BUROS
Center for Testing

Report on the Scoring of the FCAT Writing Assessment

Kurt F. Geisinger, Ph.D. and Stephen G. Sireci, Ph.D.

Buros Center for Testing

Consultants to the Florida Department of Education

May, 2011

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. and Stephen G. Sireci, Ph.D.

Buros Center for Testing

Consultants to the Florida Department of Education

May 2011

As part of the Buros Center for Testing's review of various aspects of the Florida Comprehensive Assessment Tests (FCAT) for the Florida Department of Education, three visits were made to FCAT Writing scoring sites. These site visits are summarized in Appendix A of this report. In addition, our staff and subcontractors participated in the daily calls of State Department of Education officials and the leaders of the contractor. This report summarizes our impressions of the quality of the scoring of FCAT Writing responses.

Responses to a writing prompt are written by virtually all Florida students in fourth, eighth and tenth grades as part of the FCAT. These responses were scored in Jacksonville, FL; Auburn, WA; and Brooklyn Center, MN[1], respectively. The staff of the Buros Center and one of its subcontractors visited each site for approximately a day and a half. Members of our team also listened to and participated in daily calls that transpired during the scoring process, approximately from early March until the end of the first week of April.

[1] Reports on each of these three visits are appended to this longer, more comprehensive report.

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, on-going monitoring of the scoring process and a few brief statements about the central tendency of the student performance in each grade.

Scoring Timeline and Time Frame. The time frame for the FCAT Writing scoring was documented in various contracts and notes. The work began in the scoring centers in early March 2011 and was completed April 11, 2011. This time frame was ambitious, but doable when one is working with an experienced team of professional scorers of written examination essays. Likely due to the weak economy currently being experienced by our country, Pearson was able to hire an adequate number of excellent essay scorers, many of whom are experienced. In a stronger economy, fewer excellent scorers might be available, in which case, the timeline might prove more difficult to meet. Pearson, in conjunction with Florida Department of Education professionals, were judicious in permitting the most accurate scorers to work overtime and on weekends to meet the demanding schedule.

Scorer Training. Although we were not able to observe training per se this year due to the timing of our contract's acceptance, we all observed retraining and some aspects of continuous training. Continuous training appears to be the Pearson modus operandi for scorers of writing. Initial training, whether one has previous

Pearson scoring experience or not[2], occurs over a two- to three-day period for both scoring supervisors and scorers, with supervisors trained first. After this training, supervisors and scorers must pass a test where high standards in scoring accuracy are demanded. These tests involve having the supervisors and scorers grade a carefully selected sample of essays that were previously graded by a panel of experienced, expert scorers. The grades assigned by the potential supervisors and scorers are compared to the grades provided by the expert panel . Supervisors must take three qualifying sets of essays and have an exact agreement of 75%, with none of the three sets below 60% exact agreement. The scorers, as opposed to the supervisors, need only reach an average of 70% across two samples of essays, which we believe is a rigorous 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 twenty 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 a specific type of 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. If a scorer's statistics did not meet specific criteria of scoring accuracy, he or she had to attend a retraining session. If their scoring accuracy did not improve, they were

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

removed from the pool and the scores they provided up to that point were removed as well.

The success of the scoring is ultimately completely dependent upon the quality of the scorers. Using the scoring at Brooklyn Center, MN as an example, and quoting from that site visit report, "I will begin with a bit of history describing the scoring. The training of supervisors began on March 7, 2011. Sixteen supervisors began training and 15 passed training. Fifteen supervisors continue working at the site as of our site visit; however, one left and was replaced by an exceptional scorer. On March 14, 2011, 220 scorers began training and 139 passed the training. On April 6, 129 continue(d) scoring essays." Relative to Auburn, WA, over 252 began training and approximately 200 met the standard. At Jacksonville, 268 began training and 169 were found to be qualified. At Brooklyn Center, 220 began training and 139 finished training. Clearly, Pearson is able to select an impressive group of scorers to begin the process.

Supervisors. As noted previously, supervisors met high standards to serve in this role. They are expected to work with the scorers, especially those having some difficulties. Each supervisor oversaw approximately 10-12 scorers, which we believe is a reasonable supervisory ratio.

The Work of the Scorers/Standards of Reliability and Validity. Approximately 1 in every 7 essays that a scorer reads is a validity paper rather than an essay for operational scoring. Validity papers are essays that have been

prescored by expert scorers and are embedded into the operational scoring of responses to check that scorers continue to grade in accordance with the scoring rubric. Scorers are blind to whether any paper is operational or not. Some of these essays are simply re-scorings (or second scorings) of student papers; 20% of all essays are randomly selected for rescoring so that the reliability of scoring can be assessed and concerns about some scorers identified and considered. The data for these essays are used to estimate the reliability of scoring. The index that is evaluated by Pearson is called the IRR or inter-rater reliability ratio. The standard for an acceptable IRR is 60% exact agreement. We might suggest that while it would be somewhat less interpretable, a kappa index might be a slightly preferred index, perhaps in addition to the existing IRR. The kappa adjusts the index for chance agreement. IRR values for 2011, as taken from data provided to Buros by Pearson, were 62%, 57%, and 54% for fourth, eighth and tenth grades, respectively. For the two upper grades, these values failed to meet the standards that Pearson set for itself. Percentages of scores that were scored to be non-adjacent values were 2%, 3% and 4%, respectively. Of course, while reliability—the consistency of scorers—is an important consideration in the evaluation of student writing, validity—how well the scores on a writing assessment identifies real writing ability—is the key aspect of any scoring process.

Other essays that are part of the 1 out of 7 scored as validity checks are essay papers that were taken from the pre-testing of the essay prompts and range finding. These latter papers are those that were scored in agreement by experienced scoring

directors and/or Florida educators at the rangefinding/rangefinding review meetings. Data from these essays are used to estimate the validity of essay scoring. The expected criterion for the percentage of agreement with the intended score is 70%. Again, we encourage Florida and Pearson to consider using kappa instead of, or in addition to, percentage agreement. Validity percentages were 76%, 78%, and 79%, for grades four, eight and ten, respectively. These values are quite good and are somewhat in contrast to the IRR values above and are well above the standards of performance that have been set.

Supervisors also "back-read" between 5%-10% of the essays scored by those that they supervise. This process is intended both to check the accuracy of scorers and to help provide guidance to the scorers if there is a reason to provide such instruction. We believe the back-readings, validity checks, and IRR statistics, provide an effective monitoring system and help facilitate accurate reading of FCAT essays.

The scorers are evaluated into tiers. The best scorers in terms of their validity and reliability evaluations are considered to be Tier 1 scorers and are kept on near the end of the process when fewer scorers are needed. These scorers are also those who are offered overtime when additional work is deemed necessary. When scorers are determined not to provide reliable and valid assessments of essays, they are re-trained. If their scoring performance does not improve, they are ultimately dismissed. When a scorer is dismissed due to poor scoring performance,

the scores they provided are eliminated or "reset" and then rescored by more able scorers.

While the scorers primary work is scoring the essays according to the rubrics, they also scan the essays for evidence that the student writing the essay is experiencing any sort of serious emotional difficulties (possible depression and suicide, sexual abuse and the like) and bring such essays to the attention to the scoring leaders. Similarly, if a student's essay is suspected of some sort of problem—typically intellectual dishonesty--because of two types of handwriting, suspected plagiarism, or some other similar behavior—scorers are also to bring such essays to the attention of their leaders. In either of these cases, such cases are brought to the attention of officials at the Florida DOE, who contact Florida school districts to act on the concern appropriately.

Average values. All essays are scored on a scale of 1-6. All three grade means (arithmetic averages) are slightly higher than those the previous year, although we caution that that year-to-year comparisons of essay scores should be considered with great caution, as is described below. The mean of the fourth grade essays was 4.03 (as opposed to 4.00 last year). The modal score was 4, a value assigned to 54.5% of the essays. The mean of the eighth grade essays was 4.17, relative to a 4.09 value earned last year. Forty-four percent of the eighth graders received a score of 4, again the modal value. The average value for the tenth grade students

was 3.999 (certainly rounded to 4.00) as opposed to a 3.90 value in 2010 and 45.55% of the students taking the test earned the modal score of 4.

Conclusions

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 two years, we believe that this partnership is working well, that providing valid scores of writing ability is the number one concern of the process[3], that neither politics nor pressures from the client are involved in any way, and that the work is uniformly professionally performed. Using experienced scorers is a big advantage for the State of Florida, we believe.

One must also accept, however, that scoring essays is not as exact a process as some other types of testing. 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 slight differences year-to-year occur, then the averages might well be affected. Therefore, to the extent that the rangefinding panel is not exact in assigning example essays to score points, differences across years will appear. Therefore, one should give somewhat 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.

[3] 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 score is the most important goal.

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. There are few academic areas 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.

In closing, we would like to acknowledge the willingness of Pearson to open their operations to the Buros group and its subcontractors. We believe that their openness is a strong sign of their professionalism and the quality of work they appear to be providing the State of Florida on these tests. We commend them as well as the State of Florida and its Department of Education for developing and scoring what appears to be a fine measure that should help the State assess the actual writing skills of its students.

Appendix A:

Reports on Scoring Site Visits to:

Jacksonville, Florida (C. Wells)

Auburn, Washington (R. Spies)

Brooklyn Center, Minnesota (K. Geisinger)

Notes from Observations of FCAT 4th-grade Essay Scoring

Craig S. Wells, Ph.D.
Sireci Psychometric Services, Inc.
Consultants to the Buros Center for Testing

Craig Wells observed the 4th-grade essay scoring on 4/4/11 and 4/5/11 in Jacksonville, Florida. He gathered information to evaluate the effectiveness of the scoring from primarily three sources: an interview with Rob Sights, observations of scorers and supervisors during scoring, and the scoring specifications manual. The following is a summary of his observations.

Information Obtained via Interview with Rob Sights of Pearson

The following information was gathered from an interview with the Director of the scoring site, Rob Sights (details were confirmed using the scoring manual).

Timeline. The scoring operation began on March 7th when the supervisors were trained over a four-day period starting on 3/7 and ending 3/10. The scorers were trained and selected from 3/14 to 3/17. The operational scoring began on 3/18 and was projected to be finished on 4/8. However, as of 4/4, the scoring was expected to be completed a day early (on 4/7).

Scorer training. The training started with 268 potential scorers. After the first phase of training, 169 participants were classified as qualified, of which, 140 were used for the operational scoring. Over the four-day period, the scorers were trained on an anchor set of papers and 5 practice sets of essays. All of the papers contained annotated notes to aid training and calibration. The anchor set was comprised of 18 essay responses, 3 at each score point (low to high for each score point). The practice sets contained 52 total essay responses that covered all score points. Once the training phase was completed, the scorers were assessed to determine if they qualified for operational scoring. To pass the training and become classified as a qualified scorer, the scorers were required to score two qualifying sets of 20 papers each. The sets contained essay responses for all score points (1 to 6). The scorers must have exhibited at least 70% validity agreement on average, at least 60% for each set of papers and 100% agreement within one adjacent score point. If they did not meet these criteria, then they rated a third set of 20 papers. In this case, to qualify, the scorer must have achieved at least 70% agreement on two of the sets, with no percentage agreement less than 60%. Furthermore, a scorer could not have any scores beyond one non-adjacent score point for all qualifying sets. Several of the qualified scorers had scored for many years and were professionals with a writing background.

Supervisor training. There were 17 supervisors, 15 of whom were responsible for about 10 to 12 scorers each. The other 2 supervisors were "floaters" that backup

quality control. The supervisors were judged based on the reliability and validity data of their scoring team. To qualify as a supervisor, s/he must have exhibited at least 75% agreement on two of the three qualifying sets of papers (the same papers used to qualify scorers), no agreement percentage less than 60% on any of the sets, and no more than one non-adjacent score point.

Standards for reliability and validity. Reliability was measured using inter-rater reliability (IRR) based on the percent of rescores with exact agreement; 20% of the essays were rescored (however, the student receives the first score only, except for essays that are rescored or backread by supervisors). The standard of acceptable IRR was 60% exact agreement. The site was averaging about 61% IRR through 4/5.

Validity was measured using the percent of exact agreement scoring "true score papers." A "true score" paper has a rating that was assigned by Florida educators and goes through a thorough vetting process by the customer and Pearson before it is used for validity evidence. The validity criterion was based on 70% of "true score" papers with exact matching score. A scorer rated one validity paper for every seven essays scored. They were averaging about 75% agreement through 4/5.

Monitoring scorers' performance. Multiple sources of data were used to monitor each scorer's performance daily: IRR and especially validity rates, scoring speed (too slow or too fast may be an indication of a possible issue), and backreading conducted by supervisors. For the backreading, about 5% to 10% of the papers were rescored by the supervisors for all scorers, with special attention placed on struggling scorers (e.g., those with low validity rates). Scorers were monitored using cumulative and daily data. Daily quality was monitored because it is possible that a scorer has adequate IRR and validity rates overall, but is not performing well on a specific day. In such cases, the scorer was retrained or sent home for the day.

Daily calibration and feedback. All scorers were given two essays in the morning and afternoon session that were intended to help them calibrate their scoring. The scorers were informed that the essays were for calibration purposes and immediate feedback was provided after scoring. In addition to daily online calibrations, the scorers received statistics describing their daily and cumulative performance. The statistics that were reported were agreement rates in IRR, validity and supervisor scores, as well as frequency distribution of score points for themselves and for the team. The number of papers scored was also reported.

Retraining scorers. 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 group setting. I observed two 30-minute group re-training sessions in which one or two supervisors reviewed two exemplar papers. In each re-training session, it appeared that the scorers were selected because they were making certain types of errors. For instance, in one session, the scorers were

giving too many 4's to validity papers that should have received a 5. The supervisor selected papers that addressed this issue and discussed in detail why the selected papers should receive the pre-evaluated score. Common misconceptions that may have led a scorer astray were discussed (e.g., the trainers emphasized not comparing papers to each other, but relative to the criteria and anchor item set). The scorers were encouraged to ask questions, which some of them did. At the end of each session, it appeared that the scorers had a better understanding of what a particular score meant.

Scorers were sent home if their performance was sufficiently poor. In such cases, the supervisor may review and override any score. About 14 scorers had been sent home for at least one day from the beginning of the scoring through 4/4 and their scores have been reset (i.e., rescored by another scorer or by a supervisor).

Scoring Observations

The following notes are based on Dr. Wells' observations of the scorers and supervisors during scoring. The scorers were split into two separate rooms (however, whenever general instructions were provided, they were gathered into one room so that everyone heard the same information). They were sitting in rows at large tables, all facing one direction. A team of scorers sat together in the same area as their supervisor. They were assigned seats such that they were placed in alphabetical order (using this seating arrangement allowed them to have more control over who sat next to each other). The team supervisor for the 10 to 12 scorers faced the opposite direction as the scorers. This setup allowed scorers the opportunity to contact the supervisor easily by raising her/his hand. The room was quiet and well lit. Each scorer was working on a Dell Latitude E5500 with a sufficient screen size to enlarge the essay when necessary. The essays' electronic image was clearly displayed on the screen. Anchor essays were provided to each scorer in a three-ring binder. Many of the scorers appeared to be using the anchor item sets as an aid for scoring (in fact, I noticed that many of the scorers had written extensive notes, highlighted text, and had annotations on the anchor essays). The scorers appeared to be concentrating on scoring the essays. 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 around noon.

While the scorers were scoring the essays, the supervisors were busy either re-scoring the scorers' essays to check for possible issues that may need to be addressed in retraining or answering questions raised by the scorers. In a one-hour period, I observed 24 questions raised by scorers, some of which came from the same few scorers. In most cases, the supervisor appeared to provide a satisfactory answer. In a few cases, the supervisor requested the help of the Assistant Director.

Conclusion

Given the observations of the scorers, supervisors, effective retraining of error prone scorers, and acceptable statistics such as the IRR and validity agreement rates, it appeared that the scoring of 4th-grade essays was being conducted successfully and appropriately.

Observations from FCAT 8th Grade Essay Scoring

Auburn, WA

Robert Spies, Ph.D.

Buros Center for Testing

When I arrived at 7:45 a.m. the Auburn Performance Scoring Center on April 4, 2011, I was met by the acting site manager, Niki Nelson and provided with a badge for identification and site security. She reviewed the six-point scoring rubric and the holistic method of scoring essays used with the FCAT. During that time period, scorers began arriving to take their place at individual workstations. A summary of the observations and conversations over the course of these two days has been condensed into following categories.

The following information on the scoring process used at the Auburn Performance Scoring Center was obtained from direct observation and from interviews with Niki Nelson and Scoring Director/Assistant Scoring Directors Helen Devitto, Susan Blake, and Robert Heinzman.

Scorer Recruitment: The Human Resources (HR) department in Auburn contacted previous scorers who met specific guidelines for work on the project. In addition, HR advertised via newspapers, Craig's List, Career Builder, Facebook, and Monster for new scorers. From the complete list of potential scorers, only those with sufficient writing experience were invited to participate. All potential scorers were required to hold a Bachelor's Degree in a related field and be legally able to work in the United States.

Scoring Timeline: The time frame for FCAT scoring was documented in the conference notes and verified in discussions with the above individuals. The majority of the work at the scoring centers began in early March 2011 and was completed at the Auburn training site by April 8, 2011.

Scorer Training: Over 300 scorers were invited to participate in training at the Auburn site with 252 showing up on the first day. Of this number, 111 were rehires having previous scoring experience and 143 were new hires meeting the criteria specified by the Florida Department of Education. Due to the size of the initial qualifying group, an off site location was used for initial scorer training

instead of the Auburn Performance Training Center. At this qualifying stage, a paper process was selected for logistics purposes. Scorers were provided an overview of the process and trained on anchor sets of annotated papers to establish consistent recognition of the six score points used with FCAT scoring. Five practice sets of essays were used to illustrate the scoring criteria across all six score points. When scorer training was completed for these candidates, an assessment phase was begun to determine formal qualification to participate in the FCAT scoring. To qualify, each scorer was given two sets of 20 papers that required an average of at least 70% agreement, no set scores lower than 60%, and no non-adjacent scores on either of the two sets of papers. One hundred twenty-four scores passed both rounds successfully. For those scorer candidates who failed one set but passed the other, a third set of papers was administered and only one non-adjacent score for all three sets was allowed. Approximately 200 of the original 252 first day candidates qualified as scorers. A total of 18 new hires and 20 rehires failed to qualify as scorers of the FCAT with the remaining 14 dropping out of the qualification process. To provide constant and continual feedback, scorers had continual access to their own and the overall group validity and IRR statistics during the entire scoring process.

Supervisor Training: Twenty supervisors were ultimately selected for role of scoring supervisors due to the larger than expected qualification rate (80% vs. 55%) of scorers. Seventeen of these supervisors had past supervisor experience and three were promoted from within the original group of scorer applicants. Supervisors were held to a higher standard than scorers by requiring them to successfully complete all three qualifying sets with an average of 75% exact matches to the criterion score, no set scores lower than 60%, and no more than one non-adjacent score point across all of the three sets. A similar process was used for training supervisors as previously described in the Scorer Training section. Supervisors typically had responsibility for 8-10 scorers.

Standards for Reliability: To evaluate the overall reliability of the FCAT, inter-rater reliability (IRR) estimates were used. This estimate was computed by using 20% of total essays that were scored for a second time for exact agreement with the first score. The standard previously established by the Florida Department of Education for the FCAT was set at 60%. By the afternoon of April 5th, the IRR for the 8th grade FCAT was 57% with additional essays being reloaded of problematic scorers for additional scoring. The exact specifications for this process are currently being documented at the request of the Florida Department of Education. Available data for previous eighth grade years for the FCAT IRR

estimates were between 50% and 76%.

Standards for Validity: To evaluate the overall validity evidence supporting use of the FCAT, expert panels determined essays scores (further evaluated by the Florida Department of Education and Pearson) that served as "true scores" against which to measure current scorer performance. For the current year, a validity agreement criterion of 70% was used based on a review rate of one essay per seven essays scored. At the time of this writing, an agreement rate of 78% was being observed.

Monitoring Scorers' Performance: For a variety of purposes (e.g., maintenance of scorer performance, determination of retraining needs), data were collected on all relevant aspects of individual scorer performance. Details of a scorers' statistics included their validity score (including high and low scores), IRR (including high and low scores), logged in time, total essay numbers read, backreading percentage agreement, and an overall frequency distribution. Approximately 10% of papers were subject to backreading with the highest percentages occurring in the early essay scoring stages. As statistics were amassed on individual scorers, supervisors were able to fine tune backreading to focus on validity rates, IRR rates, and frequency distributions of overall scorer ratings. Where a scorer was demonstrating fluctuation in the accuracy of their scores, supervisors would intervene with an electronic message, personal note, or visit. Scorers were sometimes instructed to take a break or to take the rest of the day off when their performance suddenly deteriorated. If performance continued at an unacceptable rate, scorer retraining was initiated.

Daily calibration and feedback: When scorers reported in the morning at the Auburn Performance Scoring Center, they started anchor review training in an effort to recalibrate their scores to a consistent level. During the two days of observation, these reviews were led by the Scoring Director or the Assistant Scoring Director and focused on distinguishing between three potential score points – 3, 4, or 5. Previous training papers were cited for reference purposes. A clear rationale for the different essay scores was articulated. Earlier calibration training began with illustrations of all six scoring points. By the time of this observation (late in the essay rating cycle) the scoring director believed it was most effective to train on

the most problematic score points.

Retraining scorers: When scorers were identified with statistical reports for consistently missing their accuracy goals, retraining was initiated. The scoring director and an assistant scoring director (Robert) conducted two retraining sessions observed on April 4. Retraining typically occurred in small groups of 10 individuals per session and offered score reviews that illustrated the differences between ratings of 3, 4, and 5. The quantity, quality, and complexity of the sentences and paragraphs supporting specific essays were highlighted for these scorers. Scorers discussed the complicating factors that may have led to other score judgments in the sample essays and engaged the scoring director with possible reasons for selecting alternative scoring decisions. The scoring director articulated additional reasons for choosing the specific essay rating. After the retraining was completed, the scoring director pointed to statistics that demonstrated the success of their retraining methods because no scorers had been subject to a second round of retraining during the course of their scoring the FCAT essays.

Observations of the Scoring Process

During the two-day observation period at the Auburn Performance Scoring Center, access to the scoring floor was carefully controlled. All scorers were issued badges to permit entry. To gain admittance, scorers walked past the main desk and deposited cell phones in a large plastic storage container. The front desk area was continually staffed and was within direct view of the site manager's office to insure additional site security. The layout of the Auburn Performance Center was formed into an L shape within one very large room for scoring the 8th grade FCAT. The room was well lit and ventilated. Each eight-foot table supported two workstations. 110 Dell GX 260 and Dell GX 280 desktop computers and 90 Dell Latitude laptop computers were available to scorers and supervisors at any one time. All computers had 17" monitors. Of these computers, the majority used wireless access, but 80 desktop computers were directly wired to the server. When essay writing was unclear or exceptionally small, scorers could enhance the image. Overall, the level of image clarity for essays was impressive.

In this large office space, supervisors faced in the same direction and were seated behind their scorers and could easily observe for pace and adherence to scoring policies. When essay score points were unclear, scorers would raise their

hands for their supervisor's attention. During the two days of observation conducted toward the last part of the scoring period, most questions had apparently already been addressed, and scorers needed few hourly clarifications. Statistics on performance were passed between scorers and supervisors, and between supervisors and the Scoring and Assistant Scoring Directors. On the first day of this observation (April 4), all available scorers worked for the full day on rating essays. However, at the end of the first day of observation (April 4) the majority of scorers were told that the work project had been completed and were excused. On the next day, only 6 scorers (selected based on high quality indicators) and all 20 supervisors continued to work on scoring the eighth grade FCAT. The same policies and procedures were obvious that day albeit in a much more limited space and with many fewer scorers. The remaining staff will apparently continue work on essays for the rest of the observation week. Essays were being recoded and were being rescored from previously scored sets. That is, Pearson utilizes a very impressive approach to re-scoring essays once it has been determined that a particular scorer is not meeting standards in terms of scoring accuracy. All of the essays scored by those scorers are re-scored by more accurate scorers who meet these high standards of scoring validity.

Conclusion

Ensuring an equitable scoring process for written essays based on a holistic, six-point rating system is a complex series of structured tasks. Within the period of this observation and from all available records documented during this essay scoring process, the Auburn Performance Scoring Center maintained high levels of scoring integrity for the 8th grade FCAT writing essays consistent with the exacting requirements of the Florida Department of Education.

Visit to Brooklyn Center, MN Scoring Site

Kurt F. Geisinger, Ph.D.

Buros Center for Testing

April 5-6, 2011

I will begin with a bit of history describing the scoring. The training of supervisors began on March 7, 2011. Sixteen supervisors began training and 15 passed training. Fifteen supervisors continue working at the site as of this date, but one left and was replaced by an exceptional scorer. On March 14, 2011, 220 scorers began training and 139 passed the training. On April 6, 129 continued scoring essays.

I arrived at about 8:00 am just as the final scorers were also arriving on April 5. Shortly after my arrival, one of the Pearson supervisors ran a review of the rubric, in this case focusing on the low essay for each of the six score points. The instruction was direct, repetitive as appropriate, and focused on both the essays at those score points (e.g., low 1, low 2, and so on) and the writing and organizational characteristics that are involved in those score points. The vast majority of the scorers were attending to the instruction. Many were looking either at their screens to read along or their notebooks that also contain the rubric and the score points. The instruction took just over one-half an hour and was a good refresher for these scorers. It seemed effective. On the 6th, the day similarly began with training on the rubric, going over a number of the critical score points. The instructor (Alex) covered the middle essays from 1-6. Once again, it seemed effective and the vast majority of scorers paid rapt attention.

Each day, the scorers quickly gravitated to their computers to score the essays. The Pearson computer system appeared quite effective. The screens provided a scanning of the handwritten essays quite clearly, and the scorers could access the rubric and the example essays in their reference library easily while scoring each essay. The room in which the scorers worked was reasonably roomy, acceptably well lit, and while scoring, the computers did not seem crowded. Scorers had ample room for their materials. It was a good space for the project. All the scorers appeared to be engaged in their tasks. Wandering among them, one sees them moving through essays somewhat quickly, or occasionally focusing on particular phases. This is the third and likely last week of their scoring. Most of the scorers also had drinks at their sides and occasionally sip some coffee, water or soda, and the supervisors all seemed to have some bowls of candies for their charges. At the beginning of the day, the scorers worked for about 2 hours, including the training before being given a break.

The scorers were not neophytes. I was told that more than half of the scorers, it was thought, scored essays for other clients of Pearson. Indeed, I was informed that some scorers have read essays at this location for more than five years.

To be selected to serve as a scorer, the standard for qualification was 60% perfect agreement and 95% adjacent agreement across three sets of essays—what Buros considers a reasonably high standard of performance, one that ensures accurate scoring overall. It is possible that an individual paper could receive a score that would not receive 100% perfect agreement by another reader, but it is unlikely that the whole state's scores could be off. Of course, the entire process is dependent upon the rangefinding process that is conducted by the Florida DOE. Supervisors, also temporary employees for the most part, must take all three qualifying sets of essays and achieve an overall exact agreement percentage of 75% with no percentage (of the three) below 60%. The scorers themselves needed only reach an average of 70% across two samples, still a very reasonable standard. Approximately 1 out of every 7 ratings is a so-called validity rating, not a scoring of an actual response. If one watches scorers discretely, one can see how seriously they are concentrating on their work. Their eyes are focused on their screens. Their faces are pictures of concentration. Although no observer can speak for every scorer, they obviously take the assignment seriously and the professionalism of the Pearson employees carries through to these temporary employees. During the day I was there, scorers received a low six essay (that is, one that had been evaluated as being a six on the scale and one that was near the bottom of the distribution of sixes). Some 83% of the scorers gave it a score of six. These empirical case studies add credence to the validity of the scoring process.

The scorers were evaluated into tiers. The best scorers in terms of their validity and reliability evaluations are considered to be Tier 1 scorers and are kept on near the end of the process when fewer scorers are needed. When scorers were determined not to provide reliable and valid assessments of essays, they were re-trained. If this process did not work, they were ultimately dismissed. When a scorer was dismissed due to poor scoring performance, the scores they provided were eliminated or "reset" and then rescored by more able scorers.

In addition to grading the essays, the scorers also scanned the essays for evidence that the student writing the essay was experiencing any sort of serious emotional difficulties (e.g., possible depression and suicide, sexual abuse, etc.) and bring such essays to the attention to the scoring leaders. Similarly, if a student's essay is suspected of some sort of problem—typically intellectual dishonesty--because of two types of handwriting, suspected plagiarism, or some other similar behavior), they were also to bring such essays to the attention of their leaders. In either of these cases, such cases were brought to the attention of officials at the Florida DOE.

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 two years. We believe that this partnership is working well, that providing valid scores of writing ability is the number one concern of the process, that neither politics nor pressures from the client are involved in any way, and that the work is professionally performed. Using experienced scorers as is the case in Brooklyn Center, MN is a big advantage for the State of Florida, we believe. One must also accept, however, that scoring essays is not as exact a process as some other types of testing. 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 slight differences year-to-year occurs, then the averages might well be affected. Therefore, one should give somewhat 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. Nevertheless, we heartily acknowledge that Florida is assessing writing. We know that to large extent teachers teach what tests test. There are few things academic more important than writing.