BUROS
Center for Testing

Operational Check of the 2010 FCAT 3rd Grade Reading Equating Results

Prepared for the

Florida Department of Education

by:

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

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

Kurt F. Geisinger, Ph.D.

May 13, 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 3rd Grade Reading 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 to the 2009 report. 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 3rd grade reading test. The results pertaining to the other tests will follow in the coming weeks. 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 3rd grade reading FCAT test. This test was chosen by the FDOE to be reviewed in part because of the high stakes nature of this assessment. More specifically, the Grade 3 Reading assessment has a state-level retention cut-point that a student must pass in order to advance to the next grade. 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 analysis, we included only examinees who received an anchor form (i.e., forms 37, 38, 39, 40). Non-standard curriculum students and students who would not receive reported scores were excluded from the data file as well. 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 Table 1; we believe that these sizes are adequate for equating purposes.

Table 1: Calibration sample size information

	Total N	**Form 17**	**Form 18**	**Form 19**	**Form 20**
Grade 3	12,654	3,175	3,161	3,134	3,184

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. A few individual items were flagged, but after further review of each item, including item-fit statistics, no item was judged to be problematic enough to be removed from item calibration.

Step 3: Item Calibration

All items (i.e., forms) were calibrated concurrently using PARSCALE-4 (Muraki & Bock, 2003) using the 3PL model. All items achieved convergence.

Table 2: Description of item calibration set

Grade	Description
3	45 core items, 27 anchor items (77 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 by 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, PARSCALE's G^2 statistic is known to be overly sensitive with large sample sizes[1], and the 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

[1] 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.

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 added.

Shifts in item position were not problematic for the anchor items in any of the anchor forms, and although a few individual items flagged as performing differently across administrations based on some of the measures, none of the individual anchor items were deemed problematic enough to warrant removal from the anchor set. The correlation between the old and new a-parameters for form 40 was lower ($r_{a,form}$ 40 = 0.506) than for other forms ($r_{a,form}$ 37 = 0.850, $r_{a,form}$ 38 = 0.879, $r_{a,form}$ 39 = 0.836). This low correlation seems to be an artifact of the range restriction of the new and old a-parameters for form 40, nonetheless, the equating results were investigated both with and without the form 40 anchor items in the anchor set.

When the form 40 anchors were removed, the overall anchor set was shortened (and perhaps the content representation suffered slightly), so we also investigated the results of adding in one of the sets of “backup” anchor items (i.e., core items that could be used as anchors). After examining items statistics for each of the backup anchor item set, we felt that the first 10 core items (from the “ZUM03” passage code) formed the best replacement anchor set, but we acknowledge that this decision was based solely on item performance statistics and that the other backup anchor set may be preferred for content representation purposes.

At last, we also reproduced the equating results with the final anchor set 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. Table 3 summarizes the possible anchor sets we chose to examine in the item parameter scaling stage.

Table 3: Description of anchor item set for 3rd grade reading

	Anchor item set	**Description**
Grade 3	#1	All “planned” anchor items used (n=27)
	#2	Form 40 anchor items removed (n=20)
	#3	Form 40 removed, Core items 1-10 added (n=30)
	#4*	All “planned” anchor items, Core items 1-11 and 13-21 added

* FDOE’s final anchor set

The Stocking and Lord (1983) equating procedure was implemented for each anchor item set described above, and the impact on final scores was examined as part of the justification for removing items from the anchor set. The equating results, including the transformation coefficients (M1 and M2) for all three anchor sets described in Table 3 are very similar. Those results are presented in more detail in Table 4.

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 4.

Before discussing the information in Table 4, 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, that that 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. The last four columns of Table 4 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 of the scale. Differences this small are likely the result of using different software packages (McKinley, personal communication, 2009), thus the consistency of the results in Table 4 lend support to the accuracy of 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 to 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 that work, was also reviewed. Overall, our impressions were that the entire process is well-organized, the statistical analyses used to identify problematic items are thorough, 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. Moreover, they have built effective checks and balances into the equating process, and while equating is a highly quantitative procedure, there is also considerable judgment involved.

Table 4: 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
3			**FDOE values**		**48.565**	**321.371**		**259**	**284**	**332**	**394**
	All core items used and 27 planned anchors used in calibration		All "planned" anchor items included in anchor set (n=27)		49.73	323.53		259.7	285.3	334.4	397.9
			Form 40 anchor items removed from anchor set (n=20)		50.23	323.93		259.4	285.3	334.9	399.0
			Form 40 removed and Core Items 1-10 added (n=30)		50.32	324.31		259.7	285.6	335.3	399.6
			FDOE's final set : all "planned" anchor items and Core Items 1-11 and 13-21 added (n=47)		49.62	321.76		258.0	283.6	332.6	396.0

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.

McKinley, R. (2010). Personal communication, May 20, 2010.

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.