BUROS
Center for Testing

Operational Check of the 2010 FCAT 10th Grade Reading and 11th Grade Science Equating Results

Prepared for the

Florida Department of Education

by:

Tzu-Yun Chin, M.S. (Doctoral Student)

Leslie H. Shaw, M.A. (Doctoral Student)

Andrew C. Dwyer, M.S. (Doctoral Student)

Carina McCormick, M.A. (Doctoral Student)

Kurt F. Geisinger, Ph.D.

May 26, 2010

Questions concerning this report can be addressed to:
Kurt F. Geisinger, Ph.D.
Buros Center for Testing
21 Teachers College Hall
University of Nebraska – Lincoln
Lincoln, NE, 68588-0353
kgeisinger2@unl.edu

Operational Check of the 2010 FCAT 10th Grade Reading and 11th Grade Science Equating Results

Purpose of the report

In order to assist in ensuring the accuracy of the scores from the 2010 FCAT examinations, the Buros Center for Testing was asked by the Florida Department of Education (FDOE) to review the procedures used in their 2010 FCAT item calibration, scaling, and equating procedures and to perform an independent check of their calibration and equating results using data and anchor item parameter estimates provided by the FDOE. This study is similar to one conducted by Buros in 2009, and this report follows a highly similar format. Although we have been asked to examine the equating results of three 2010 tests (3rd grade reading, 10th grade reading, 11th grade science), in this draft report we only address the results from the operational check of the 10th grade reading and 11th grade science tests[1]. This draft report (1) provides an outline of the procedures we used to operationally calibrate the items and link their parameter estimates to the FCAT reporting scale, (2) summarizes our results, and (3) compares them to results obtained by the FDOE.

Operational item calibration and scaling

We carried out an operational check of the item calibration, scaling and equating of the 10th grade reading and 11th grade science FCAT tests. These tests were chosen by the FDOE to be reviewed. Specifically, the Grade 10 Reading assessment was chosen

[1] The report for the 3rd grade reading assessment was delivered on May 13, 2010.

because of the high stakes nature in which the assessment has a state-level graduation cut-point that plays a role in whether or not a student graduates with a standard high school diploma or a Certificate of Completion. The Grade 11 Science had item-type changes that the open-ended items were replaced by multiple-choice and gridded-response items in 2010. This section provides a short description of the procedures we followed to arrive at our equating results.

Step 1: Calibration data files

The FDOE provided us with a calibration data set, but minor data cleaning was still required for each grade. For our analyses, we included only examinees who received a 10th grade reading or 11th grade science anchor form (i.e., forms 37, 38, 39, 40 and forms 17, 18, 19, 20, respectively). Non-standard curriculum students and students who would not receive reported scores were excluded from the both data files as well. Additional filters were applied for selecting the calibration sample for the 10th grade reading assessment. For 10th grade reading, the additional filters were (1) selecting students from the designated calibration schools and (2) excluding students who had missing scores for the performance tasks. We understand that this practice is consistent with procedures followed by the FDOE. The sample sizes of the calibration samples for each grade are reported in Tables 1 and 2.

Table 1: Calibration sample size information for 10th grade reading

	Total N	**Form 37**	**Form 38**	**Form 39**	**Form 40**
Grade 10	14,927	3,732	3,738	3,700	3,757

Table 2: Calibration sample size information for 11th grade science

	Total N	**Form 17**	**Form 18**	**Form 19**	**Form 20**
Grade 11	24,854	6,200	6,187	6,216	6,351

Step 2: Classical item statistics

Classical item statistics were computed to identify any items that might need to be removed from the item calibration step. For our analysis, we computed item p-values and corrected item-total correlations. Items with extreme p-values or low item-total correlations were flagged for further psychometric review using the same flagging criteria described in the calibration and equating specifications document provided by the FDOE. No items were judged to be problematic enough to be removed from item calibration in the 10th grade reading test. However, after reviewing the classical item statistics for the 11th grade science test, field test items from all four anchor forms were removed from the item calibration (i.e., Item 9 from forms 17, 18, 19, and 20). These items were excluded due to their relatively low corrected item-total correlations.

Step 3: Item Calibration

All items (i.e., forms) were calibrated concurrently using PARSCALE-4 (Muraki & Bock, 2003). Three-parameter logistic model (3PL) was applied to the multiple-choice items and the two-parameter partial credit model (2PPC) was applied to the short-response, extended-response, and gridded-response items. All estimations achieved convergence.

Table 3: Description of item calibration set

Grade (Subject)	Description
10 (Reading)	45 core items, 27 anchor items (77 total)
11 (Science)	51 core items, 29 anchor items (80 total)

Step 4: Item ICC’s and Item Fit Statistics

Item fit statistics as well as graphs of the ICC’s (expected vs. observed) were examined to determine if any item should be removed from the item calibration.

PARSCALE's G^2 statistic was computed and employed to assess item fit. The ICC's and residual plots were created using the software program ResidPlots-2 (Liang, T., Han, K. T, & Hambleton, R. K., 2008). Although a few items were identified as having a significant lack of fit by the PARSCALE's G^2 statistic, the G^2 statistic is known to be overly sensitive with large sample sizes[2]. The corresponding ICC plots did not indicate major problems with any of the items, therefore no items were removed from the item calibration or from the subsequent item scaling based on these measures.

Step 5: Year-to-year anchor item performance

A number of methods were used to determine if the anchor items were behaving differently in the current administration than they had in previous administrations. All of the methods used in the 2009 check were again used in 2010, including delta plots of item difficulties, correlations and scatter plots of new and old anchor item parameter estimates, and correlations and scatter plots between shifts in item difficulty and shifts in item position (within the test). The root mean squared deviations (RMSD) of the a- and b-parameters (see Gu, Lall, Monfils, & Jiang, 2010 for a detailed description) were also evaluated.

For the 10th grade reading assessment, shifts in item position were not problematic for the anchor items, and although a few individual items flagged as performing differently across administrations based on some of the indices, none of the individual anchor items were deemed problematic enough to warrant removal from the anchor set.

[2] Each core item, for example, was administered to all students in the calibration sample, while each anchor item was administered to roughly one-fourth of the students in the calibration sample. As a result, a higher proportion of core items were artificially flagged as being misfitting.

This corresponds to FDOE's final anchor set decision that included all planned anchor items but no backup core items in the final equating solution.

For the 11th grade science assessment, shifts in item position were again not problematic for the anchor items. However, the correlation between the old and new *a*-parameters was low for the entire anchor set (r_a=.610). Examining the *a*-parameter correlation by anchor form shows that the value was especially low for form 17 ($r_{a,form\ 17}$=.138). In addition, item 36 on form 18 exhibits relatively large p-value difference between the current calibration sample (*p*=.641) and the previous sample (*p*=.536). This large *p*-value change for item 36 on form 18 was also identified by the delta plot. Therefore, we considered equating the grade 11th science assessment with and without form 17 and/or item 36 on form 18. It should be noted that removing form 17 dramatically reduces the length of the anchor set which may introduce unstable equating coefficients and reduce the degree of the content representativeness. Therefore, the equating solutions associated with removing form 17 may not be preferred unless additional backup core items can be added to the anchor set.

Finally, we also reproduced the equating results with the final anchor set for the 11th grade science assessment selected by FDOE. The advantage of performing this analysis—done after our independent analyses—was that it demonstrates the variance due purely to software differences. Tables 4 and 5 summarize the possible anchor sets we chose to examine in the item parameter scaling stage.

Table 4: Description of anchor item set for 10th grade reading

Anchor item set	Description
#1*	All “planned” anchor items used (n=27)

* FDOE’s final anchor set

Table 5: Description of anchor item set for 11th grade science

Anchor item set	Description
#1	All “planned” anchor items used (n=29)
#2	Item 36 on Form 18 removed (n=28)
#3	Form 17 removed (n=21)
#4	Form 17 and Item 36 on Form 18 removed (n=20)
#5*	Item 47 on Form 20 removed (n=28)

* FDOE’s final anchor set

The Stocking and Lord (1983) equating procedure was implemented for each anchor item set described above. Table 6 contains the equating results for both the 10th grade reading and 11th grade science in more detail.

Step 6: Item parameter scaling

The item parameters were placed on the FCAT reporting scale by the program IRTEQ (Han, K. T., 2007). This program uses Stocking & Lord (1983) methodology to adjust the item parameter estimates obtained in Step 3 so that they are on the FCAT reporting scale. A separate item parameter scaling “run” was performed for each anchor set that was considered (described in Step 5).

Step 7: Examination of results

For each item parameter scaling “run”, the linking coefficients (M1 and M2) from the Stocking & Lord (1983) procedure were compared to the final coefficients computed by the FDOE in their operational scaling and equating. Software differences and slight differences in anchor sets were expected to produce slightly different linking coefficients, so differences between the scale scores that would be obtained using the FDOE linking coefficients and scale scores that would be obtained using our linking coefficients were calculated at important points along the FCAT reporting scale (specifically, the "cutoff" points that separate the FCAT reporting categories). The FCAT cut scores were first transformed back to the theta scale using the M1 and M2 coefficients produced by FDOE. The cut points on the theta scale were then transformed to the FCAT scale using the M1 and M2 obtained with different anchor sets from our investigation. This procedure was followed to provide a means for determining the magnitude/importance of the differences between our linking coefficients and the FDOE linking coefficients. The results are reported in Table 6.

Before discussing the information in Table 6, it is important to note that just as in 2009, we expected slight differences between our results and the values provided by the FDOE due to differences in calibration and equating software. FDOE and its subcontractors used MULTILOG for calibration while we used PARSCALE in this project. It has been demonstrated in the psychometric literature that different IRT software packages usually produce different, though highly similar, parameter estimates (e.g., Childs and Chen, 1999). With a simulation study, DeMars (2002) further concluded that MULTILOG and PARSCALE could both recover item parameters well and the selection between the two software packages for a specific project can be made based on considerations other than estimation accuracy. In addition to calibration software packages, calibration options that may assist parameter estimation (e.g., prior distributions, starting values, and RIDGE estimator) and convergence criteria within a given IRT calibration software package can also introduce differences in parameter estimates. Besides different calibration software, we also used a different software package (i.e., IRTEQ) for implementing the Stocking & Lord procedure in order to obtain the linking coefficients. The difference in the equating programs can also result slightly different linking coefficients. One possible advantage of using different software packages can be the additional evaluation of the stability of the parameter estimates and linking coefficients. Moreover, duplication of analyses with the same software basically only has the potential to identify errors of analysis or judgment. Using different software permits consideration of the two above-named types of errors but also considers differences across standard software packages.

Because evaluating the content representation of the anchor set was outside the scope of this project, we had no reason to remove (or add) items from the anchor set based on item content. We believe, based on observation and conversations we have had with the FDOE, the content reviewing task has been handled appropriately by the FDOE.

Taking software differences into consideration, our item calibration and equating results should be considered consistent with the results obtained by FDOE and its subcontractors, especially for the score range below the fourth cut point (4-5 cut). The last four columns of Table 6 show the magnitude of the differences between scale scores calculated by FDOE and scale scores calculated by Buros at several important points along the FCAT reporting scale. Generally speaking, the Buros-calculated scale scores are always within a few points of the FDOE-calculated scores, especially at points located towards the center and the lower end of the scale. Although the score differences around the fourth cut point is larger than ideal, the consistency of the results in the majority of the scale score range lend support to the 2010 FCAT scores.

Overall impressions of the calibration, scaling, and equating procedures

Despite knowing that software differences would likely lead to slightly different results, carrying out the operational calibration and equating is an important step in the evaluation of the calibration and equating procedures used by the FDOE. The *FCAT 2010 Calibration and Equating Specifications* (2010) document provided to us by the FDOE, which served as a guide when we performed this work, was also reviewed. Overall, our impressions were that the entire process is well-organized, the statistical analyses used to identify problematic items are adequate, the organizations involved in the operational work are all nationally recognized testing firms composed of high quality staff, and the responsibilities of those organizations are clearly defined and set up in a way to ensure the accuracy of the scores on the 2010 FCAT. While equating is a highly quantitative procedure, there is also considerable judgment involved; we believe FDOE has utilized the resources and expertise at their disposal to come to reasonable and justifiable calibration, scaling and equating conclusions.

Table 6: Scaling and equating results for each calibration/anchor set combination

Grade	**Description of Calibration and Scaling items**				**Transformation Coefficients**			**Differences between Scale Scores in the area of achievement level cut points**			
	Calibration items		Anchor items		M1	M2		Cut 1-2	Cut 2-3	Cut 3-4	Cut 4-5
10 (Reading)	**FDOE values**				**50.764**	**319.836**		**259**	**284**	**332**	**394**
	All core items used and 27 planned anchors used in calibration (n=72)		FDOE’s final set: All "planned" anchor items included in anchor set (n=27)		52.39	321.65		258.9	284.7	334.2	398.2
Grade	**Description of Calibration and Scaling items**				**Transformation Coefficients**			**Differences between Scale Scores in the area of achievement level cut points**			
	Calibration items		Anchor items		M1	M2		Cut 1-2	Cut 2-3	Cut 3-4	Cut 4-5
11 (Science)	**FDOE values**				**45.347**	**312.398**		**259**	**284**	**332**	**394**
	All core items used and 29 planned anchors used in calibration (n=80)		All “planned” anchor items used (n=29)		43.85	312.93		261.3	285.5	331.9	391.8
			Item 36 on Form 18 removed (n=28)		44.26	312.42		260.3	284.7	331.6	392.1
			Form 17 removed (n=21)		42.96	312.34		261.8	285.4	330.9	389.6
			Form 17 and Item 36 on Form 18 removed (n=20)		43.39	311.68		260.6	284.5	330.4	389.8
			FDOE’s final set: Item 47 on Form 20 removed (n=28)		44.16	313.50		261.5	285.8	332.6	393.0

References

FCAT 2010 Calibration and Equating Specifications (January, 2010). Final Version.

Childs, R. A., & Chen, W.-H. (1999). Software note: Obtaining comparable item parameter estimates in MULTILOG and PARSCALE for two polytomous IRT models. *Applied Psychological Measurement*, *23*, 371-379.

DeMars, C. E. (2002). Recovery of Graded Response and Partial Credit Parameters in MULTILOG and PARSCALE. Paper presented at the annual meeting of the National Council for Measurement in Education, Chicago, IL. (ERIC Document Reproduction Service No. ED476138)

Gu, L., Lall, V.F., Monfils, L., Jiang, Y. (2010). Evaluating anchor items for outliers in IRT common item equating: A review of the commonly used methods and flagging criteria. Paper presented at the annual meeting of the National council on Measurement in Education (NCME). April 29-May 3, 2010, Denver, CO.

Han, K. T. (2007). IRTEQ [Computer software]. Amherst, MA: Center for Educational Assessment, University of Massachusetts at Amherst.

Liang, T., Han, K. T, & Hambleton, R. K. (2008). ResidPlots-2 [Computer software]. Amherst, MA: Center for Educational Assessment, University of Massachusetts at Amherst.

Muraki, E., & Bock, R. D. (2003). PARSCALE-4: IRT item analysis and test scoring for rating-scale data [Computer software]. Chicago: Scientific Software International.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.