BUROS
CENTER FOR TESTING

2014 Audit II Report:
Calibration, Equating, and Scoring

Prepared by:

Tzu-Yun Chin, Ph.D.

John P. Poggio, Ph.D.

Consultants to the Florida Department of Education

June 2014

For questions or comments, please contact:
Tzu-Yun Chin, Ph.D.
tchin@buros.org
(402) 472-1414

Background

The Florida Department of Education (FDOE) contracted with the Buros Center for Testing (Buros) to serve as an independent external reviewer for the 2014 Florida EOC assessments. For the 2013 and 2014 assessment operations, Buros was charged with auditing several key measurement components of the Florida Statewide Assessment Program. This report documents our observations and findings of the 2014 scaling operations of the 2014 Florida Comprehensive Assessment Tests 2.0 (FCAT 2.0) for Reading, Mathematics, and Science and for the End-of-Course (EOC) exams. We also include our analyses for the eight (8) contracted assessments in this report.

Multiple test forms are constructed each year for each FCAT 2.0 and EOC assessment. Each year, all forms are randomly administered to students with the exception of students who need special accommodations. In general, the test forms were composed of a set of core items that occurred constantly (same relative position without alteration) on every form of the assessment and a set of items unique to each form. These form-specific items were incorporated to pilot new items or to serve as external linkages for scaling. Students' scores on the FCAT 2.0 assessments were derived based on their performance on the core items only (i.e., the form-specific items did not contribute to student scores) whereas the EOC assessments included a few scoring items that were form-specific. Depending on the assessment, some core items were also designated to be internal anchors for the purpose of equating. There were other core items that could be used as additional internal anchors if needed. These core items may serve as backups when the designated anchor items did not exhibit the desired psychometric properties or they may supplement the planned anchor set to form additional linkage. However, the planned anchor

sets were adequate for all 2014 FCAT 2.0 and EOC assessments, thus none of the backup anchors were called upon for this year's equating operation.

All FCAT 2.0 and EOC items were dichotomously scored (i.e., correct/incorrect). Item response theory (IRT) is used for calibration, scaling, and scoring. FDOE applies the two-parameter logistic (2PL) model to the gridded-response and fill-in response items, whereas the three-parameter logistic (3PL) model is applied to the multiple-choice items (one best choice selected-response items).

In order to ensure the quality of the statewide assessments, FDOE implements multiple layers of accuracy checks for the calibration, scaling, and scoring operations. The calibration, scaling, and scoring specifications were delineated in the *2014 FCAT2.0/EOC Calibration, Equating, and Scaling Specifications* (Version 0.5; referred to as the *Specifications*, hereafter). This document is agreed upon between FDOE and NCS Pearson, Inc. (Pearson). A team of experts was assembled for the multiple tasks involved in the processes described in the *Specifications*. The parties who participated in the 2014 calibration were FDOE, Pearson, Human Resources Research Organization (HumRRO), Test Development Center (TDC), and the Buros Center. Pearson was the primary contractor responsible for the calibration, scaling, and scoring for the FCAT 2.0 and EOC assessments. As the vendor for the FCAT 2.0 and EOC exams, Pearson was also responsible for other testing activities such as test construction and item review, administration, data preparation, and reporting. HumRRO was contracted to perform independent checks in the calibration, scaling, and scoring analyses for all grade levels and subject areas. In addition to Pearson and HumRRO, FDOE staff also replicated and verified all of the important psychometric analyses that Pearson and HumRRO conducted.

Analysis results from the three parties (Pearson, HumRRO, and FDOE) were compared by the independent teams. When a discrepancy occurred, the team searched for the source of the error systematically, then addressed and fixed the problem with the analyses of item(s). Item parameters, equating coefficients, and score estimates could not be approved until achieving a close, if not exact, match among all three parties. Due to slight differences in software programs or computing algorithms employed by different parties, a close match would generally refer to a third or fourth decimal place difference in item parameter estimates. TDC staff served as the content experts for the team. They were consulted when content issues were raised for specific items or when the content coverage and balance of an anchor item set needed to be evaluated.

The vast majority of the calibration, scaling, and scoring activities and tasks regarding the 2014 FCAT 2.0 and EOC assessments occurred between May 2 and May 23, 2014. During this period of time, Buros consultants observed and participated in the decision-making processes related to the psychometric operations. We monitored the email exchanges among the calibration team members, were invited into their discussions on the daily calibration conference calls, and contributed actively to the input being offered and the decisions being made.

For each FCAT 2.0 and EOC assessment, there was at least one planned conference call. This year, each conference call included participants from Pearson, FDOE, HumRRO, TDC, and Buros. During the conference calls, Pearson and HumRRO staffs summarized the preliminary results of their calibration, scaling, and scoring analyses including test design considerations and item properties, and then on the call or shortly thereafter following further analyses as may have been suggested, recommended the final equating solution(s). Feedback and reaction was sought from all conference call participants. According to the feedback received, Pearson conducted additional analyses if necessary. In order to make the final calibration and equating decisions,

FDOE carefully reviewed the reports and statistics produced by Pearson, HumRRO, and Buros and consulted with the team.

In addition, FDOE sought services from Buros to provide an extra level of quality assurance for eight selected assessments. Some of these eight assessments were chosen for the high stakes associated with resultant test scores. Some other assessments were chosen due to significant education policy changes in the state that could potentially augment the psychometric operations. In the role of independent reviewer, Buros conducted the analyses using different algorithms and software packages for item calibration, scaling, and scoring. The objective of the analyses conducted by Buros was not to duplicate by using the identical methods and procedures of the other parties' results, but to triangulate and independently and separately evaluate the consistency and accuracy of the final student scores. With different software packages and estimation algorithms, discrepancies are to be expected at all levels of estimates and statistics and can further lead to different decisions regarding the set of anchor items and the student scores. However, if the scaling of the FCAT 2.0 and EOC assessments can be demonstrated as robust to the differences introduced by methodologies, FDOE and the public can have strong confidence in the scores reported to students and families.

In general, the methodology and analyses Buros conducted involved the following activities:

1) selection of students to be included in the calibration sample according to the *Specifications*,
2) item calibration,
3) review calibration results regarding convergence and the standard errors of estimates (item-model fit was investigated if concerns arose),

4) scaling,
5) review of the anchor item statistics and flagging anchor items for removal,
6) scaling with the reduced anchor-item set, if applicable,
7) scoring with the final scaled item parameter estimates, and
8) estimation of the scale score statistics and the proficiency rate for the group of students selected for impact analysis according to the *Specifications.*

The technical descriptions of the psychometric methodology that Buros adopted for this project are included in the Appendix. It should be noted that the content coverage and balance of the anchor items was not considered in the analyses Buros conducted. However, the operational scaling lead by FDOE did include TDC content experts. TDC content experts reviewed the FDOE's final anchor sets for their representation of the content standards as well as reporting categories and weighed-in on the appropriateness of alterations being contemplated on test score validity.

The eight assessments selected by FDOE for Buros to review were: (1) FCAT 2.0 Grade 3 Reading, (2) FCAT 2.0 Grade 8 Reading, (3) FCAT 2.0 Grade 10 Reading, (4) FCAT 2.0 Grade 6 Mathematics, (5) FCAT 2.0 Grade 8 Mathematics, (6) FCAT 2.0 Grade 5 Science, (7) Algebra I EOC, and (8) Geometry EOC. The verification results of the scaling for these eight assessments are presented in the following sections.

Review of the 2014 FCAT 2.0 Grade 3 Reading Assessment

The base scale for the FCAT 2.0 Grade 3 Reading assessment was established in 2011. In 2014, there were 15 test forms, and each form included 53 items (45 core items and 8 form-specific items). All items were multiple-choice items, and the assessment was delivered solely via paper and pencil. The original anchor item set included 13 internal anchor items and 19 external anchor items.

The calibration sample included 203,189 students after applying the exclusion rules delineated in the *Specifications*. After a holistic review of the anchor items, all anchor items were retained. Students were scored using the scaled item parameters, and then the theta scores were converted to the FCAT 2.0 reporting scale. Buros then conducted the impact analysis of the scaling on student scores. We followed the same selection criterion presented in the *Specifications* for including students in the impact analysis. Table 1 presents the scaling results in terms of student scores, and Table 2 presents the proficiency rates. In Table 1 and Table 2, we included both Buros' results and the final results obtained by FDOE for comparison purposes. The differences in the results were due to different calibration/scoring algorithms and software packages implemented by FDOE and by Buros. As presented in Table 1 and 2, the average scale scores were nearly identical regardless of the scaling approaches, and the differences in the proficiency rates were minimal (within 2%). Thus, our separate and distinct psychometric evaluation confirmed for the Grade 3 Reading Assessment results obtained from the other contractors and FDOE.

Table 1. Scale Score Statistics for Grade 3 Reading (n = 203,200) [Note 1]

	# of Anchor Items Removed	Scale Score	
		Mean	*SD*
Buros' results	0	199.0	22.0
FDOE's results	0	200.3	22.0

Table 2. Percentage of Students at Each Proficiency Level and the Proficiency Rates (%) for Grade 3 Reading (n = 203,200) [Note 1]

	Proficiency Level (%)					Proficiency
	Level 1	Level 2	Level 3	Level 4	Level 5	Rates (%)
Buros' results	19.8	24.9	23.1	22.7	9.6	55.3
FDOE's results	18.7	24.4	23.3	23.4	10.2	56.9

Review of the 2014 FCAT 2.0 Grade 8 Reading Assessment

The base scale for the FCAT 2.0 Grade 8 Reading assessment was established in 2011. In 2014, there were 18 test forms, and each form included 53 items (45 core items and 8 form-specific items). All items were multiple-choice items, and the assessment was delivered online. The original anchor item set included 12 internal anchor items and 20 external anchor items.

The calibration sample included 196,311 students. After a holistic review of the anchor items, four items were identified for removal. Buros then conducted the impact analysis of the scaling on student scores. In addition to our scaling analysis, we conducted extra scaling and scoring using the final anchor set approved by FDOE. Table 3 presents the scaling results in terms of student scores, and Table 4 presents the results in terms of proficiency rates.

As presented in Table 3, the average scale scores were similar regardless of the scaling approaches. In addition, the differences in the proficiency rates shown in Table 4 were quite small (within 2%). Thus, our separate and distinct psychometric evaluation confirmed for the Grade 8 Reading Assessment results obtained from the other contractors and FDOE.

Table 3. Scale Score Statistics for Grade 8 Reading (n = 196,663) [Note 2, Note 3]

	# of Anchor Items Removed	Scale Score	
		Mean	*SD*
Buros' results	4	237.5	23.2
Buros' estimates using FDOE approved anchor set	1	237.5	23.0
FDOE's results	1	237.9	23.1

Table 4. Percentage of Students at Each Proficiency Level and the Proficiency Rates (%) for Grade 8 Reading (n = 196,663) [Note 2, Note 3]

	Proficiency Level (%)					Proficiency
	Level 1	Level 2	Level 3	Level 4	Level 5	Rates (%)
Buros' results	18.0	24.7	25.3	19.4	12.7	57.4
Buros' estimates using FDOE approved anchor set	17.8	24.8	25.5	19.4	12.5	57.4
FDOE's results	18.0	24.8	25.5	19.2	12.5	57.2

Review of the 2014 FCAT 2.0 Grade 10 Reading Assessment

The base scale for the FCAT 2.0 Grade 10 Reading assessment was established in 2011. In 2014, there were 16 test forms, and each form included 53 items (45 core items and 8 form-specific items). All items were multiple-choice items, and the assessment was delivered online. The original anchor item set included 13 internal anchor items and 19 external anchor items. The calibration sample included 186,339 students. After a holistic review of the anchor items, four items were identified for removal. Buros then conducted the impact analysis of the scaling on student scores. In addition to our scaling analysis, we conducted extra scaling and scoring using the final anchor set approved by FDOE, in which two anchor items were removed. Table 5 presents the scaling results in terms of student scores, and Table 6 presents the results in terms of proficiency rates. As presented in Table 5 and 6, the average scale scores and the proficiency rates were almost identical regardless of different scaling solutions. Thus, our separate and distinct psychometric evaluation confirmed for the Grade 10 Reading Assessment results obtained from the other contractors and FDOE.

Table 5. Scale Score Statistics for Grade 10 Reading (n =187,100) Note 2, Note 3

	# of Anchor Items Removed	Scale Score	
		Mean	*SD*
Buros' results	4	245.9	21.3
Buros' estimates using FDOE approved anchor set	2	245.9	21.2
FDOE's results	2	246.2	21.1

Table 6. Percentage of Students at Each Proficiency Level and the Proficiency Rates (%) for Grade 10 Reading (n =187,100) Note 2, Note 3

	Proficiency Level (%)					Proficiency
	Level 1	Level 2	Level 3	Level 4	Level 5	Rates (%)
Buros' results	17.3	27.2	22.0	21.8	11.6	55.5
Buros' estimates using FDOE approved anchor set	17.1	27.4	22.2	21.9	11.4	55.5
FDOE's results	17.3	27.5	22.3	21.8	11.0	55.2

Review of the 2014 FCAT 2.0 Grade 6 Mathematics Assessment

The base scale for the FCAT 2.0 Grade 6 Mathematics assessment was established in 2011. In 2014, there were 14 test forms and each form included 52 items (44 core items and 8 form-specific items). The test forms included both multiple-choice and gridded-response items. The anchor item set included 28 external anchor items. The assessment was administered online.

The calibration sample for Grade 6 Mathematics included 192,034 students. After a holistic review of the anchor items, all anchor items were retained. The impact analysis results are shown in Table 7 and Table 8. As shown in these tables, the resultant average scale scores and the proficiency rates were almost identical despite using different software and estimation algorithm. Thus, our separate and distinct psychometric evaluation confirmed for the Grade 6 Mathematics Assessment results obtained from the other contractors and FDOE.

Table 7. Scale Score Statistics for Grade 6 Mathematics (n =192,685) [Note 1]

	# of Anchor Items Removed	Scale Score	
		Mean	*SD*
Buros' results	0	226.5	22.9
FDOE's results	0	226.7	22.6

Table 8. Percentage of Students at Each Proficiency Level and the Proficiency Rates (%) for Grade 6 Mathematics (n = 192,685) [Note 1]

	Proficiency Level (%)					Proficiency
	Level 1	Level 2	Level 3	Level 4	Level 5	Rates (%)
Buros' results	22.9	23.3	24.3	18.9	10.6	53.8
FDOE's results	23.0	22.7	23.9	19.1	11.3	54.3

Review of the 2014 FCAT 2.0 Grade 8 Mathematics Assessment

The base scale for the FCAT 2.0 Grade 8 Mathematics assessment was established in 2011. In 2014, there were 14 test forms, and each form included 56 items (48 core items and 8 form-specific items). The test forms included both multiple-choice and gridded-response items, and the assessment was delivered via paper and pencil. The original anchor item set included 31 external anchor items and no internal anchor items.

The calibration sample included 158,358 students. FDOE's test-taking policy implemented last year (2013) continues in 2014. According to the policy, the middle school students who enroll in Algebra 1 or Geometry must take the corresponding EOC assessments, and these students do not have to take the FCAT 2.0 Mathematics test. As a consequence, a significant portion of Grade 8 students, who presumably had middle to high mathematics ability, did not take FCAT 2.0 Grade 8 Mathematics.

After a holistic review of the anchor items, one item was identified for removal. Buros then conducted the impact analysis of the scaling on student scores. In addition to our scaling analysis, we conducted extra scaling and scoring using the final anchor set approved by FDOE. As shown in Table 9, the average scale scores were almost identical regardless of the scaling approaches. The proficiency rates (Table 10) were also quite similar and the differences were within 2%. Thus, our separate and distinct psychometric evaluation confirmed for the Grade 8 Mathematics Assessment results obtained from the other contractors and FDOE.

Table 9. Scale Score Statistics for Grade 8 Mathematics (n = 158,434) Note 2, Note 3

	# of Anchor Items Removed	Scale Score	
		Mean	*SD*
Buros' results	1	238.8	20.4
Buros' estimates using FDOE approved anchor set	0	238.7	20.5
FDOE's results	0	238.5	19.8

Table 10. Percentage of Students at Each Proficiency Level and the Proficiency Rates (%) for Grade 8 Mathematics (n = 158,434) Note 2, Note 3

	Proficiency Level (%)					Proficiency
	Level 1	Level 2	Level 3	Level 4	Level 5	Rates (%)
Buros' results	26.9	24.1	29.0	12.8	7.1	49.0
Buros' estimates using FDOE approved anchor set	27.1	24.0	28.8	12.8	7.3	48.9
FDOE's results	27.8	25.2	28.9	11.8	6.3	47.0

Review of the 2014 FCAT 2.0 Grade 5 Science Assessment

The base scale for the FCAT 2.0 Grade 5 Science assessment was established in 2012. In 2014, there were 31 test forms, and each form included 66 items (56 core items and 10 form-specific items). All items were multiple-choice items. The assessment was delivered via paper and pencil. The original anchor item set included 32 external anchor items and no internal anchor items.

The calibration sample included 170,253 students. After a holistic review of the anchor items, all anchor items were retained. Table 11 presents the scaling results in terms of student scores, and Table 12 presents the results in terms of proficiency rates. As presented in these two tables, the average scale scores and the proficiency rates were very close between Buros' and FDOE's results. Thus, our separate and distinct psychometric evaluation confirmed for the Grade 5 Science Assessment results obtained from the other contractors and FDOE.

Table 11. Scale Score Statistics for Grade 5 Science (n = 170,255) [Note 1]

	# of Anchor Items Removed	Scale Score	
		Mean	*SD*
Buros' results	0	201.2	21.7
FDOE's results	0	201.6	21.3

Table 12. Percentage of Students at Each Proficiency Level and the Proficiency Rates (%) for Grade 5 Science (n = 170,255) [Note 1]

	Proficiency Level (%)					Proficiency
	Level 1	Level 2	Level 3	Level 4	Level 5	Rates (%)
Buros' results	20.5	24.9	27.7	13.4	13.6	54.6
FDOE's results	20.2	25.3	28.1	13.2	13.2	54.5

Review of the 2014 Algebra 1 EOC Assessment

The base scale for the 2014 Algebra 1 EOC assessment was established in 2011. In 2014, there were 36 test forms. Each test form included one of the 4 core item sets (54 core items in each set) and a set of 8 form-specific field-test items. The Algebra 1 included both multiple-choice and filled-in response items, and the EOC was delivered online. The original anchor item set had 29 internal anchor items and no external anchor items.

The calibration sample included 196,740 students. After a holistic review of the anchor items, all anchor items were retained. Table 13 presents the scaling results in terms of student scores, and Table 14 presents the results in terms of proficiency rates. As shown in these two tables, the average scale scores were nearly identical regardless of the scaling approaches, and the differences in the proficiency rates were minimal (within 2%). Thus, our separate and distinct psychometric evaluation confirmed for the Algebra 1 EOC results obtained from the other contractors and FDOE.

Table 13. Scale Score Statistics for Algebra 1 EOC (n = 205,832) [Note 1]

	# of Anchor Items Removed	Scale Score	
		Mean	*SD*
Buros' results	0	407.3	29.0
FDOE's results	0	407.7	28.8

Table 14. Percentage of Students at Each Proficiency Level and the Proficiency Rates (%) for Algebra 1 EOC (n = 205,832) [Note 1]

	Proficiency Level (%)					Proficiency Rates (%)
	Level 1	Level 2	Level 3	Level 4	Level 5	
Buros' results	11.5	21.3	38.9	14.3	14.1	67.3
FDOE's results	11.5	22.4	38.1	13.8	14.2	66.2

Review of the 2014 Geometry EOC Assessment

The base scale for the 2014 Geometry EOC assessment was established in 2012. In 2014, there were 36 test forms. Each test form included 1 of the 4 core item sets (54 core items per item set) and a set of 8 form-specific field-test items. The assessment was delivered online and consisted of both multiple-choice items and fill-in response items. The original anchor item set included 28 internal anchor items and no external anchors.

The calibration sample included 171,011 students. After a holistic review of the anchor items, two items were identified for removal. For this assessment, FDOE also removed the same anchor items Buros identified. Table 15 presents the scaling results in terms of student scores, and Table 16 presents the results in terms of proficiency rates. As shown in these two tables, the average scale scores were nearly identical regardless of the scaling approaches, and the differences in the proficiency rates were minimal (within 2%). Thus, our separate and distinct psychometric evaluation confirmed for the Geometry EOC Assessment results obtained from the other contractors and FDOE.

Table 15. Scale Score Statistics for Geometry EOC (n = 174,673) [Note 1]

	# of Anchor Items Removed	Scale Score	
		Mean	*SD*
Buros' results	2	403.5	30.1
FDOE's results	2	403.0	30.1

Table 16. Percentage of Students at Each Proficiency Level and the Proficiency Rates (%) for Geometry EOC (n = 174,673) [Note 1]

	Proficiency Level (%)					Proficiency
	Level 1	Level 2	Level 3	Level 4	Level 5	Rates (%)
Buros' results	11.4	23.1	30.1	20.8	14.7	65.6
FDOE's results	11.9	24.0	30.1	20.2	13.8	64.1

Other Observations

During our planning, study and analyses, we found some estimation difficulties with the pseudo-guessing parameters (*c* parameters) and the discrimination parameters (*a* parameters) of some multiple-choice items. After close examination of the item parameter estimates and the empirical response curves for those items exhibiting estimation difficulties, it was noticed that the empirical response curves often lacked clear inflection points, lower asymptotes (lower limits), or a uniform smooth rise, for those items. We also observed those items often had *c* parameter estimates approaching zero, thus indicating inability to estimate guessing. In several calibrations, we had to fix some of the *c* parameters at an arbitrary value to achieve stable estimates for other parameters. The percentages of such items for Science (Grade 5), Mathematics (Grade 6 and 8), and EOC (Algebra I and Geometry) calibrations were between 0-2% while the numbers for Reading calibrations (Grade 3, 8 and 10) were around 10%.

Although these phenomena might be due to the algorithm implemented by the software package configuration we used (IRTPRO 2.1, SSI, 2011), we understand that Pearson also occasionally observed elevated standard error of estimates for some anchor items. These items did not affect the equating operation this year. However, we strongly advise FDOE, whenever possible, not to include items with elevated standard error of estimates as equating items in the anchor item sets planned for the future. The elevated standard error of estimates indicates potential instability of the estimated parameters, thus including those items may threaten the integrity of future anchor sets and resultant student scores.

Conclusions

Buros monitored the calibration, scaling, and scoring processes for the 2014 FCAT 2.0 and EOC assessments throughout the May 2014 study period. Access to all data, methods and procedures was made available to Buros. The psychometric team assembled by FDOE implemented multiple layers of quality controls to ensure the accuracy of data handling. Multiple statistics and graphs were reviewed and deliberated by the team in order to make anchor item evaluation and anchor set selection. Questions surfaced, expert information was provided if not immediately, then in hours. FDOE demonstrated effective leadership during the course by inviting and facilitating open and thorough discussions as well as by thoughtfully and wisely considering different opinions whenever possible. The content experts from TDC were also actively engaged during all conference sessions in order to safeguard the final anchor set with proper content coverage and a sufficient number of anchor items for each reporting category. Invariably, the team on occasion found it necessary to discuss specific core items that exhibited unexpected response patterns or item statistics. Based on our observation, we found the calibration, equating, and scoring activity and operation was conducted according to professional standards and best practices. In all respects, we find the FDOE process of study, audit, verification and then independent third party review and follow-along to set the benchmark for high stakes testing in the US. FDOE is committed undoubtedly to safeguarding the rights of students, educators, policy makers, and the public. The Florida assessment programs at all levels and content areas can and should serve as the model for all other states relying on assessment to support decisions being made.

In addition, Buros performed independent analyses for eight FCAT 2.0 or EOC assessments selected by FDOE. The objective of the analyses Buros conducted was not to

replicate but to verify the stability of FDOE's final results using different estimation methods. As presented in the previous sections, we were able to achieve very satisfactory convergence regarding student scores and group proficiency levels although the anchor sets selected by Buros and FDOE sometimes had minor discrepancies.

In conclusion and in summary, from our observations and verification, we have confidence in the resultant 2014 EOC student scores, not just for the assessments we evaluated, but the collection of comparable FDOE assessments.

Report Notes

Note 1. The differences between Buros' results and FDOE's results were due to different calibration/scoring algorithm and software.

Note 2. Buros did not provide content review of the anchor items. However, TDC content experts reviewed the FDOE's final anchor sets for their representation of the content standards as well as reporting categories. Differences in the anchor item removals were often due to content consideration.

Note 3. The differences between Buros' results and the results from Buros' estimates using FDOE-approved anchor set were due to different anchor item retention; the differences between FDOE's results and the results from Buros' estimates using FDOE-approved anchor set were due to the impacts of different calibration/scoring algorithm and software; finally, the differences between Buros' results and FDOE's results were due to different anchor item retention plus different calibration/scoring algorithm and software.

Appendix

Methodology

The analyses were performed in five steps. These steps were as follows:

- Calibration
- Scaling
- Flagging
- Scoring

1. CALIBRATION

Calibration was conducted using IRTPRO 2.1 (Cai, Thissen, & du Toit, 2011). The 3PL model was employed for the multiple-choice items, and the 2PL model was used for the gridded-response or filled-in-response items. Log-normal prior was used for the discrimination (*a*) parameter and Beta prior was used for the pseudo-guessing (*c*) parameter. Bock-Aitkin Expectation Maximization (EM) algorithm (BAEM; Bock & Aitkin, 1981) was used for calibration. All initial calibration runs of the eight assessments converged. However, some assessments have items with extremely low *c* estimates along with very large standard errors of estimates from the initial run; in these cases, the accuracy of their *a* parameter estimates also suffered. In order to obtain stable *a* estimates, the *c* parameters for these items were fixed at a value very close to zero (logit c=-10).

2. SCALING

Scaling was performed with the Stocking-Lord algorithm using IRTEQ (Han, 2009). The minimization statistic was computed using uniformly distributed values of theta over the range -3.0 to 3.0.

3. FLAGGING

Item flagging was performed using three statistics: D^2 (Wells, Hambleton & Meng, 2011), the weighted root mean squared difference (WRMSD), and the weighted mean absolute difference (WMAD) statistics. All statistics were computed using equally spaced values of theta over the range -4.0 to 4.0. Anchor item removal was deliberated with a holistic approach that the aforementioned statistics along with other item information were considered (e.g., ICC curves).

4. SCORING

Scoring was conducted using IRTPRO 2.1 using the Maximum a Posteriori (MAP) estimator. To approximate the maximum likelihood estimator FDOE adopted, a diffused prior, N(0,100), was used as the theta prior.

REFERENCE

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459.

Cai, L., du Toit, S. H. C., & Thissen, D. (2011). IRTPRO: Flexible professional item response theory modeling for patient reported outcomes (Version 2.1) [Computer software]. Chicago, IL: Scientific Software International.

Han, K. T. (2009). IRTEQ: Windows application that implements IRT scaling and equating [Computer software]. *Applied Psychological Measurement*, 33(6), 491-493.

Wells, C. S., Hambleton, R. K., & Meng, Y. (2011). *An examination of two procedures for identifying consequential item parameter drift.* (Center for Educational Assessment Research Report No. 761). Amherst, MA: University of Massachusetts.