for work on this project. In addition, HR advertised for new scorers via public media, job sites, and social media. From the complete list of potential scorers, only those meeting specific qualifications were invited to participate. All candidate scorers at a minimum were required to hold a Bachelor's degree and be legally permitted to work in the United States. Verification of candidate qualifications was made through the National Student Clearinghouse. In the third wave of scorer candidate recruitment, two temporary work agencies (i.e., Advantage, AppleOne) were also used to add potential candidates. Of 168 third wave candidates, somewhat over 50% were reported by the Auburn HR Department to have been recruited by these agencies.

Operational scoring process

The 8th grade FCAT writing assessment consisted this year of a prompt that described both a precise writing situation and specific directions for writing a persuasive essay. Students were given 45 minutes to prepare their written response.

FCAT essays written by 4th, 8th, and 10th grade students in Florida were scanned and turned into electronic documents. After essays were processed, each essay was grouped and assigned to the appropriate Pearson PSC. Each PSC used the ePEN scoring system that automatically placed essays into a queue and controlled the order of their presentation.

Two scorers evaluated each essay and assigned a score of one to six. Essays were to be evaluated in terms of the integration of four writing elements: organization, focus, conventions, and support. When the two scores were identical or adjacent, the score average was used. When scores were not adjacent, the essay was referred to the resolution process that involved another round of evaluation. When the third score matched either one of the previous scores or was adjacent to one of the scores, the final score was the average of the matching or adjacent scores. In very unusual cases where the third resolution score was not adjacent to either of the two previous scores, additional processing with a fourth score was required.

Child in danger alerts: Danger alerts were to be reported by scorers to their supervisors upon viewing the writings of troubled children. These responses were then forwarded to the Florida DOE for investigation.

Testing irregularity escalation: All observed irregularities in testing were forwarded to the Florida DOE for investigation.

Quality Management

Reliability: Reliability in test measurement refers to the stability of a score or scores. In the case of FCAT scoring, the IRR (inter-rater reliability) is defined as the exact agreement between the first and second score of the two independent raters who evaluated the same test response. All essays were rated at least twice during the FCAT scoring process. The cumulative standard used by the FDOE was 60% inter-rater agreement for the 8th grade writing assessment.

Validity: Validity in test measurement describes the process of gathering evidence to support the interpretation of test scores. In the case of the FCAT, validity evidence was built using the process of designating a "true score" for the

essays written by Florida students. To determine the "true score" of an essay, experts with specialized training from the FDOE and from Pearson rated specific essays. The percent of scorers' exact agreement with those validity scores defines the validity agreement. The cumulative standard used by the FDOE was over 70% for the 8th grade writing assessment.

Scorer qualification and retention: After scorers passed their qualification sets, they entered into the "pseudoscoring" phase for a period of two days (expanded to three days in specific circumstances) when they were provided extra practice time. If scorers accuracy fell below minimum standards, they were administered a 10-paper calibration set of essays. Warnings were issued via ePEN with a scorer exception when quality indicators fell below the required standards. When candidates failed to maintain cumulative validity expectations during pseudoscoring and did not pass their targeted calibration set, they were released from the project without entering the live scoring process. Candidates passing the pseudoscoring process to become active scorers were subject to continuing validity checks. Pearson's ePEN software computed statistics based on cumulative scores for every 10 validity essays as part of the "scorer exception process" that continued throughout FCAT scoring. Scorers were expected to maintain at least 70% exact agreement and 90% adjacent agreement. When scorers fell below 60% agreement and/or 90% adjacent agreement, an alert was issued. Scorers were encouraged to speak to their supervisor about improving their level of performance when they received this alert, and supervisors often tried to intervene prior to that point. Subsequently, when scorers fell below the 60% agreement level, they received a targeted ten-paper calibration set to determine their continuing eligibility to score FCAT essays. When individuals no longer qualified to score the FCAT, their previously scored responses were rescored by other scorers in an effort to maintain quality control.

Observations of ongoing scorer training

Daily calibrations

Each day typically began with an anchor review followed by a targeted review designed for scorers to distinguish between specific score points. Frequently, this instruction was more focused on the middle range of scores that had been problematic for the entire group. Scorers used the morning large group instructional time to ask questions or discuss their concerns about some of the fine points of scoring. Scoring directors would provide organizational updates, logistical

clarifications, and group statistics on the Auburn PSC. To inject some variation to the normal routine, scoring directors requested volunteers recruited from available scorers to read essays to the group. In an attempt to improve (or at least maintain morale), lotteries were held of food items or gift coupons to break up the repetition of scoring.

At the discretion of the scoring directors, specific score points would also be discussed in the afternoon along with an occasional elaboration on different writing elements (e.g., conventions, organization) that were subject to potential misunderstanding. Consistently in most afternoons, small group instruction would be held for scorers identified by supervisors or by scoring directors as needing additional training in specific scoring areas.

Daily calibrations were subject to a fair amount of flexibility depending on the perceived needs of the scorers. During one of the observation days (March 28), the ePEN system crashed in the early morning hours and was completely unavailable over the course of a full day. The scoring directors responded with a complete anchor review, followed by a focused review illustrating the low end and high end of a specific score point. Because scoring could not proceed without ePEN, scorers were dismissed around noon when it became obvious the system would not be available for the rest of the day. The third wave training class continued at the Auburn PSC without any discernable impact.

Supervision

In conjunction with the ePEN scorer exception process providing supervisors and scoring directors with IRR and validity data, one of the primary supervisor roles was to read at least 5% of scorer essays. Supervisors had a shared responsibility with scoring directors for the effective remediation of scorers when IRR and validity scores reached borderline areas and scorers were in danger of missing their performance goals. The Auburn PSC eventually hired 24 individuals to act as supervisors.

During the second observation period, supervisors primarily worked either to identify scorers with lower validity and lower IRR statistics or worked directly with scorers to improve their overall performance. Supervisors would approach individual scorers with notes, would sometimes speak privately with scorers, or use electronic notification procedures to inform scorers of their concerns. In addition to monitoring the statistics of scorers, supervisors also were required to maintain their own scoring statistics (i.e., 70% exact agreement and 95% adjacent agreement across 11 validity papers). If supervisors failed to maintain their scoring statistics,

they were no longer allowed to continue in their supervisory role. Scoring directors automatically monitored the work of supervisors through ePEN.

Retraining and scorer disqualification

The training and retention of scorers after their initial qualification round existed at three different levels. On a more global level, anchor reviews and calibrations (delivered online) were initiated on a daily basis in large groups by scoring directors and content specialists. On an individual level, supervisors would intervene with scorers when the accuracy of a scorer's ratings began to deteriorate based on ePEN statistics and the backreading of scorers' essays. Finally, at a small group level for individuals at risk of being discharged, scoring directors would highlight essays in a specific range of score points. During small groups sessions, scoring directors could also clarify the integration of the four writing elements (focus, organization, support, and conventions) in greater detail. Identification of participants for small group sessions was typically made by supervisors based on their backreading agreement or by scoring directors based on their ePEN statistics (IRR, validity).

COMMENTS

The task of scoring the 2012 FCAT 8th grade persuasive writing essay proved a difficult undertaking as described below. Scorers applied the identical six-point holistic method that was the same as the previous year with increased expectations for the writing elements of support and conventions. Rather than 20% second reads for FCAT essays designated in the previous year, in 2012 100% of FCAT essays were scored twice. This FDOE policy provided an important assurance to individual students and all stakeholders that the assigned writing scores were trustworthy and consistent.

Lower supervisor and scorer qualification rates compared to the previous year were immediately noted at the Auburn PSC. Supervisors had been expected to qualify at a 90% rate (19 of 21) that was even higher than the 81% qualification rate (17 if 21) observed in 2011. However, only 62% (13 of 22) of supervisors qualified during the training period completed one week prior to the first wave of candidate training. Instead of 70% of candidates projected to succeed as scorers (181 of 259 candidates), only 55% of first day candidates (including those repeating training) across four waves of training (312 of 565) eventually qualified. Compared to last year, higher numbers of qualified scorers also dropped out or did not meet their quality metrics for validity and targeted calibration testing, resulting in their disqualification from

FCAT scoring. Four waves of scorer training were ultimately required in order to qualify an adequate numbers of eligible scorers. The delay in the completion of 8th grade FCAT scoring (finishing May 2, 2012 rather than April 20, 2012) was related to the gradual acquisition of enough qualified scorers.

To their credit, the Florida Department of Education and Pearson acted quickly and employed innovative strategies to add potential candidates to successive waves of training classes. To add to training numbers, some first wave candidates who initially failed to qualify as FCAT scorers were allowed to retrain in the second training wave, and pseudoscoring was extended for a third day (from two days) for some scorers. Temporary work agencies were also solicited to bolster candidate numbers. These strategies without question added successful candidates to the ranks of scorers. However, qualifying reviewers and keeping them qualified throughout the time needed to score all 8th grade FCAT continued to prove a very difficult task.

The delay experienced with the completion of the 2012 8th grade FCAT scoring was likely not the result of training, materials, experience, or the level of FDOE and Pearson support. With the 2011 8th grade FCAT scoring, these two parties produced a timely result with substantially the same personnel. Based on exchanges that took place during regular telephone conferences over the course of 8th grade scoring, it was evident that FDOE and Pearson took prompt and appropriate actions to rectify the number of qualified scorers needed. It was also unlikely that the increase in second reads for essays compared to last year (100% compared to 20%) was a substantive factor in this delay. Both 4th grade and 10th grade also required 100% second reads and did not experience significant delays in their project completion date.

Instead, the reason for this scoring completion delay is more likely due to a combination of three interrelated factors. First, higher expectations for the 2012 FCAT writing essay were obvious compared to the same exam last year. Secondly, when compared to the 2011 8th grade expository essay, the 2012 8th grade persuasive essay required students to demonstrate considerable organizational and conceptual skills. Statistics were not available to compare previous FCAT results, but the complexity of persuasive writing for students has been consistently documented in the research literature.

Finally, the 2011 8th grade writing prompt (visiting a favorite place) was much more concrete for students compared to the 2012 writing prompt of making a specific recommendation to a principal regarding school policy. When viewed together,

these three factors were considered likely to produce a decline in student FCAT scores and account for the majority of the difficulty experienced during 8th grade FCAT scoring in terms of the qualification and retention of scorers.

Site Visit Report on Operational Scoring of FCAT 10th grade Writing

Anja Römhild

Buros Center for Testing

From April 4th to April 6th, 2012 I visited the Pearson scoring center in Mesa, AZ to observe and evaluate on-going operational scoring activities for the 2012 FCAT 10th grade Writing Assessment. This report provides a summary of my observations and impressions. The information gathered for this report is based on conversations with various Pearson project team members including the lead scoring director, Alex Atrubin; the associate scoring directors, Anita Cook and Milton Eichacker; one scoring supervisor, Joseph Townsend, two Pearson HR representatives, the site manager, Betsy Newville; and Pearson content specialist, Mel Jurgens. In addition, information was gathered from the Handscoring Specifications document and from observations of the activities of the scoring directors, supervisors, and scorers.

1. Observations of life scoring process

1.1. Logistic implementation

Time frame.

The operational scoring window for the FCAT 10th grade Writing Assessment is approximately 4½ weeks starting March 16 and ending April 18. During this window, an expected 202,683 essays were scored. Pearson planned for 193 scorers and 20 supervisors to accomplish this task. The initial training wave did not qualify the expected number of scorers prompting two additional training waves from March 17-21 and March 26-29. With the additional training waves, the project eventually recruited 249 scorers and 21 supervisors. The initial shortfall of qualifying scorers caused a lag in scoring output at the beginning of the operational scoring window. To make up ground, Pearson offered overtime to scorers and supervisors during the first three weeks of scoring. By the end of the third week and by the end of this site visit, the Mesa site was back on schedule having completed more than 60% of essays.

Facilities.

The Pearson scoring center in Mesa is set up to accommodate multiple scoring projects simultaneously. At the time of this site visit, one other scoring project was ongoing. It was apparent that the activities from this project did not interfere with FCAT scoring operations. For the FCAT Writing project, the center utilized one very large room and two smaller sectioned-off areas in adjacent rooms to seat scorers and supervisors. Scorers work on computer desks with 15-inch monitors and adequate room for notebooks and materials. The desks are arranged in rows with between 6 and 15 seats per row. Supervisors usually sit at the end of a row, next or near to their scoring team. Though scoring directors have a separate office, two of them usually work on computers in the main room and within easy reach for supervisors and scorers. Scoring directors and supervisors were frequently sought out by scorers and supervisors to ask questions and discuss papers. Noise level in all rooms was low and a calm work environment was maintained throughout.

Pace, breaks, overtime.
Scorers were expected to work from 8 am to 4:30 pm on weekdays with two 15-minute breaks around 10 am and 2:15 pm and one 30-minute lunch break. In addition, scorers were allowed to take short individual breaks as needed. To monitor productivity, scorers' time logged into the ePEN system was monitored by supervisors and scoring directors. Scorers with consistently low productivity (percent time logged into ePEN) and/or consistently low rates of scored papers (10 or fewer per hour) were approached by their supervisors or scoring directors to improve their scoring output. The payment structure for scorers also provided some incentive to strive for high scoring productivity. In addition to a minimum hourly rate for time spent scoring and time spent in training, scorers also earn an amount for every paper scored. A subset of scorers was hired through an agency; these scorers only received an hourly wage without additional performance-based payments.

The site offered overtime to scorers and supervisors during the first three weeks of operational scoring. At the peak of activities, overtime was offered for one hour in the morning from 7 am to 8 am, in the evening from 4:30 pm to maximally 8pm, and on weekends. As the project progressed and scoring output began to improve, overtime was reduced and limited to supervisors and scorers meeting higher scoring quality standards.

Security.
All staff and visitors are required to wear a badge inside the building. Scorers are given photo-id badges which they return when leaving the project. Visitors wear

non-photo-id badges and sign in and out of a visitor log book for the duration of their visit. The facility has one main entrance with a reception area where badges are checked upon entering. All scoring-related materials used by scorers and supervisors (i.e., notebooks with anchor and calibration papers) remain in the building. Scorers and supervisors in the main scoring room leave materials on their desks at the end of each work day. The room is locked overnight. Those scorers and supervisors working in adjacent rooms with access from other areas of the building lock their materials in a secure box which is stored in the locked main room at the end of the day.

1.2. Scorer and supervisor recruitment

Pearson recruited a portion of the scorer candidates and all supervisor candidates for the FCAT 10th grade Writing Assessment from its own pool of experienced scorers and supervisors, some of which were said to have participated in previous scoring projects of Writing assessments. Additional scorer candidates were recruited through online announcements, radio and tv ads, and through referrals from previous Pearson scorers. All scorer and supervisor candidates were required to have at least a Bachelor's degree and needed to have completed college level writing-based course work. Professionals from Florida's Test Development Center (TDC) approved the education qualifications of each individual scorer.

For the first wave of scorer training, 130 invited trainees had previous scoring experience. The remaining 170 were new hires. Only 146 of the 300 candidates had qualified to score for this project, falling short of the 193 scorers needed. Due to the lower than expected qualification rates from the first wave of training, two additional trainings were conducted. The second wave of training included 48 repeat trainees from the first wave, who had narrowly missed the qualification criteria. These repeat trainees were identified by the scoring directors and TDC staff. For the third wave of training, Pearson employed the services of an agency to recruit 95 additional candidates, who had no prior scoring experience, and invited 32 additional candidates from its own pool of previous scorers.

1.3. Operational scoring process

The FCAT 10th Grade Writing assessment consisted of a single writing prompt to which Florida 10th graders composed a handwritten response on a maximum of two pages within a 45-minute time limit. The responses were scored on a six-point holistic rubric which describes levels of achievement in terms of four writing

elements: focus, organization, support, and conventions. Scorers were instructed to evaluate the overall quality of the response and to consider the integration of the four writing elements. The scoring criteria were modified for the 2012 FCAT 10th grade Writing to include expanded expectations on the correct use of writing conventions and use of supporting detail.

Starting in 2012, all essays were independently scored by two readers. When the two assigned scores are identical or adjacent, the final score is computed as the average of those scores. With this decision rule, it is possible for students to receive a half-point score as their final essay score. For example, an essay given a score of 4 by one reader and a score of 5 by the second reader would be assigned a final essay score of 4.5. When an essay received two non-adjacent scores, then the essay was automatically assigned to resolution scoring. A supervisor or a scorer with high scoring accuracy performance would read the essay and assign a resolution score. The final essay score was then computed as average of the resolution score and the identical or adjacent original score or scores. Should the resolution score be non-adjacent to either of the original scores, then the essay was escalated to adjudication scoring where a final score would be determined and approved by a FDOE representative.

Aside from rubric-based score points, essay responses may be assigned a condition code marking a response as blank; as illegible, incomprehensible and/or insufficient content; as off-topic; or as a response in a foreign language. All essays were scanned for potential alerts concerning a child in danger or testing irregularity. Essays suspected of such content were brought to the immediate attention of the scoring director who issues an alert report to the FDOE.

Central to the scoring operations of the FCAT 10th Grade Writing was Pearson's scoring system ePEN which facilitates the viewing and scoring of the essays and the delivery of various training and calibration tools to scorers via web interface. Various tools in ePEN allow scorers to zoom, drag, or fit-to-screen the essay they are reading; to access a reference library with anchor and calibration papers; and to access the scoring rubric, the writing prompt, annotations available for some validity and calibration papers, a glossary, and a help system. Scorers typed scores into a graphically displayed score scale or select condition codes. Messaging tools allowed scorers to send a paper for review. Overall, the navigation design of the ePEN system appears to be straightforward. Given the number of navigation tools and windows available on the screen of the 15-inch-sized monitors, space to view an essay was somewhat restricted but the viewing experience is enhanced by tools such as the zoom and drag functions.

1.4. Quality monitoring and scorer disqualification

The scoring quality of the FCAT 10th Grade Writing was managed through test papers called validity papers, monitoring of scorer performance metrics, remediation and training measures targeted at scorers whose scoring performance falls below scoring quality standards, and through daily calibration efforts focused on the entire pool of scorers.

Scorer performance was evaluated through measures of score agreement with validity papers and measures of inter-rater reliability. Validity papers were embedded in a scorer's queue of essays where approximately one in seven essays read by the scorer is a validity paper. Validity agreement was computed as percent perfect agreement with the paper's pre-assigned and FDOE approved validity score. During operational scoring, scorers needed to maintain a daily minimum validity agreement of 70%. Scorers who fell below the standard are monitored and might be given one or more remediation actions. Scorers whose cumulative validity performance falls below 60% perfect agreement or 90% perfect and adjacent score agreement receive an initial warning along with targeted training measures. A final warning was issued if the scorer's validity performance does not meet the 60% perfect agreement and 90% perfect and adjacent score agreement after an additional 10 validity papers. The scorer would then be administered a targeted 10-paper calibration set which he or she must pass with 70% perfect agreement and 100% perfect and adjacent score agreement or be released from the project. All essays scored by scorers who were released from the project for scoring quality reasons were reset and rescored.

While validity agreement was the primary metric by which scorers are evaluated, supervisors also monitored scorers' inter-rater reliability which is computed as percent agreement between an essay's first and second score. Intervention needs for individual scorers are determined based on the IRR that is percent perfect and adjacent score agreement, i.e. a scorer's tendency to assign non-adjacent scores.

Scorer performance was monitored primarily via the above-mentioned scoring quality metrics and through the backreading of scored essays by supervisors. Supervisors are expected to backread approximately 5% of a scorer's work. While

backreading does not override operational scores, it allows supervisors to identify problematic patterns or misunderstandings by individual scorers.

2. Observations of ongoing scorer training

2.1. Daily calibrations

Each morning, scoring directors conducted a focused anchor review that targeted specific score points or score point ranges. The morning reviews were also frequently used to address logistical issues such as overtime and to update scorers on the progress of the project with information on project-wide daily and cumulative inter-rater reliability, validity agreement, and completion rate. While on site, a full group calibration followed the anchor review every morning. For example, on the first morning of the site visit, scorers were given a paper exemplifying the low end of score point 3. After scorers read and scored the paper independently, the associate scoring director Anita Cook reviewed the paper for the entire group. Scorers used the opportunity to ask questions.

Additional calibrations were administered throughout the day, many of them online vie ePEN. Unlike validity papers, scorers were notified when they are reading a calibration paper. The papers are annotated to provide instant feedback. Online calibrations may be directed at the entire group or to specific scorers in need of additional training. Scorers who joined the project after live scoring began, for example those who qualified in the third wave of training, were required to complete all online calibrations that had been administered up to that point. In addition to online calibrations, scorers also receive instant feedback from annotated validity papers they may have missed.

2.2. Supervision

At the time of the site visit, the project employed 21 scorer supervisors for a total of 216 scorers. From a cursory look across the room, it appeared that supervisors worked with teams of 8 to 12 scorers. Their primary responsibilities were to monitor the daily and cumulative scoring quality metrics of individual scorers and to backread a portion of the essays scored by their team. In addition, supervisors were expected to score essays for at least 2 hours a day, assist scoring directors with identifying suitable validity papers, and serve as point person for those scorers on

their team who have questions concerning specific papers. All scorers are subject to backreading, though frequently supervisors target specific scorers with performance problems. Backreading couild not produce an override of original scores; however, it can trigger a resetting of scores that are non-adjacent to the backreading score. Based on the monitoring of scoring quality metrics and backreading, supervisors identify needs for additional training, which they may address individually or coordinate with the scoring directors. While on site, I observed supervisors frequently engaged in one-on-one conversations with scorers. Supervisors keep a scorer intervention log documenting a scorer's performance issues and the training efforts provided.

The scoring performance and backreading activity of each supervisor was recorded within ePEN, and can therefore be monitored by the scoring directors, who are responsible for backreading the work of supervisors.

2.3. Retraining

On one occasion, a series of anomalous papers required additional calibration of the entire group. At the start of the project, a series of essays with predominantly narrative structures required specific instruction on how to apply the scoring rubric to these essays. In addition to ongoing calibration activities with the entire scorer group, scoring directors address scoring performance issues through targeted small and large-group trainings as well as through targeted online calibrations. Scorers might be selected for additional training based on supervisor recommendation or are automatically selected on the basis of performance metrics (e.g., falling below the 70% validity agreement rate) missed calibration sets or validity papers.

Training and remediation efforts could also include individual feedback from supervisors or scoring directors. For example, supervisors might directly observe individual scorers during essay scoring and provide immediate feedback. Some training opportunities were built into the ePEN system. During the site visit, targeted online calibration sets were also given to small groups of scorers with specific remediation needs. These small-group trainings were conducted every day by either the lead scoring director Alex Atrubin or the associate scoring director Anita Cook. These trainings focused on contrasting specific score points.