BUROS
Center for Testing

Replication Analyses for Florida's K-12 Statewide Assessment Program

Final Report

Contractor:

Buros Center for Testing

The University of Nebraska-Lincoln

Subcontactor:

Sireci Psychometric Services, Inc.

Buros Center for Testing

August 4, 2011

Sireci Psychometric Services, Inc.
43 Whittier Street
Florence, MA 01062, USA

Table of Contents

I. Introduction

Calibration, scaling, and equating are perhaps the most technical aspects of large-scale educational testing programs. When students' educational progress is being monitored over time, either using growth or status models, equating becomes a fundamental validity issue. If the equating is not done properly, it cannot be determined whether changes in test scores over time are due to students' progress or due to differences in test difficulty or an equating error. For this reason, the State of Florida Department of Education (FDOE) includes important quality control checks on the calibration, scaling, and equating processes for the Florida K-12 Statewide Assessment Program. This assessment program includes the Florida Comprehensive Assessment Tests 2.0 (FCAT) and end-of-course (EOC) exams. These quality control checks help ensure that when the FDOE reports FCAT test results, they are accurate.

As part of the quality control checks on the FCAT, the FDOE contracted with the Buros Center for Testing (its Institute for Assessment Consultation and Outreach) to completely replicate the calibration, scaling, and equating analyses for selected FCAT assessments. As included as part of its contract with the State of Florida, Department of Education, Buros explicitly subcontracted with Sireci Psychometric Services (SPS), Inc. to complete the this portion of the work on the FCAT. The vendor for the Florida Assessment Program is Pearson, who was responsible for test administration, scoring, item calibration, scaling, and equating, among other testing activities. In addition to the SPS/Buros replications, the FDOE also contracted with HumRRO to replicate the analyses in all grades and subject areas. In this report, we summarize our comprehensive series of replication analyses on the tests selected by the FDOE to undergo an extra level of quality control.

The specific tests included in our analyses were,

- Grade 3 Reading
- Grade 8 Math
- Grade 8 Science
- Grade 10 Reading
- Algebra I End-of-Course (EOC) Exam

This report includes a separate chapter that summarizes our analyses for each of these five exams. Although there were some differences across the exams due to the differing nature of each exam, the process was similar with respect to receipt of the data and most statistical analyses. Essentially, for each exam, the process was as follows.

Pearson alerted SPS when the data for an exam was posted to a secure sFTP site. The data were subsequently downloaded for analysis. The analyses conducted included descriptive statistics, calibration using item response theory (IRT), evaluation of anchor items (where relevant), analysis of model fit, equating analyses, and comparing item, equating, and test score results with those obtained by Pearson and HumRRO.

During the scaling, equating, calibration period, at least one conference call was held for each exam to discuss the data, the progress of the analyses, and the results. The Reading, Math, and Science tests were discussed on calls occurring from May 5, 2011 through May 13, 2011. The calls to discuss the Algebra I EOC exam occurred during May 25-26, 2011. SPS/Buros participated on all the calls for all exams, not just the ones on which we performed replication analyses. Each call included representatives from Pearson, the FDOE, HumRRO, and SPS/Buros. Through these calls, we noted the great care taken by Pearson to make data available, to summarize the results of all analyses, to conduct and summarize additional analyses

when questions arose, and to facilitate the discussion among the various participants. Pearson posted their results to the sFTP site and these results were discussed and compared with our own. We were impressed with the efficiency and accuracy in which Pearson posted the data, addressed any problems, completed analyses, and facilitated the conference calls.

The separate chapters in this report provide the details regarding our analyses for each grade. In general, the analysis procedures for each test were as follows.

1. Download the data and recreate the calibration sample by applying the exclusion rules determined by FDOE and Pearson.
2. Completely rescore students' responses to core items (and anchor items when applicable) using the scoring keys provided by Pearson.
3. Perform classical item analyses to flag any items with questionable statistics.
4. Calibrate the items using the IRT models specified by Pearson/FDOE.
5. Check whether the item parameters we computed were identical to those computed by Pearson.
6. Inspect items using (a) Pearson/FDOE flagging criteria, and (b) IRT residual analysis to identify any potentially problematic items.
7. Perform equating analyses. For Grade 3 Reading, Grade 10 Reading and Grade 8 Math assessments, equipercentile equating was conducted to link the 2011 scale score distribution to the 2010 distribution. For the Grade 8 Science assessment, the Stocking and Lord method was used to place the 2011 scale onto the base scale. For the Algebra I EOC assessment, equating was not performed.
8. Compare the equating functions, raw score and scale score distributions we computed with those computed by Pearson.

9. For the Grade 3 and Grade 10 Reading and the Grade 8 Math, create raw-to-scale score conversion tables and compare with those created by Pearson.

To complete these analyses, several software packages were used including Bilog-MG, Multilog, RAGE-RGEQUATE, ResidPlots, and STUIRT. Citations for and descriptions of these software packages are included in the chapters summarizing each report. Bilog-MG was used for the Grade 3 Reading exam because it included only multiple-choice items. Multilog, an IRT program that can handle a mixed-format test (i.e., multiple-choice and free-response items), was used for the other exams.

General Conclusion/Summary

Although our conclusions with respect to each test analyzed appear in the specific chapter corresponding to each test, based on our results, we can conclude the calibration, scaling, and equating results, and the final scores reported for students on the FCAT 2.0, Algebra I EOC, and Grade 8 Science assessments are accurate. We independently replicated all analyses for the tests described in this report, and did supplementary analyses to investigate whether problems existed that may have affected the results. We found that all problematic items were flagged by Pearson and were adequately discussed and resolved by all parties. We agreed with all decisions that were made throughout the process and we confirm that the calibration, scaling, and equating analyses were done correctly and that the resulting scores are accurate. Details supporting these conclusions can be found in the following chapters.

II. Replication of the 2011 FCAT 2.0 Grade 3 Reading Assessment

In this chapter we describe our replication analyses conducted on the 2011 FCAT 2.0 Grade 3 Reading assessment. The procedures involving the Grade 3 Reading assessment (as well as the Grade 10 Reading and Grade 8 Math assessments) differed from previous years because the 2011 test represented a new test with respect to content, which required a new scale to be established. Therefore, particular attention was placed on the accuracy and appropriateness of the item parameter calibrations, and on the accuracy and appropriateness of the equipercentile equating that was used to link the scale score distribution for 2011 to the 2010 distribution.

Calibration Sample and Demographic Variables

The initial calibration file we downloaded from the sFTP site contained 178,500 students. We first excluded students who had a school type of 10, 11, 14, or 99 and those who used large print or Braille. From the excluded sample, we selected only students with reportable scores (Score Flag = 1 in the data file). Using the exclusion rules, we were able to create the same calibration sample reported by Pearson (N = 175,893). The demographics for the calibration sample, shown in Table II-1, were also identical to that reported by Pearson.

Rescoring Item Responses and Flagging Items via Classical Item Statistics

Once the calibration sample was created, we next rescored the raw core item responses and compared them to the scored item responses provided in the data file. There were 45 core items, which were all dichotomously scored, multiple-choice items. The raw item responses were rescored using the answer key provided in the test map file. A correct response was given a 1 while an incorrect response was given a 0. Each rescored core item was identical to the scored items in the file provided by Pearson.

Table II-1

Demographics for Grade 3 Reading Calibration Sample

		Pearson	SPS/Buros
Gender	Female	85,313	85,313
	Male	90,414	90,414
	Unknown	166	166
Ethnicity	Asian	4,388	4,388
	Black	42,443	42,443
	Hispanic	52,256	52,256
	American Indian/Alaskan	740	740
	Multiracial	5,527	5,527
	Native Hawaiian/Pacific	145	145
	Unknown	378	378
	White	70,016	70,016

We next used classical item statistics to flag potentially problematic items. Items with the following characteristics (determined by Pearson) were flagged:

- Classical item discrimination ($r_{pbi\text{-}c}$) was less than 0.2,
- Classical item difficulty (*p*-value) was greater than 0.9 or less than 0.15,
- An incorrect option was selected by more than 40% of the sample,
- The *p*-value on any one form differed from the overall *p*-value by more than $|0.08|$.

Using the above criteria, one item was flagged (SEQ Item 35) because the *p*-value was greater than 0.90 ($p = 0.903$). Pearson did not flag item 35, but instead flagged item 5 because more than 40% of the examinees responded to one of the incorrect options on a several of the forms; however, because the overall percentage of students who chose the incorrect option was less than 40%, we did not flag this item. Furthermore, we agree that both items should have been included in further analyses.

Item Parameter Calibration

Item parameter calibration was conducted using the computer program BILOG-MG[1] (Zimowski, Muraki, Mislevy, & Bock, 2003) on the calibration sample. The three-parameter logistic model (3PLM) was used for the multiple-choice items. The default prior[2] was implemented only for the *c* parameter in the 3PLM. Sample BILOG-MG code is provided in Appendix II-A.

BILOG-MG successfully converged within 22 EM cycles. The item parameter estimates and their corresponding standard errors were reasonable values. Furthermore, the item parameter estimates were nearly identical to those reported by Pearson; the *a*-, *b*-, and *c*-parameter estimates correlated with the values reported by Pearson to 0.99 and any absolute differences between respective *a*-, *b*- and *c*-parameter estimates were less than 0.01.

Items were flagged for detailed inspection using the following criteria provided by Pearson and FDOE: $a < 0.5$, $2.0 < b < -2.0$, or $c < .05$. Table II-2 reports the flagged items and the reason for being flagged.

Table II-2

Items Flagged Using Pearson/FDOE Criteria

Item (SEQ)	Reason	*a*	*b*	*c*	Model Fit
1	$b < -2$ $c < 0.05$	0.69	-2.02	0.02	Good
5	$c < 0.05$	0.56	-0.24	0.01	Acceptable
25	$c < 0.05$	0.99	-1.59	0.02	Good
32	$c < 0.05$	0.70	-0.82	0.02	Good
42	$c < 0.05$	0.81	-0.76	0.02	Good
45	$c < 0.05$	0.82	-1.66	0.01	Acceptable

[1] BILOG-MG was used to replicate the Grade 3 Reading in part because the commercial version of MULTILOG that was available to us at the time could not handle the large sample size, and since the Grade 3 Reading test involved only multiple-choice items, the BILOG program was appropriate. However, for the other tests, we were able to use Pearson's extended version of MULTILOG, which could handle very large sample sizes. Nevertheless, replicating the procedures using another calibration computer program in this case was an interesting activity that lends more credence to the results.

[2] The prior for the *c* parameter in Bilog-MG is based on the beta distribution with parameter values of 6 and 16.

The item parameter calibration and model fit of the flagged items were further inspected to determine if they should be excluded from additional analyses. Model fit was examined via an inspection of raw residuals around the item characteristic curve (ICC) that is defined by the item parameter estimates. The computer program ResidPlots (Liang, Han, & Hambleton, 2008) was used to examine model fit. Reasonable or acceptable model fit occurs when the majority of the observed proportions are randomly distributed around the ICC, with very few observed points falling far from the ICC. For each flagged item, the item parameter estimation and model fit was acceptable in that the observed proportions were close to the expected value given by the ICC (Appendix II-B contains model fit plots for each flagged item). Therefore, given that there were no key check issues with these items and that the item statistics and model fit was acceptable, we agreed that these items should be included in the equipercentile equating.

In addition to inspecting the model fit for the flagged items, we examined the model fit for all items. The model exhibited excellent to acceptable fit for most of the items. However, there were a few items that exhibited a small magnitude of misfit. For example, items (SEQ) 7, 28, 42 and 43 exhibited a small magnitude misfit. Figures II-1 to II-4 provide plots for the observed and expected ICC plots. The solid curve represents the ICC for the item, which shows the probability that students will answer the item correctly, given their IRT proficiency estimate (i.e., $\hat{\theta}$). The circles shown in the plots represent the proportion of students within a specific interval on the IRT scale that actually did answer the item correctly. Comparing the distance between the circles and the curve indicates the closeness in "fit" between what the IRT model predicts for the item and how students actually performed on the item. The vertical lines represent the confidence bands for these (conditional) observed proportion correct statistics based on 3 standard errors. Items in which the observed proportions fall far from the model-

based ICC (i.e., outside the confidence bands) are indicative of poor fit. For these three items, the misfit was small. Therefore, including these items in the equipercentile equating will have a negligible effect and removing the items may have a detrimental effect on reliability. Therefore, given all of the analyses to this point, we agree that all of the core items should be included in the equipercentile equating.

Figure II-1. *Model fit plot for item (SEQ) 7*

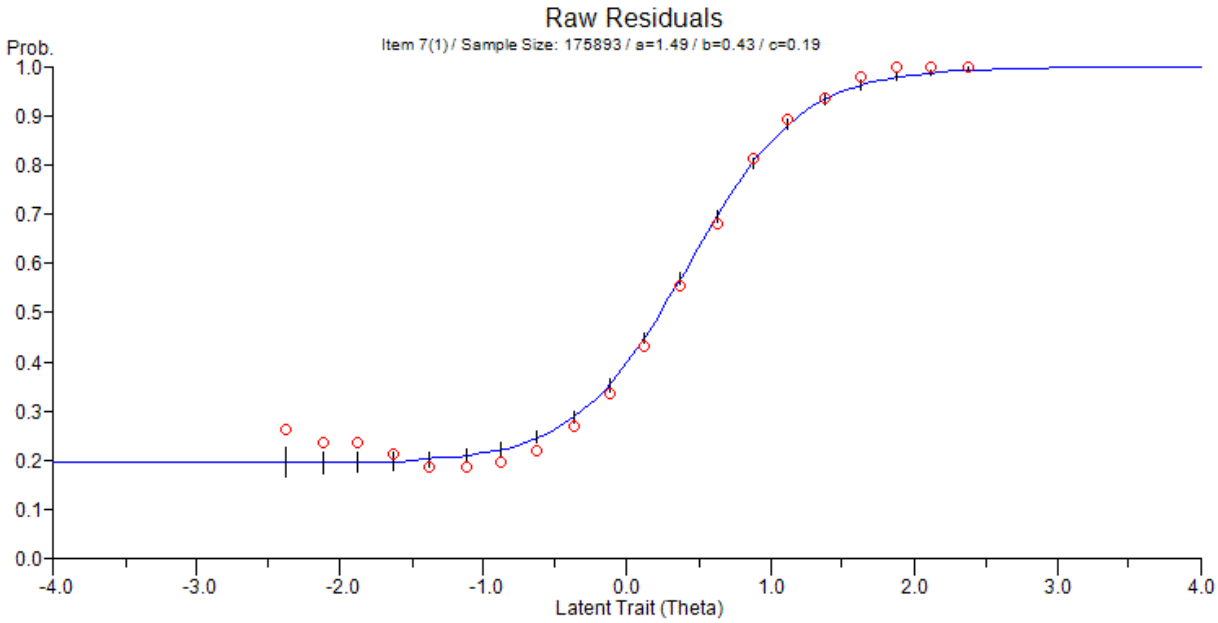


Figure II-2. *Model fit plot for item (SEQ) 36*

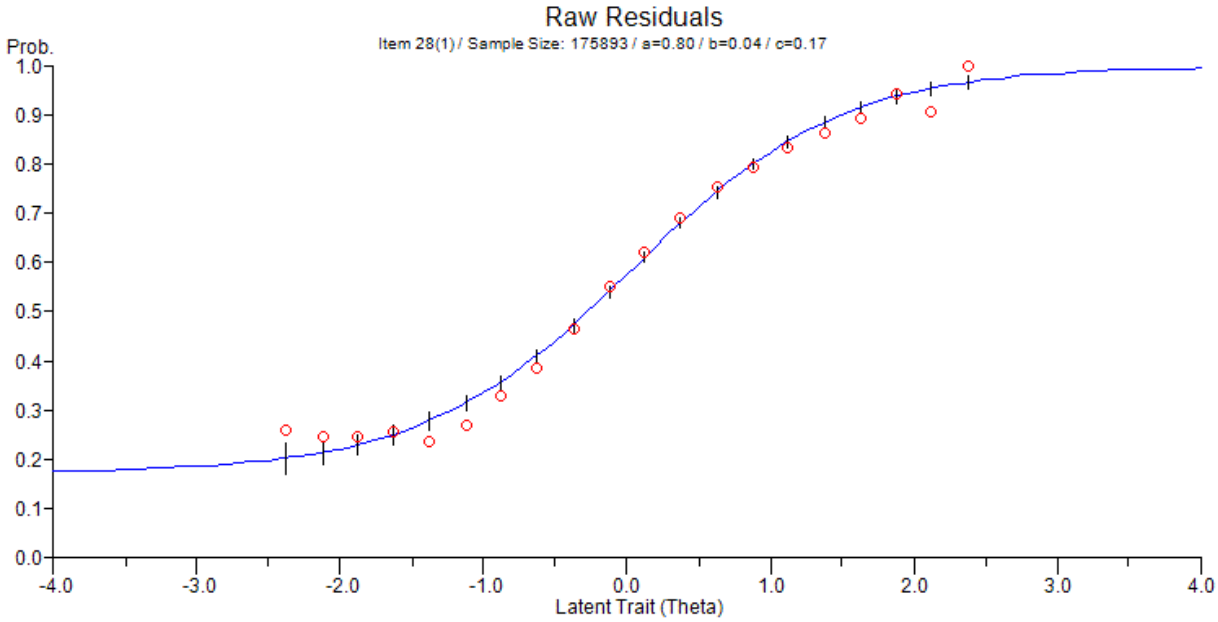


Figure II-3. *Model fit plot for item (SEQ) 50*

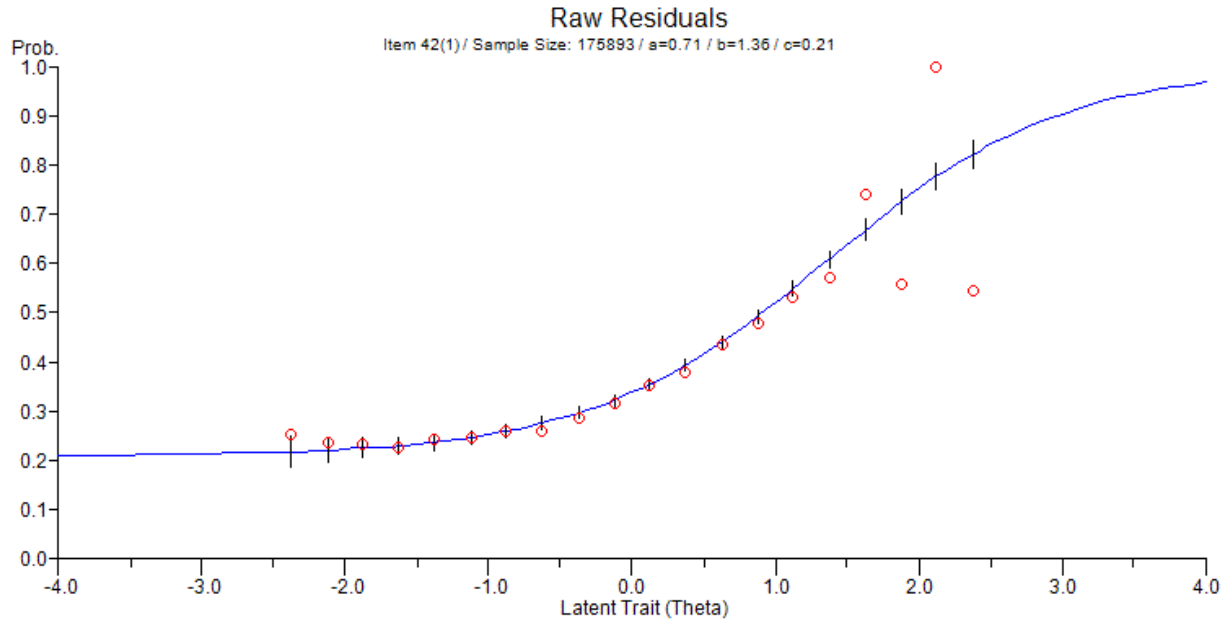
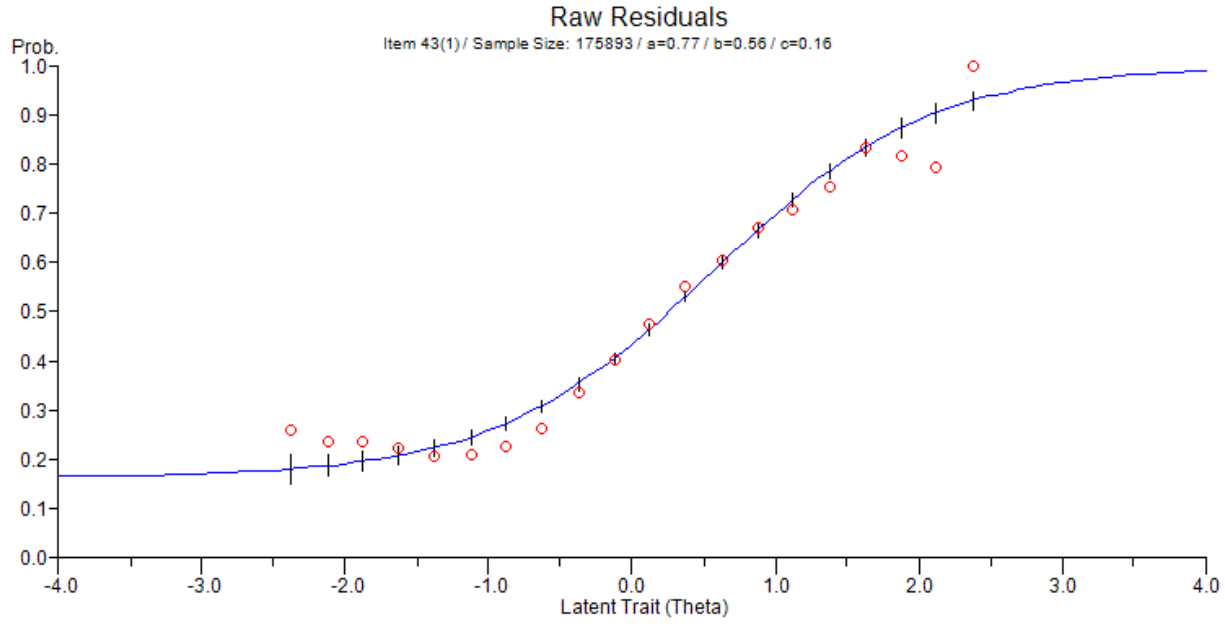


Figure II-4. *Model fit plot for item (SEQ) 51*



Scale Scores and Equipercentile Equating

Before estimating the scale scores for each student, the new item parameter estimates (a_{New}, b_{New} and c_{New}) were transformed onto a scale with a mean of 300 and standard deviation of 50 as follows:

$$a_{SS} = \frac{a_{New}}{50},$$

$$b_{SS} = b_{New} * 50 + 300$$

and

$$c_{SS} = c_{New}.$$

The newly transformed item parameters (a_{SS}, b_{SS}, and c_{SS}) were very similar to those reported by Pearson and were used to estimate the students' IRT ability estimate. The computer program IRT Score Estimation (Chien, Hsu, and Shin, 2011) was used to perform maximum likelihood estimation. Table II-3 reports the descriptive statistics for the raw scores and scale scores.

Table II-3

Descriptive Statistics for Proficiency Estimates

	Raw Score	IRT Scale Score
Mean	30.93	301.45
Standard Deviation	9.49	56.84
Kurtosis	-0.56	1.37
Skewness	-0.61	0.09

The equipercentile equating was conducted to adjust the 2011 scale score distribution such that it was equivalent to the 2010 distribution using the computer program RAGE-RGEQUATE (Zeng, Kolen, Hanson, Cui & Chien, 2005). The frequency distribution for the scale scores was created and used with the scale score frequency distribution from 2010 to create a conversion table linking the 2011 scale scores to the 2010 scale scores for the Grade 10 Reading test. The operational postsmooth limit was used in the equating. The postsmooth lower limit was based on the following equation: (percent of examinees who received a 100 scale score + 0.5)/100. The postsmooth upper limit was based on the following equation: (percent of examinees who received a 500 scale score + 0.5)/100. In this case, the lower and upper limit was

0.011[3] and 0.0168, respectively. Therefore, the equipercentile was run twice - once for each limit. The final conversion table was a combination of the two equatings - equated scale scores for 100 to 300 were based on the postsmooth lower limit value and the equated scale scores for 300 to 500 were based on the postsmooth upper limit.

The final conversion table was nearly identical to that constructed by Pearson. Our equated scores correlated 0.9999 to those reported by Pearson and differed by no more than 1 point for any equated score. The proportion of exact agreement exceeded 95%. To illustrate the high level of agreement, Figure 5 compares the Pearson and Sireci Psychometric Services (SPS) equated scale scores from the conversion table to the identity line that represents exact agreement.

The solid line represents the identity line and runs perfectly through the majority of the equated scores. Although a few equated scores differed by one point, the difference was likely due to rounding error in the item parameter estimates and scales scores as well as the fact that we used BILOG-MG instead of MULTILOG. Because the conversion table was nearly identical, we agree with Pearson's final conversion table.

After creating the conversion table, we transformed the 2011 scale scores onto the 2010 scale. The descriptive statistics for the equated scores, shown in Table II-4, were nearly identical to those reported by Pearson.

[3] The lower postsmooth limit was based on the 2010 frequency distribution because the percent of students with a scale score of 100 was higher for 2010 compared to 2011.

Figure II-5. Plot comparing Pearson and SPS equated scale scores in conversion table.

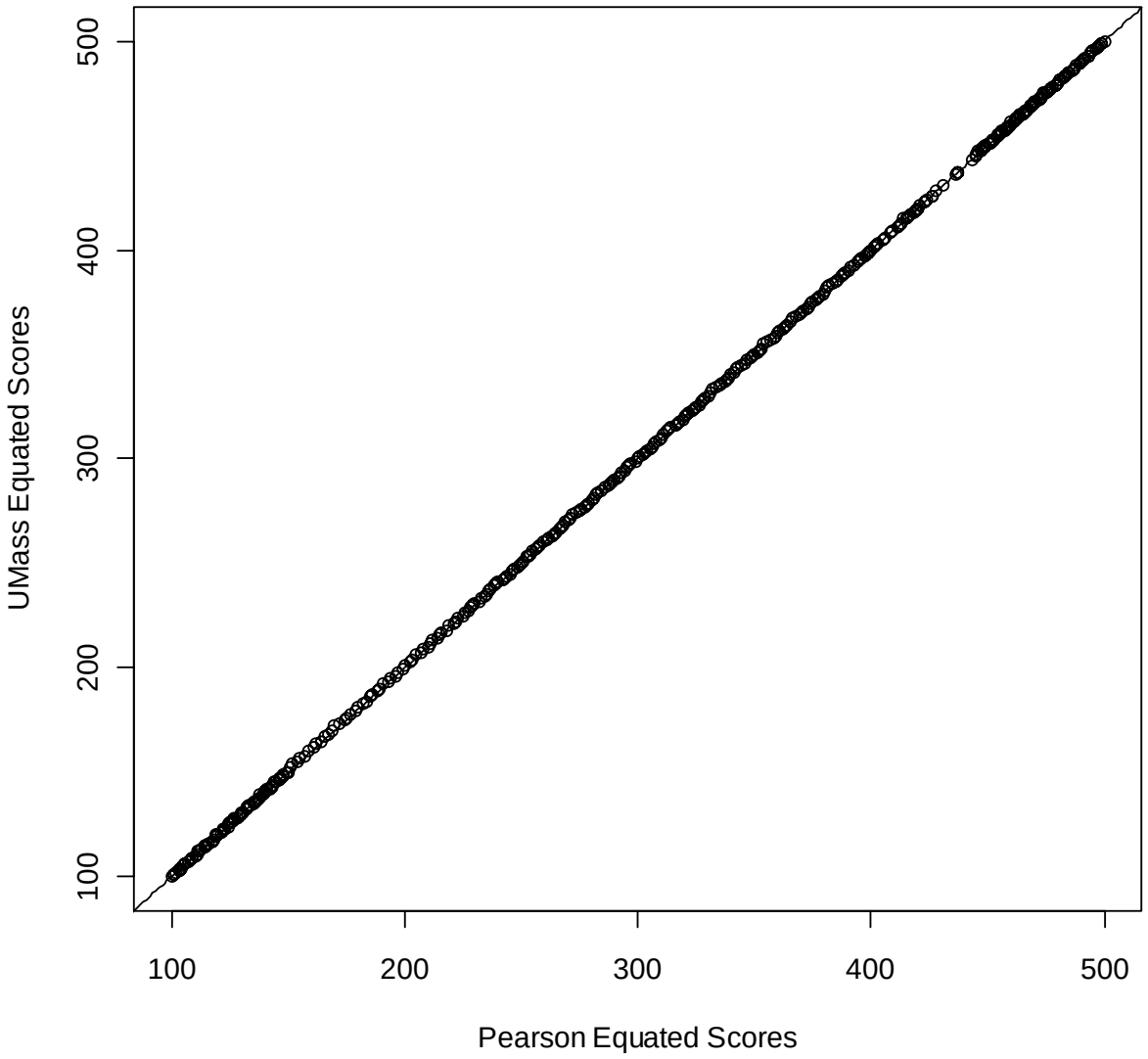


Table II-4

Descriptive Statistics for Equated Proficiency Estimates

	Pearson	SPS
Mean	314.37	314.36
Standard Deviation	61.19	61.19
Kurtosis	0.96	0.96
Skewness	-0.16	-0.16

Conclusion

In summary, SPS was able to successfully replicate Pearson's operational procedures and results including creating the calibration sample given the exclusion/inclusion rules, scoring the raw item responses, verifying the quality of the items (item statistics and model fit), reproducing identical item parameter estimates, and the (nearly) identical conversion table and scale score distribution in regard to the 2001 FCAT 2.0 Grade 3 Reading Assessment. Given this successful replication, we feel confident that the operational procedures were conducted correctly.

Appendix II-A

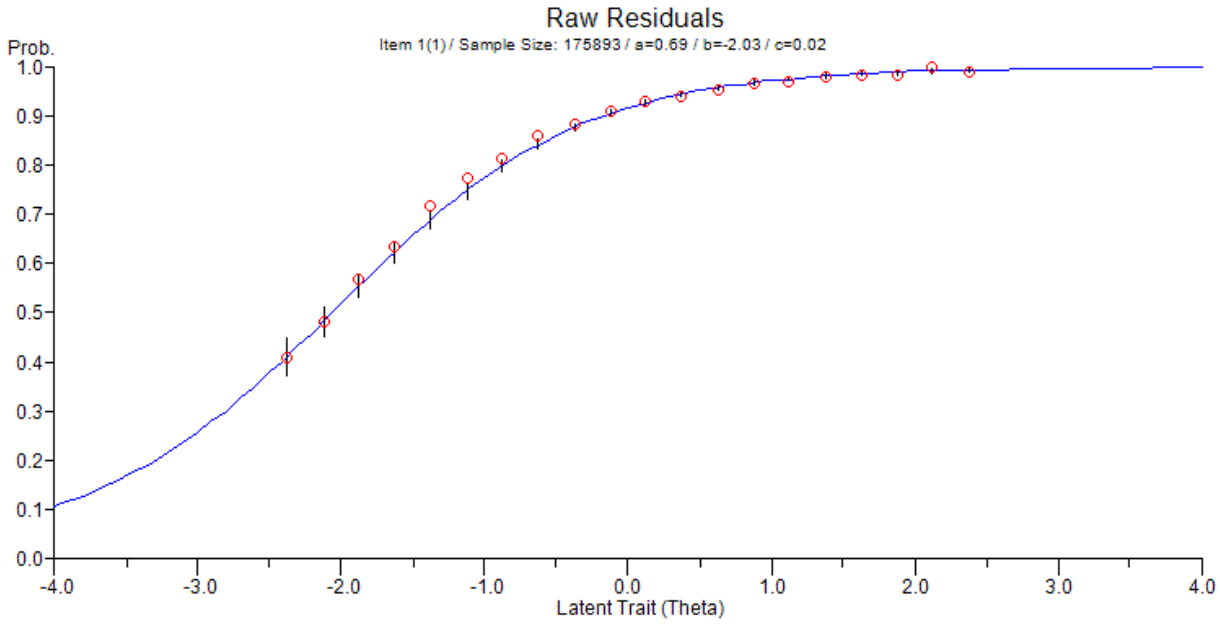
Sample BILOG Code for Grade 3 Reading

```
FCAT 2.0 2011
Grade 3 Reading
>GLOBAL DFN='G3RDG.DAT', NPA=3, SAVE;
>SAVE PAR='G3RDG.PAR', SCO='G3RDG.SCO';
>LENGTH NITEMS = (45);
>INPUT NTOTAL = 45, NIDCHAR = 4, NALT = 4;
>ITEMS INUMBERS = 1(1)45,INAMES = (i01(1)i45);
>TEST TNAME = 'FCAT',INUMBER=(1(1)45);
(4A,T1,45A1)
>CALIB NQPT=41,CYCLES=120,GPRI,NOTPR,NOSPR,NEW=4,CRIT=0.0010;
>SCORE METHOD=1, NOPRINT;
```

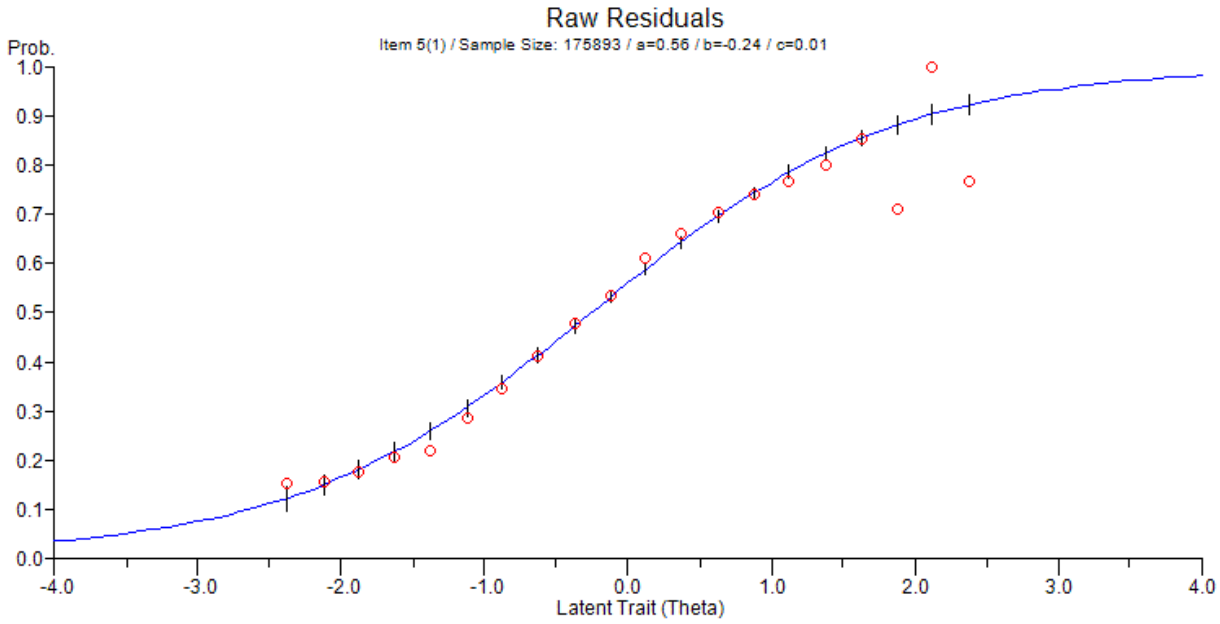
Appendix II-B

Model Fit Plots for Items Flagged During Item Calibration Inspection

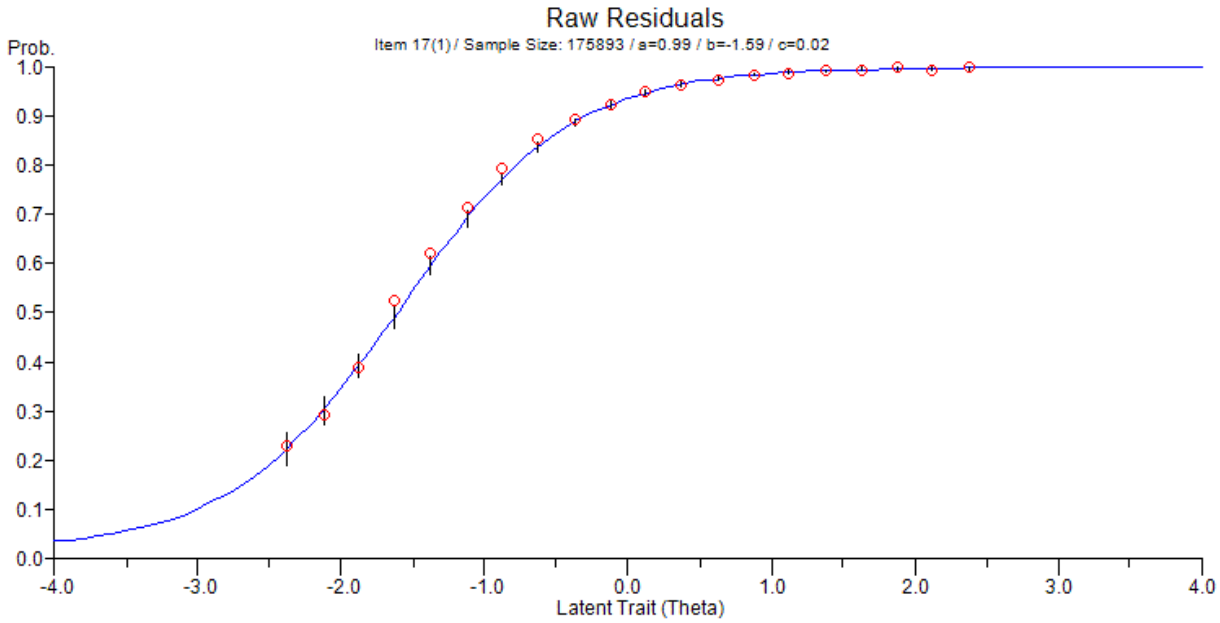
Item (SEQ) 1



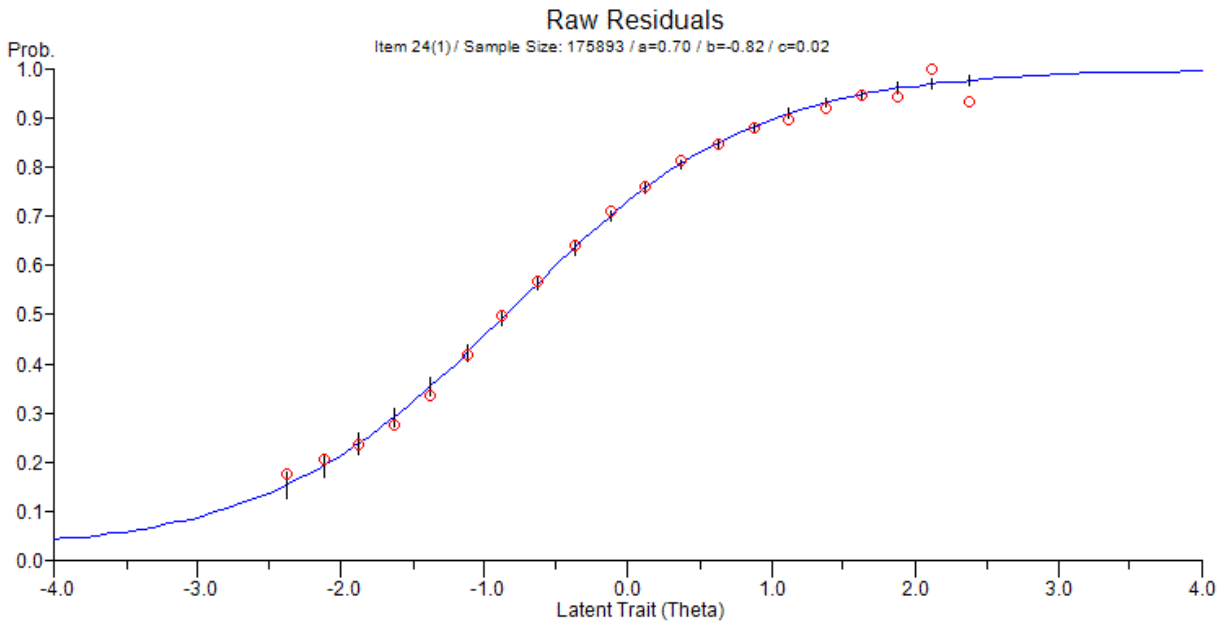
Item (SEQ) 5



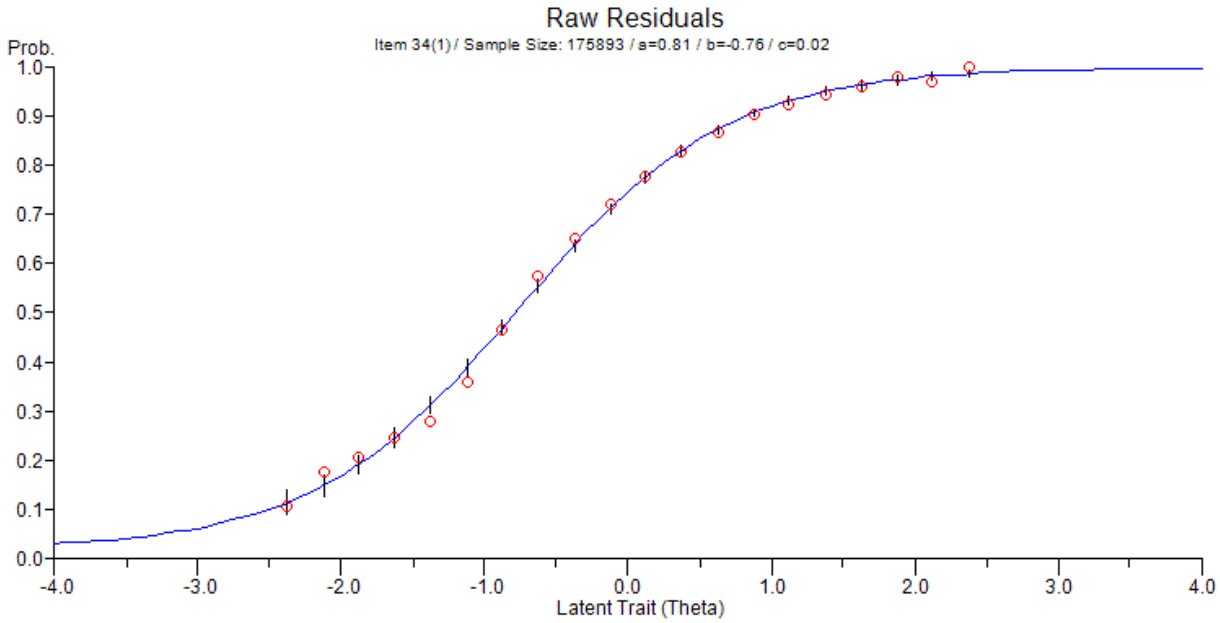
Item (SEQ) 25



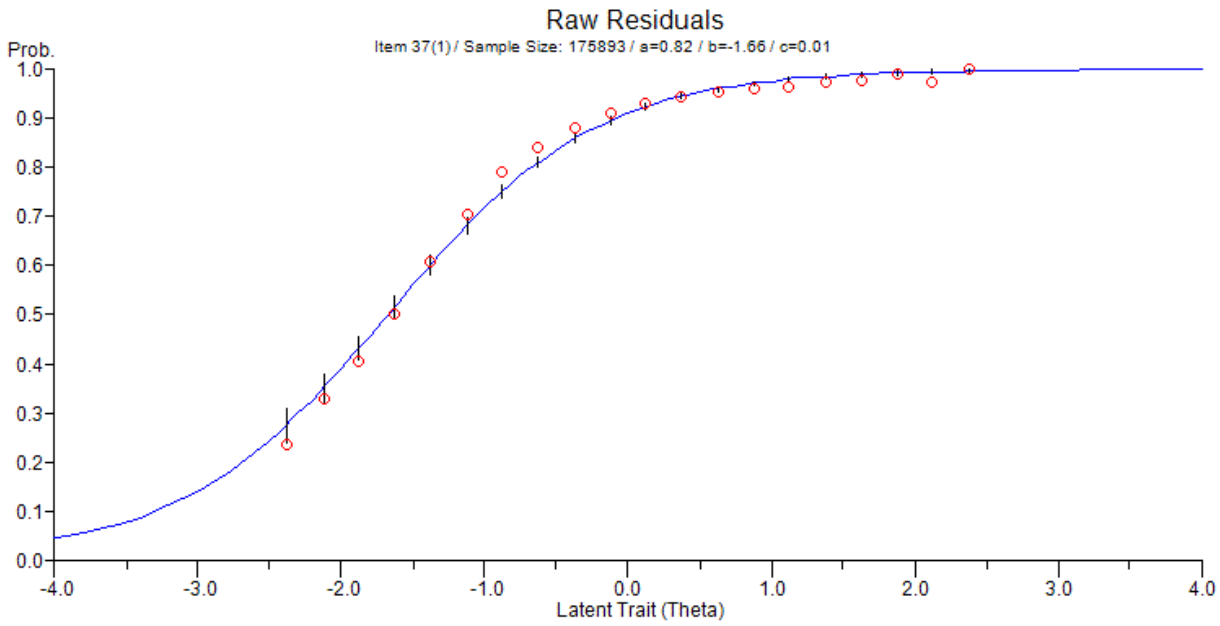
Item (SEQ) 32



Item (SEQ) 42



Item (SEQ) 45



III. Replication of the 2011 FCAT 2.0 Grade 8 Math Assessment

In this chapter, we summarize the results of all analyses related to the item analyses, calibration, scaling, and equipercentile equating of the Grade 8 Math FCAT 2.0. We begin with a description of how we processed the data and finish with our conclusion regarding the degree to which our results and Pearson's converged.

Calibration Sample and Demographic Variables

The initial calibration file we downloaded from the sFTP site contained 178,793 students. We first excluded students who had a school type of 10, 11, 14, or 99 and those who used large print or Braille. From the excluded sample, we selected only students with reportable scores (Score Flag = 1 in the data file). Using the exclusion rules, we were able to create the same calibration sample reported by Pearson (N = 176,627). The demographics for the calibration sample, shown in Table III-1, were also identical to that reported by Pearson.

Table III-1

Demographics for Grade 8 Math Calibration Sample

		Pearson	SPS
Gender	Female	86,651	86,651
	Male	89,822	89,822
	Unknown	154	154
Ethnicity	Asian	4,591	4,591
	Black	39,482	39,482
	Hispanic	48,772	48,772
	American Indian/Alaskan	751	751
	Multiracial	4,712	4,712
	Native Hawaiian/Pacific	185	185
	Unknown	432	432
	White	77,702	77,702

Rescoring Item Responses and Flagging Items via Classical Item Statistics

Once the calibration sample was created, we next rescored the raw core item responses and compared them to the scored item responses provided in the data file. There were 48 core items, 31 multiple-choice and 17 gridded items, all dichotomously scored. The raw item responses were rescored using the answer key provided in the test map file. A correct response was given a 1 while an incorrect response was given a 0. Each rescored core item was identical to the scored items in the file provided by Pearson.

We next used classical item statistics to flag potentially problematic items. Items with the following characteristics (determined by Pearson) were flagged:

- Classical item discrimination ($r_{\text{pbi-c}}$) was less than 0.2,
- Classical item difficulty (*p*-value) was greater than 0.9 or less than 0.15,
- An incorrect option was selected by more than 40% of the sample,
- The *p*-value on any one form differed from the overall *p*-value by more than $\left|0.08\right|$.

Using the above criteria, one item was flagged (item 44, CID: 100000083729) because the *p*-value on one of the forms differed from the overall *p*-value by 0.084. Although this item was not reported by Pearson as being flagged using the above criteria, the item appeared to be functioning appropriately (e.g., reasonable item discrimination) and we agree that it should have been included in the following analyses. Pearson reported one item that was flagged (item 46, CID: 100000084467) apparently because more than 40% of the examinees responded to one of the incorrect options on two forms. Because the overall percentage of students who chose the incorrect option was less than 40%, we did not flag this item. Furthermore, we agree that this item should have been included in further analyses.

Item Parameter Calibration

Item parameter calibration was conducted using the computer program MULTILOG (Thissen, 2003) on the calibration sample. The three-parameter logistic model (3PLM) was used for the multiple-choice items and the two-parameter logistic model (2PLM) was used for the gridded items. A prior for the *c* parameter, which is based on the normal distribution with a mean of -1.4 and a standard deviation of 1 on the logit metric, was implemented only for the *c* parameter in the 3PLM. Sample MULTILOG code is provided in Appendix III-A.

MULTILOG successfully converged within 24 EM cycles. The item parameter estimates provide by MULTILOG were transformed onto the logistic metric so that we could compare them to the estimates reported by Pearson. The item parameter estimates were transformed as follows:

- For 2PLM

$$a_{\text{New}} = \frac{a_{\text{MLG}}}{1.7}$$

$$b_{\text{New}} = b_{\text{MLG}}$$

- For 3PLM

$$a_{\text{New}} = \frac{a_{\text{MLG}}}{1.7},$$

$$b_{\text{New}} = \frac{-b_{\text{MLG}}}{a_{\text{MLG}}}$$

and

$$c_{\text{New}} = \frac{\exp\left[c_{\text{MLG}}\right]}{1+\exp\left[c_{\text{MLG}}\right]}.$$

All of the transformed item parameter estimates and their corresponding standard errors were reasonable values. Furthermore, the item parameter estimates were identical to those reported by Pearson.

Items were flagged for detailed inspection using the following criteria provided by Pearson and FDOE: $a < 0.5$, $2.0 < b < -2.0$, or, for multiple-choice items, $c < .05$. Table III-2 reports the flagged items and the reason for being flagged.

Table III-2

Items Flagged Given Above Criteria

Item	Reason	*a*	*b*	*c*	Model Fit
5	$c < 0.05$	0.51	-1.31	0.01	Acceptable
9	$a < 0.5$, $c < 0.05$	0.45	0.44	0.04	Good
19	$a < 0.5$	0.46	0.51	0.15	Good
24	$c < 0.05$	0.67	-1.28	0.01	Good
25	$c < 0.05$	0.77	-1.46	0.04	Acceptable

The item parameter calibration and model fit of the flagged items were further inspected to determine if they should be excluded from additional analyses. Model fit was examined via an inspection of raw residuals around the item characteristic curve (ICC) that is defined by the item parameter estimates. The computer program ResidPlots (Liang, Han, & Hambleton, 2008) was used to examine model fit. Reasonable or acceptable model fit occurs when the majority of the observed proportions are randomly distributed around the ICC, with very few observed points falling far from the ICC. For each flagged item, the item parameter estimation and model fit was acceptable in that the observed proportions were close to the expected value given by the ICC. III- B contains model fit plots for each flagged item.

In addition to inspecting the model fit for the flagged items, we examined the model fit for all items. The model exhibited excellent fit for most of the items, and the fit was acceptable

for all items. Therefore, given all of the analyses to this point, we agree that all of the core items should be included in the equipercentile equating.

Scale Scores and Equipercentile Equating

Before estimating the scale scores for each student, the new item parameter estimates (a_{New}, b_{New} and c_{New}) were transformed onto the FCAT score scale as follows:

$$a_{SS} = \frac{a_{New}}{50},$$

$$b_{SS} = b_{New} * 50 + 300$$

and

$$c_{SS} = c_{New}$$

The newly transformed item parameters (a_{SS}, b_{SS}, and c_{SS}) were nearly identical to those reported by Pearson (barring rounding error) and were used to estimate the students' IRT ability estimate. The computer program IRT Score Estimation (Chien, Hsu, and Shin, 2011) was used to perform maximum likelihood estimation. Table III-3 reports the descriptive statistics for the raw scores and scale scores.

Table III-3

Descriptive Statistics for Proficiency Estimates

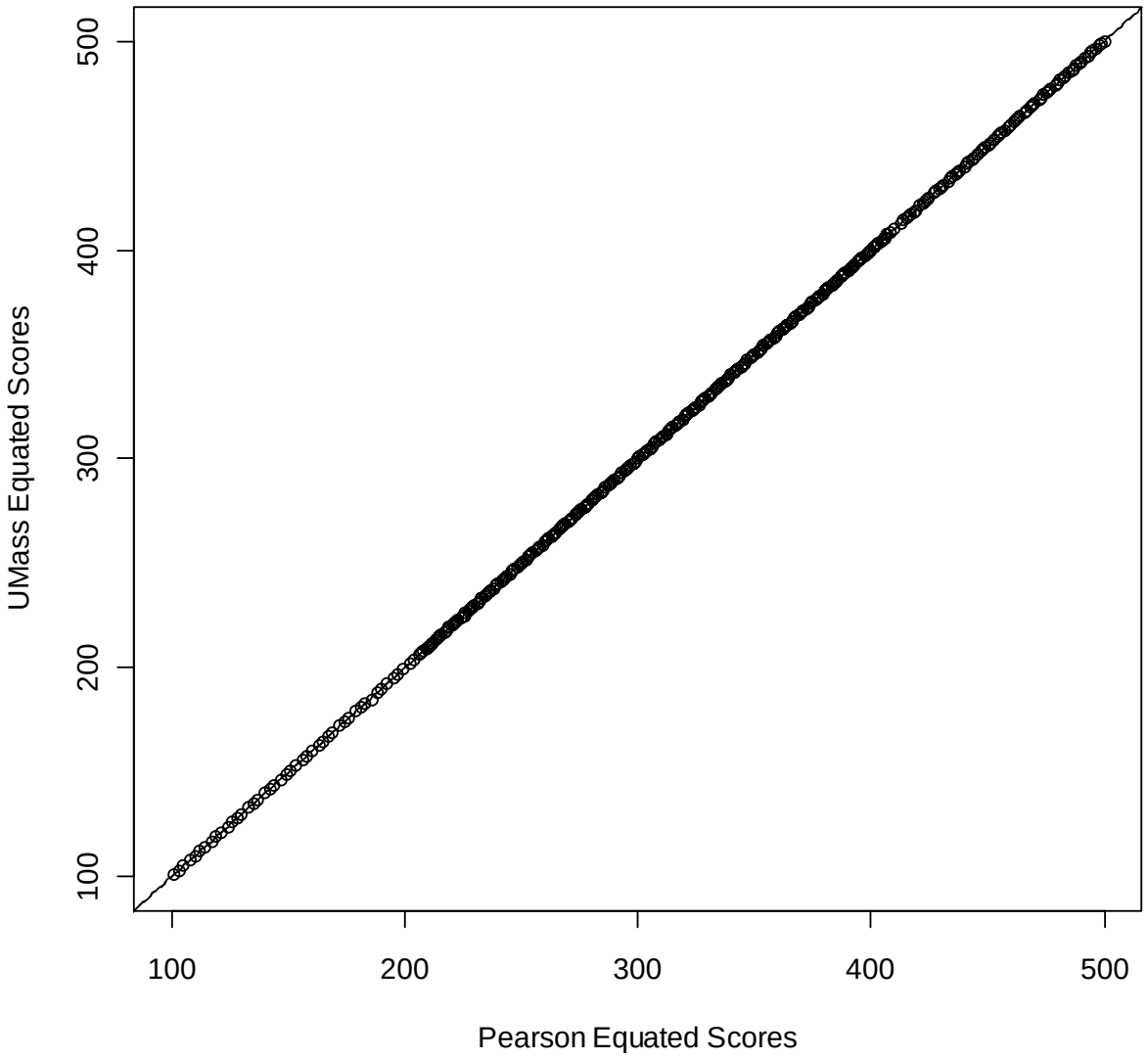
	Raw Score	IRT Scale Score
Mean	26.84	299.32
Standard Deviation	10.51	56.44
Kurtosis	0.92	1.29
Skewness	0.04	-0.24

The equipercentile equating was conducted to adjust the 2011 scale score distribution such that it was equivalent to the 2010 distribution using the computer program RAGE-RGEQUATE (Zeng, Kolen, Hanson, Cui & Chien, 2005). The frequency distribution for the

scale scores was created and used with the scale score frequency distribution from 2010 to create a conversion table linking the 2011 scale scores to the 2010 scale scores for the Grade 8 Math test. The operational postsmooth limit was used in the equating. The postsmooth lower limit was based on the following equation: (percent of examinees who received a 100 scale score + 0.5)/100. The postsmooth upper limit was based on the following equation: (percent of examinees who received a 500 scale score + 0.5)/100. In this case, the lower and upper limit was 0.012 and 0.010, respectively. Therefore, the equipercentile was run twice - once for each limit. The final conversion table was a combination of the two equatings - equated scale scores for 100 to 300 were based on the postsmooth lower limit value and the equated scale scores for 300 to 500 were based on the postsmooth upper limit.

The final conversion table was nearly identical to that constructed by Pearson. Our equated scores correlated 0.9999 with Pearson's reported equated scale scores and differed by no more than 1 point for any equated score. The proportion of exact agreement exceeded 99%. To illustrate the high level of agreement, Figure III-1 compares the Pearson and SPS equated scale scores from the conversion table to the identity line. If our scale scores are in agreement, then the points should fall directly on the identity line, which represents exact agreement. The solid line in Figure III-1 represents the identity line and runs perfectly through the majority of the equated scores. Although a few scale scores differed by one point, this was likely due to rounding error in the item parameter estimates used in the IRT scoring program and in rounding the scale scores. Because the conversion table was nearly identical, we agree with Pearson's final conversion table.

Figure III-1. Plot comparing Pearson and SPS equated scale scores.



After creating the conversion table, we transformed the 2011 scale scores onto the 2010 scale. The descriptive statistics for the equated scores, shown in Table III-4, were identical (to the second decimal place) to those reported by Pearson.

Table III-4

Descriptive Statistics for Equated Proficiency Estimates

	Pearson	SPS
Mean	324.55	324.55
Standard Deviation	45.24	45.24
Kurtosis	4.53	4.53
Skewness	-0.83	-0.83

Conclusion

In summary, we were able to successfully replicate Pearson's operational procedures and results including creating the calibration sample given the exclusion/inclusion rules, scoring the raw item responses, verifying the quality of the items (item statistics and model fit), reproducing identical item parameter estimates, and the (nearly) identical conversion table and scale score distribution on the 2001 FCAT 2.0 Grade 8 Mathematics Assessment. Given this successful replication, we feel confident that the operational procedures were conducted correctly.

Appendix III-A

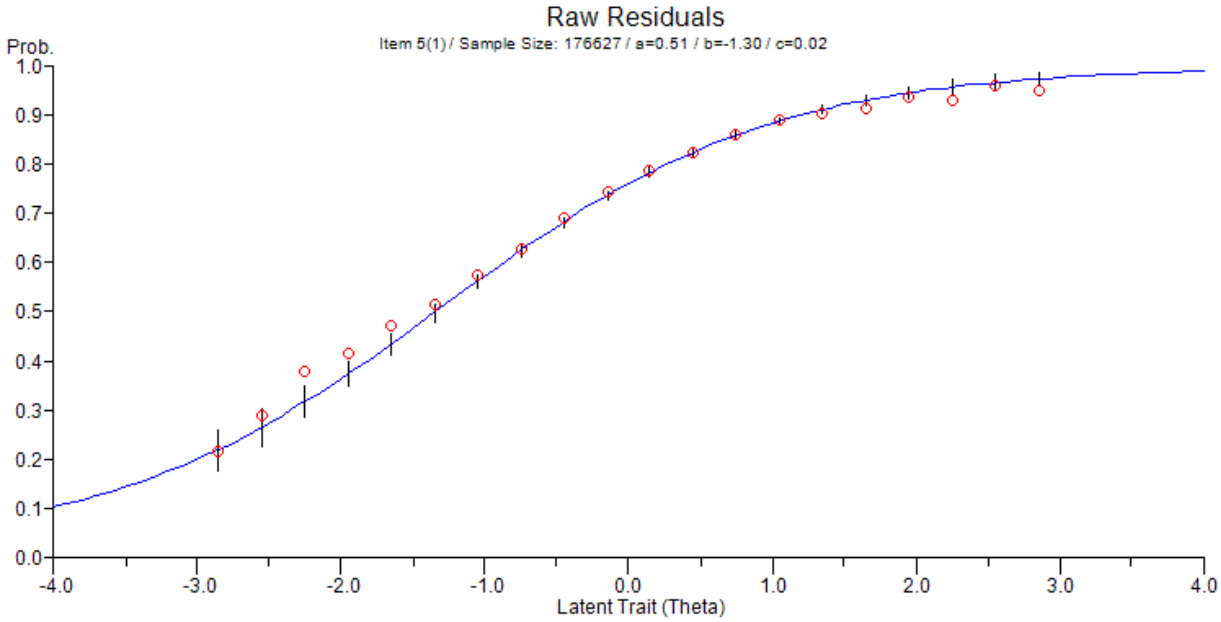
Sample MULTILOG Code for Grade 8 Math

```
FCAT 2.0 2011
Grade 8 Math
>PRO RAN,IND,NIT=48,NEX=176627,NGR=1,DAT='G8Math.DAT';
>TES ITE=(1(1)31),L3;
>TES ITE=(32(1)48),L2;
>PRIORS ITE=(1(1)31),CJ,PAR=(-1.4,1.0);
>EST NCY=120;
>SAVE;
>END;
2
01
1111111111111111111111111111111111111111111111
N
(48A1)
```

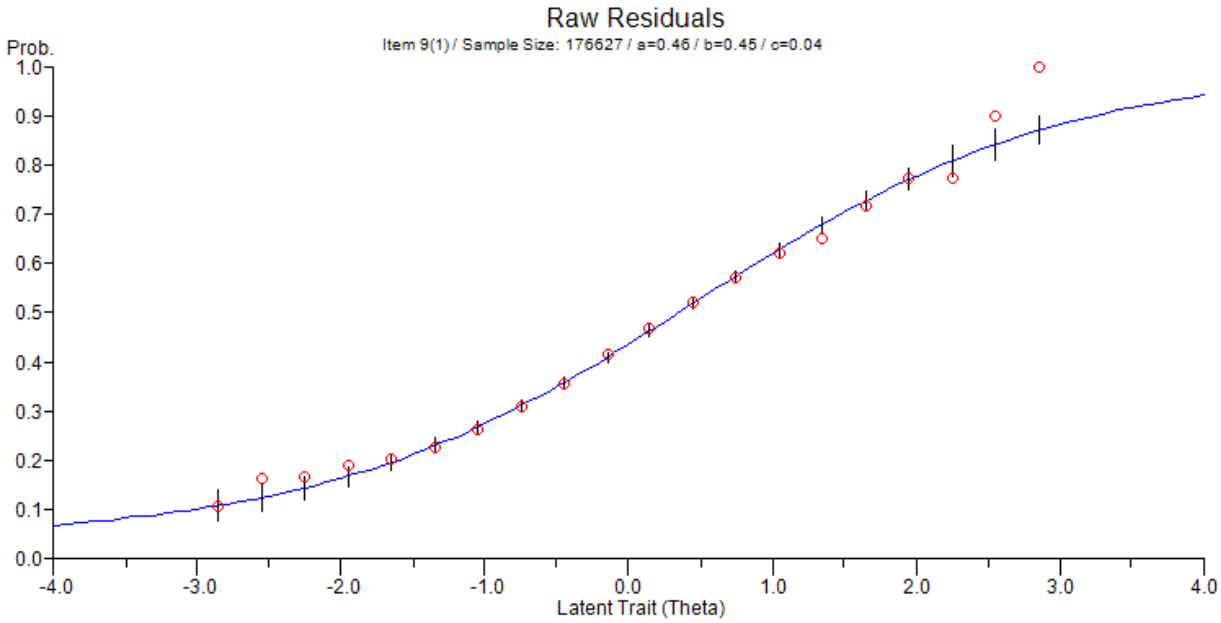
Appendix III-B

Model Fit Plots for Items Flagged During Item Calibration Inspection

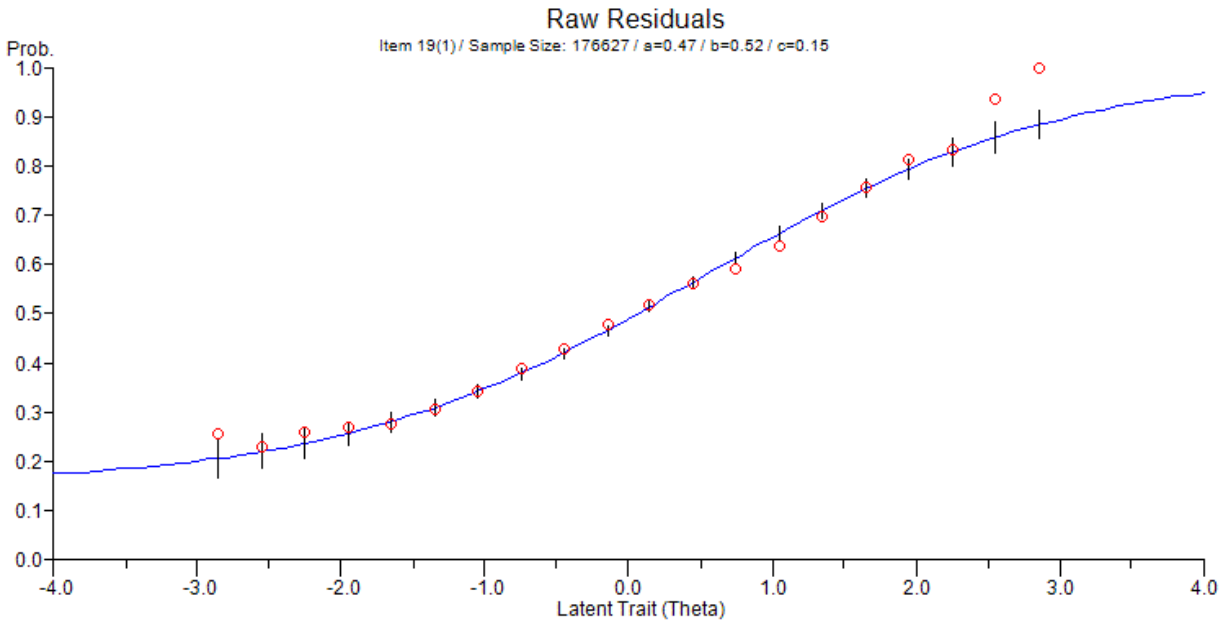
Item 5



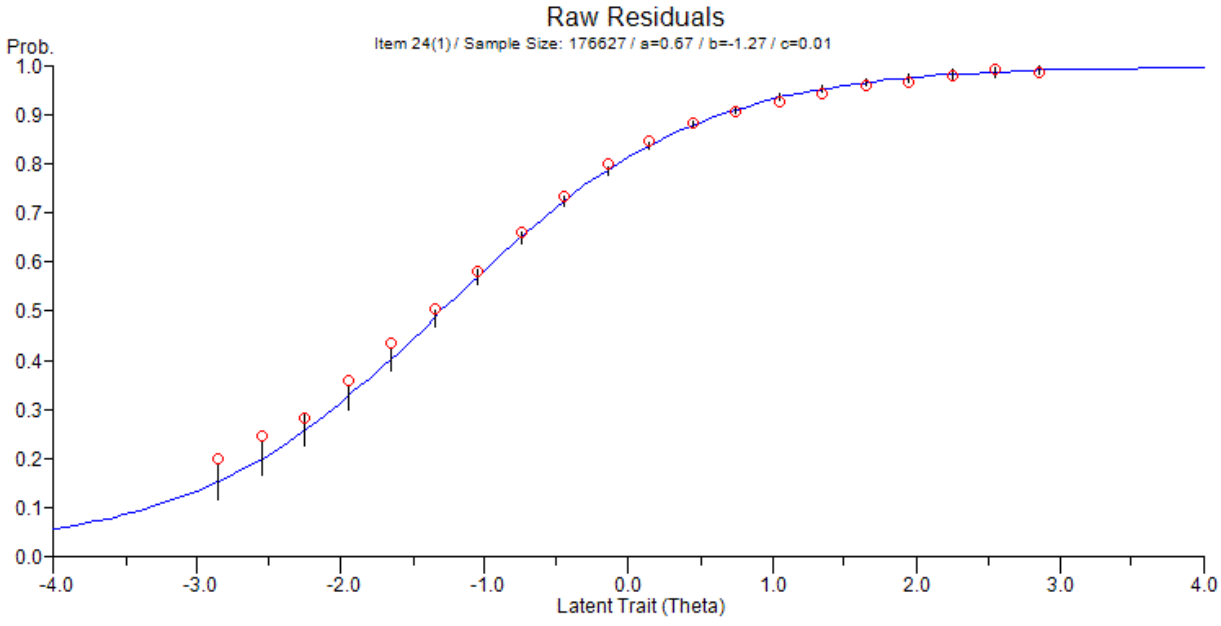
Item 9



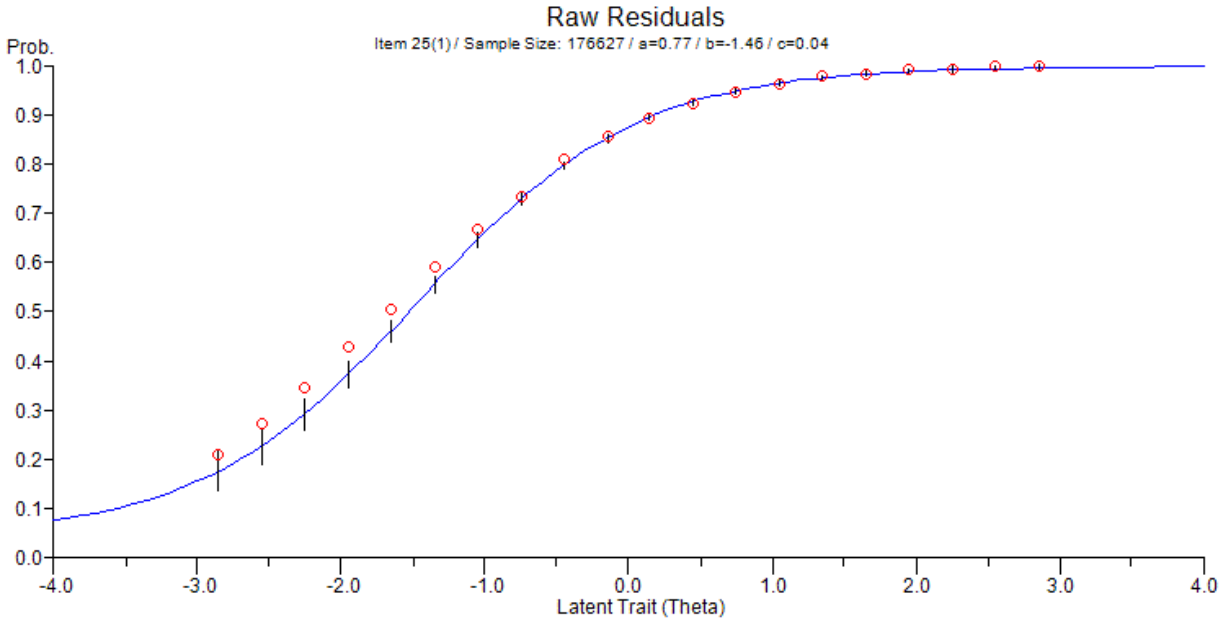
Item 19



Item 24



Item 25


Raw Residuals
Item 25(1) / Sample Size: 176627 / a=0.77 / b=-1.46 / c=0.04
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

IV. Replication of the 2011 Grade 8 Science Assessment

In this chapter we describe the analyses and report on the results for the Grade 8 Science assessment. Our descriptions begin with our receipt and cleaning of the data and end with our conclusions regarding the degree to which our results were congruent with those of Pearson Assessment.

Calibration Sample and Demographic Variables

The initial calibration file we downloaded from the sFTP site contained 180,472 students. We first excluded students who had a school type of 10, 11, 14, or 99 and those who used large print or Braille. From the excluded sample, we selected only students with reportable scores (Score Flag = 1 in the data file), who were part of the standard curriculum, and who took one of the four anchor forms (forms 30, 31, 32, and 33). Using the exclusion/inclusion rules, we were able to create the same calibration sample reported by Pearson (N = 18,687). The demographics for the calibration sample, shown in Table IV-1, were also identical to that reported by Pearson.

Table IV-1

Demographics for Grade 10 Reading Calibration Sample

		Pearson	SPS
Gender	Female	9,621	9,621
	Male	9,052	9,052
	Unknown	14	14
Ethnicity	Asian	546	546
	Black	4,077	4,077
	Hispanic	4,812	4,812
	American Indian/Alaskan	74	74
	Multiracial	510	510
	Native Hawaiian/Pacific	15	15
	Unknown	45	45
	White	8,608	8,608

Rescoring Item Responses and Flagging Items via Classical Item Statistics

Once the calibration sample was created, we next rescored the raw core and anchor item responses and compared them to the scored item responses provided in the data file. Each anchor form was comprised of 51 core items, 47 multiple-choice and 4 gridded-response items, all dichotomously scored. Each anchor form had 7 or 8 anchor items with a total of 30 anchor items, 27 multiple-choice and 3 gridded-response items. The raw item responses for the core and anchor items were rescored using the answer key provided in the test map file. A correct response was given a 1 while an incorrect response was given a 0. Each rescored core and anchor item was identical to the scored items in the file provided by Pearson.

We next used classical item statistics to flag potentially problematic items. Items with the following characteristics (determined by Pearson) were flagged:

- Classical item discrimination ($r_{\text{pbi-c}}$) was less than 0.2,
- Classical item difficulty (*p*-value) was greater than 0.9 or less than 0.15,
- An incorrect option was selected by more than 40% of the sample,
- The *p*-value on any one form differed from the overall *p*-value by more than $\left|0.08\right|$.

Using the above criteria, one core item was flagged (Item 56) because the proportion of examinees who selected option D was greater than 0.40 (0.42). Pearson also flagged item 56 for the same reason. No other items were flagged. After further investigation, we agreed that item 56 should have been included in further analyses.

Item Parameter Calibration

Item parameter calibration was conducted using the computer program MULTILOG (Thissen, 2003) on the calibration sample and anchor forms. The three-parameter logistic model (3PLM) was used for the multiple-choice items and the two-parameter logistic model (2PLM)

was used for the gridded items. A prior for the *c* parameter, which is based on the normal distribution with a mean of -1.4 and a standard deviation of 1 on the logit metric, was implemented only for the *c* parameter in the 3PLM. Sample MULTILOG code is provided in Appendix IV-A.

MULTILOG successfully converged within 60 EM cycles. The item parameter estimates provide by MULTILOG were transformed onto the logistic metric so that we could compare them to the estimates reported by Pearson. The item parameter estimates were transformed as follows:

- For 2PLM

$$a_{\text{New}} = \frac{a_{\text{MLG}}}{1.7}$$

$$b_{\text{New}} = b_{\text{MLG}}$$

- For 3PLM

$$a_{\text{New}} = \frac{a_{\text{MLG}}}{1.7},$$

$$b_{\text{New}} = \frac{-b_{\text{MLG}}}{a_{\text{MLG}}}$$

and

$$c_{\text{New}} = \frac{\exp\left[c_{\text{MLG}}\right]}{1+\exp\left[c_{\text{MLG}}\right]}.$$

All of the transformed item parameter estimates and their corresponding standard errors were reasonable values. Furthermore, the item parameter estimates were essentially identical to those reported by Pearson.

Items were flagged for detailed inspection using the following criteria provided by Pearson and FDOE: $a < 0.5$, $2.0 < b < -2.0$, or $c < .05$. Table IV-2 reports the flagged items and the reason for being flagged.

Table IV-2

Items Flagged Given Above Criteria

Item	Reason	*a*	*b*	*c*	Model Fit
12 (core)	$c < 0.05$	0.75	-1.01	0.03	Good
14 (core)	$c < 0.05$	0.98	-1.33	0.04	Good
36 (core)	$a < 0.50$	0.47	-0.60	0.10	Good
60 (anchor)	$a < 0.50$	0.49	0.26	0.14	Good

The item parameter calibration and model fit of the flagged items were further inspected to determine if they should be excluded from additional analyses. Model fit was examined via an inspection of raw residuals around the item characteristic curve (ICC) that is defined by the item parameter estimates. The computer program ResidPlots (Liang, Han, & Hambleton, 2008) was used to examine model fit. Reasonable or acceptable model fit occurs when the majority of the observed proportions are randomly distributed around the ICC, with very few observed points falling far from the ICC. For each flagged item, the item parameter estimation and model fit was good in that the observed proportions were close to the expected value given by the ICC (Appendix IV-B contains model fit plots for each flagged item). Therefore, given that there were no key check issues with these items and that the item statistics and model fit was acceptable, we agreed these items should *not* be excluded from further analyses.

In addition to inspecting the model fit for the flagged items, we examined the model fit for all items. The model exhibited excellent to acceptable fit for all of the items. Model fit is particularly important for the anchor items since poor fit can negatively influence linking the

2011 scale to the base scale, which can in turn influence performance classification. Therefore, given all of the analyses to this point, we agreed that no item should be excluded due to poor fit.

Anchor Stability Check

Before performing the final linking between the 2011 and base scale, we examined anchor item stability to determine if any of the anchor items were exhibiting differential performance (i.e., drifting from their initial item parameter values). We used a measure that summarizes the difference between the item characteristic curves (ICCs) given the 2011 and base scale item parameter estimates. The statistic, denoted D^2, is a weighted sum of the differences in the ICCs and was computed as follows:

$$D^2 = \sum_{q=1}^{40} w_q \left(\hat{P}_q^{2011} - \hat{P}_q^{Base} \right)^2$$

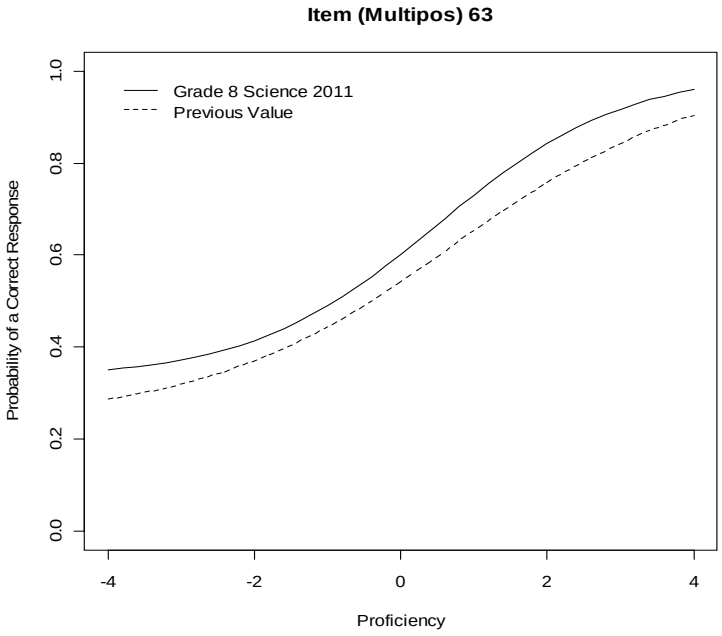
w_q represents the normalized weight at quadrature point q and is based on the height of the standard normal distribution. $\hat{P}_q^{2011}$ and $\hat{P}_q^{Base}$ represent the probability of a correct response at quadrature point q.

The following iterative procedure was used to flag potentially problematic anchor items. First, the item parameter estimates for 2011 were first placed onto the base scale using the Stocking and Lord (1983) procedure. Once the item parameter estimates were on the same scale, D^2 was computed for each anchor item. Items that had a D^2 beyond three standard deviations from the mean were initially flagged as drift. One anchor item ("Multipos 70") exhibited a D^2 that was 3.11 standard deviations above the mean. Second, the original item parameter estimates for 2011 were again placed onto the base scale using the Stocking and Lord method, however, excluding the flagged item from the linking. D^2 was computed for all items using the newly transformed item parameter estimates. Two anchor items were flagged as exhibiting drift: items

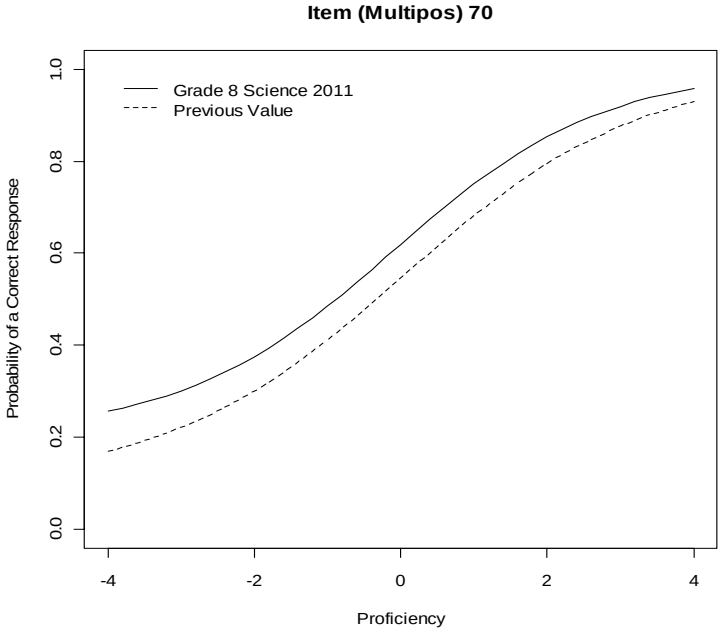
63 and 70. A third iteration was performed, but no additional items were flagged. Figure 1(a) and 1(b) show the 2011 and base scale ICCs for the flagged items. It was apparent that both items were easier in 2011.

Figure IV-1. ICCs for items flagged as drifting.

(a)



(b)



Pearson flagged both of these items and excluded them from the final linking. We agree that both of the items should be excluded from the final linking.

Stocking and Lord Transformation and Scale Scores

Before estimating the scale scores for each student, the 2011 item parameter estimates (a_{2011}, b_{2011} and c_{2011}) were transformed onto Pearson's original item parameter estimate scale so that we could compare the estimates to those reported by Pearson to determine if the item parameter calibration was successful. The item parameter estimates were transformed as follows:

$$a_{SS} = \frac{a_{2011}}{50},$$

$$b_{SS} = b_{2011} * 50 + 300$$

and

$$c_{SS} = c_{2011}.$$

The newly transformed item parameters (a_{SS}, b_{SS}, and c_{SS}) were nearly identical to those reported by Pearson, supporting their item parameter calibration results.

The item parameter estimates were then placed onto the based scale using the Stocking and Lord constants, which were computed using the computer program STUIRT (Kim & Kolen, 2004). The Stocking and Lord constants were nearly identical to those reported by Pearson (SPS: slope = 55.91, intercept = 323.79; Pearson: slope = 55.92, intercept = 323.80). The item parameter estimates were placed onto the based score scale as follows:

$$a_{SS} = \frac{a_{2011}}{55.91},$$

$$b_{SS} = b_{2011} * 55.91 + 323.79$$

and

$$c_{SS} = c_{2011}.$$

Once the 2011 item parameter estimates were placed onto the base score scale, we estimated the examinees' ability via maximum likelihood estimation using the computer program IRT Score Estimation (Chien, Hsu, and Shin, 2011). Table IV-3 reports the descriptive statistics for the scale score distribution for each anchor form. As can be seen, we were able to reproduce the mean and standard deviation of the proficiency distribution.

Table IV-3

Descriptive Statistics for Proficiency Estimates on Each Anchor Form

		SPS		Pearson	
Form	Sample Size	Mean	SD	Mean	SD
30	5,399	316.45	66.42	316.46	65.77
31	5,383	316.12	66.20	316.13	65.65
32	5,322	317.22	66.65	317.23	65.14
33	5,312	317.16	66.28	317.16	66.36

Conclusion

In summary, SPS was able to successfully replicate Pearson's operational procedures and results including creating the calibration sample given the exclusion/inclusion rules, scoring the raw item responses, verifying the quality of the items (item statistics and model fit), reproducing identical item parameter estimates, flagging the same items in the anchor stability check and the (nearly) identical scale score distribution on the Grade 8 Science Assessment. Given this successful replication, we feel confident that the operational procedures were conducted correctly.

Appendix IV-A

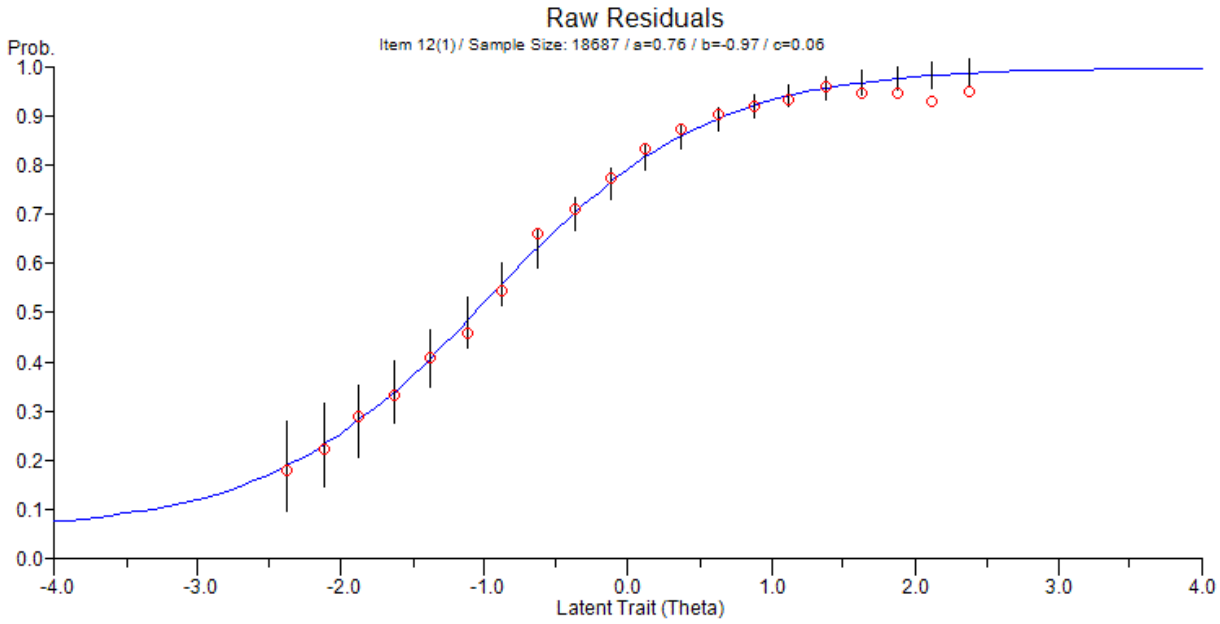
Sample MULTILOG Code for Grade 8 Science

```
FCAT 2011
Grade 8 Science
>PRO RAN,IND,NIT=81,NEX=18687,NGR=1,DAT='G8Science.DAT';
>TES ITE=(1(1)74),L3;
>TES ITE=(75(1)81),L2;
>PRIORS ITE=(1(1)74),CJ,PAR=(-1.4,1.0);
>EST NCY=120;
>SAVE;
>END;
2
01
11111111111111111111111111111111111111111111111111111111111111111111
11111111111111111
N
(81A1)
```

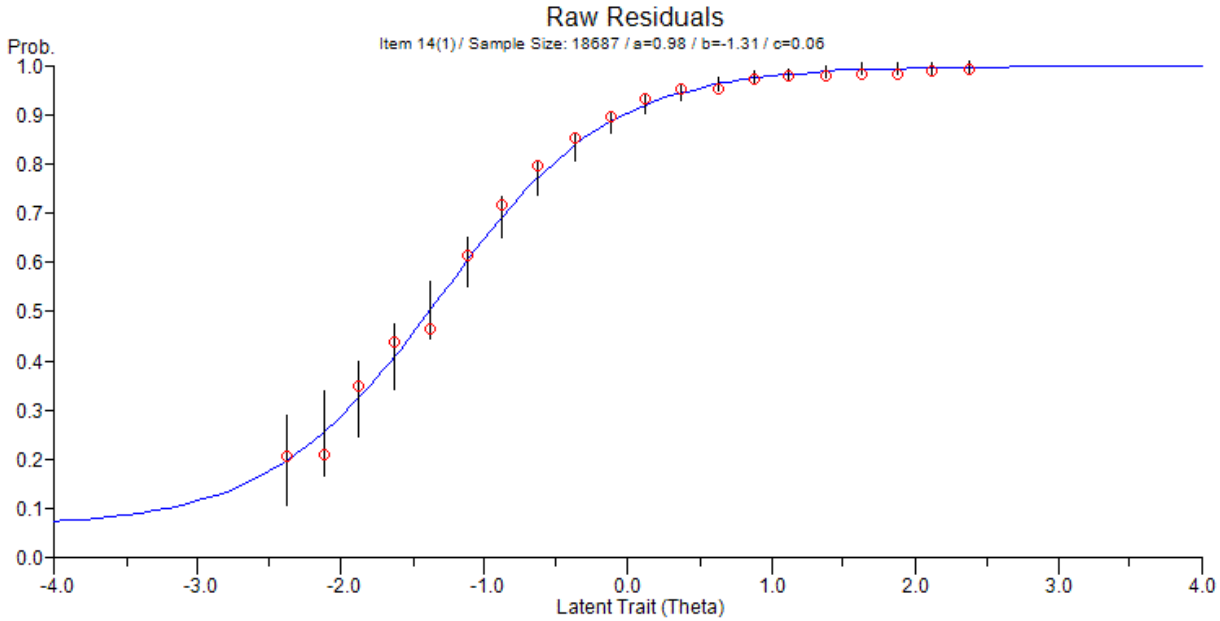
Appendix IV-B

Model Fit Plots for Items Flagged During Item Calibration Inspection

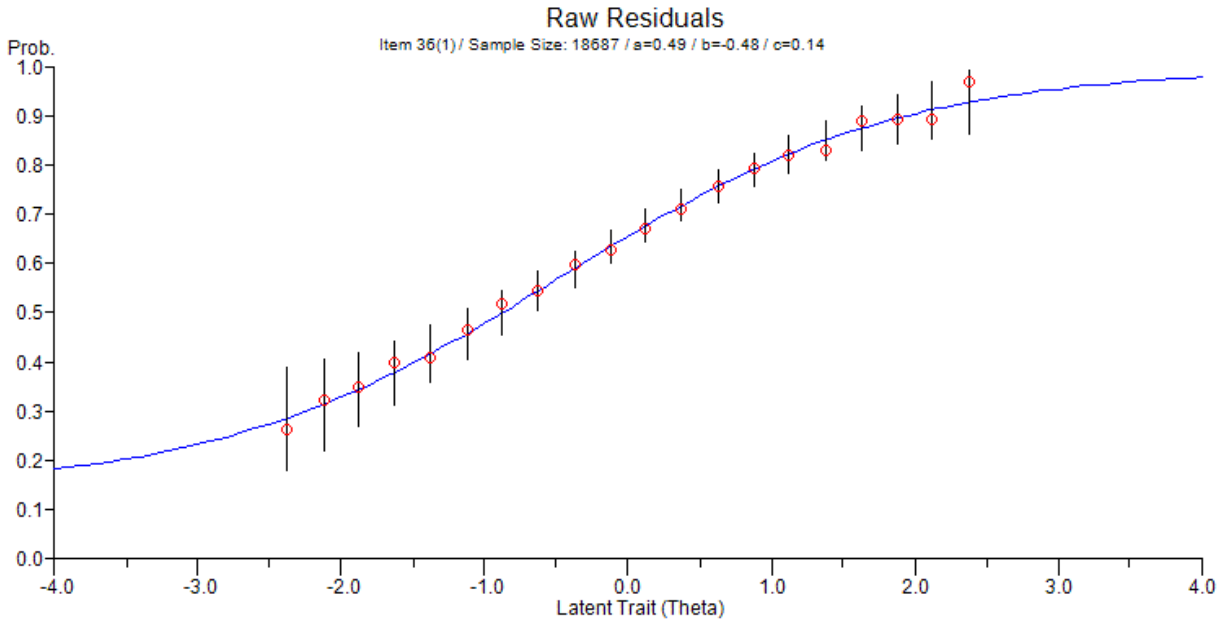
Item 12



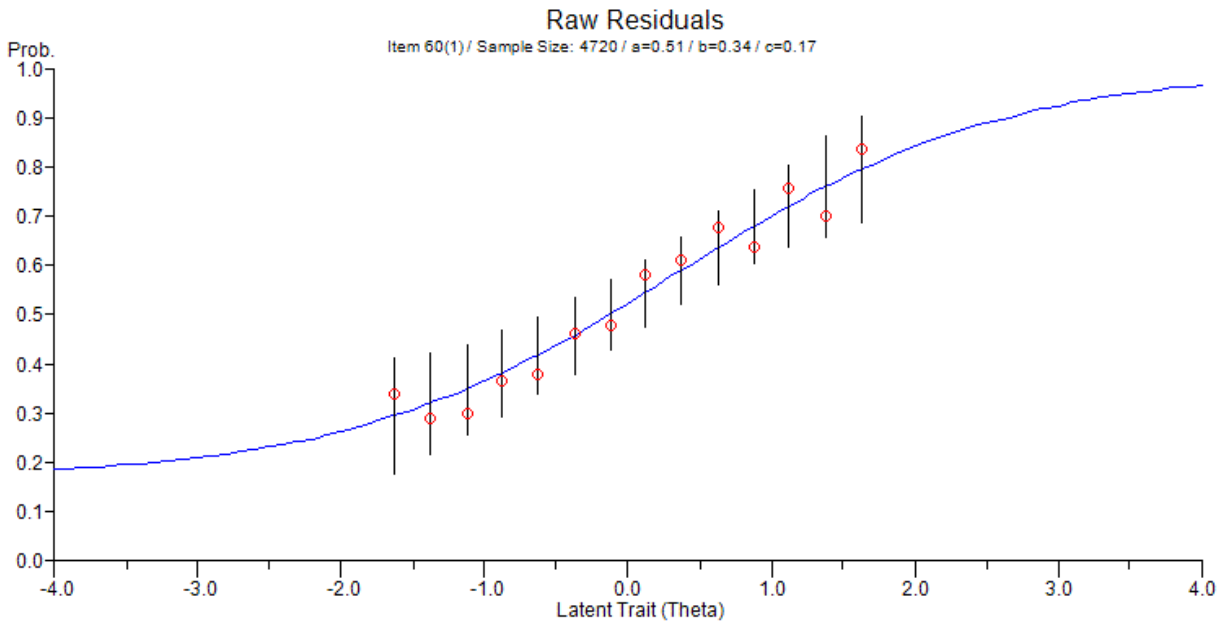
Item 14



Item 36



Item 60



V. Replication of the 2011 FCAT 2.0 Grade 10 Reading Assessment

In this chapter, we present and discuss the results of all analyses conducted on the Grade 10 Reading assessment. Like the previous chapters, we begin with processing of the data and end with our conclusions regarding how well our results compared with Pearson's results.

Calibration Sample and Demographic Variables

The initial calibration file we downloaded from the sFTP site contained 193,568 students. We first excluded students who had a school type of 10, 11, 14, or 99 and those who used large print or Braille. From the excluded sample, we selected only students with reportable scores (Score Flag = 1 in the data file). Using the exclusion rules, we were able to create the same calibration sample reported by Pearson (N = 171,246). The demographics for the calibration sample, shown in Table V-1, were also identical to that reported by Pearson.

Table V-1

Demographics for Grade 10 Reading Calibration Sample

		Pearson	SPS
Gender	Female	85,513	85,513
	Male	85,461	85,461
	Unknown	272	272
Ethnicity	Asian	4,564	4,564
	Black	37,172	37,172
	Hispanic	46,164	46,164
	American Indian/Alaskan	691	691
	Multiracial	4,682	4,682
	Native Hawaiian/Pacific	186	186
	Unknown	578	578
	White	77,254	77,254

Rescoring Item Responses and Flagging Items via Classical Item Statistics

Once the calibration sample was created, we next rescored the raw core item responses and compared them to the scored item responses provided in the data file. There were 45 core items, all dichotomously-scored, multiple-choice items. The raw item responses were rescored using the answer key provided in the test map file. A correct response was given a 1 while an incorrect response was given a 0. Each rescored core item was identical to the scored items in the file provided by Pearson.

We next used classical item statistics to flag potentially problematic items. Items with the following characteristics (determined by Pearson) were flagged:

- Classical item discrimination ($r_{\text{pbi-c}}$) was less than 0.2,
- Classical item difficulty (*p*-value) was greater than 0.9 or less than 0.15,
- An incorrect option was selected by more than 40% of the sample,
- The *p*-value on any one form differed from the overall *p*-value by more than $\left|0.08\right|$.

Using the above criteria, one item was flagged (SEQ Item 37) because the proportion of examinees who selected option D was greater than 0.40 (0.45). Pearson also flagged item 37 for the same reason. In addition, Pearson flagged item (SEQ) 43 because more than 40% of the examinees responded to one of the incorrect options on a few of the forms; however, because the overall percentage of students who chose the incorrect option was less than 40%, we did not flag this item. Furthermore, we agree that both items should have been included in further analyses.

Item Parameter Calibration

Item parameter calibration was conducted using the computer program MULTILOG (Thissen, 2003) on the calibration sample. The three-parameter logistic model (3PLM) was used for the multiple-choice items and the two-parameter logistic model (2PLM) was used for the

gridded items. A prior for the *c* parameter, which is based on the normal distribution with a mean of -1.4 and a standard deviation of 1 on the logit metric, was implemented only for the *c* parameter in the 3PLM. Sample MULTILOG code is provided in Appendix V-A.

MULTILOG successfully converged within 39 EM cycles. The item parameter estimates provide by MULTILOG were transformed onto the logistic metric so that we could compare them to the estimates reported by Pearson. The item parameter estimates were transformed as follows:

- For 2PLM

$$a_{\text{New}} = \frac{a_{\text{MLG}}}{1.7}$$

$$b_{\text{New}} = b_{\text{MLG}}$$

- For 3PLM

$$a_{\text{New}} = \frac{a_{\text{MLG}}}{1.7},$$

$$b_{\text{New}} = \frac{-b_{\text{MLG}}}{a_{\text{MLG}}}$$

and

$$c_{\text{New}} = \frac{\exp\left[c_{\text{MLG}}\right]}{1+\exp\left[c_{\text{MLG}}\right]}.$$

All of the transformed item parameter estimates and their corresponding standard errors were reasonable values. Furthermore, the item parameter estimates were nearly identical to those reported by Pearson.

Items were flagged for detailed inspection using the following criteria provided by Pearson and FDOE: $a < 0.5$, $2.0 < b < -2.0$, or $c < .05$. Table V-2 reports the flagged items and the reason for being flagged.

Table V-2

Items Flagged Given Above Criteria

Item (SEQ)	Reason	*a*	*B*	*c*	Model Fit
1	$c < 0.05$	0.60	-1.60	0.02	Good
2	$c < 0.05$	0.67	-1.78	0.02	Good
3	$c < 0.05$	0.46	-1.77	0.02	Good
	$a < 0.5$				
4	$a < 0.5$	0.48	-1.20	0.09	Good
18	$c < 0.05$	0.66	-2.09	0.02	Good
	$b < -2$				
31	$a < 0.5$	0.44	0.57	0.31	Acceptable
34	$c < 0.05$	0.55	-1.62	0.03	Acceptable
36	$c < 0.05$	0.55	-1.30	0.04	Good
41	$c < 0.05$	0.65	0.10	0.03	Good

The item parameter calibration and model fit of the flagged items were further inspected to determine if they should be excluded from additional analyses. Model fit was examined via an inspection of raw residuals around the item characteristic curve (ICC) that is defined by the item parameter estimates. The computer program ResidPlots (Liang, Han, & Hambleton, 2008) was used to examine model fit. Reasonable or acceptable model fit occurs when the majority of the observed proportions are randomly distributed around the ICC, with very few observed points falling far from the ICC. For each flagged item, the item parameter estimation and model fit was acceptable in that the observed proportions were close to the expected value given by the ICC (Appendix V-B contains model fit plots for each flagged item). Therefore, given that there were no key check issues with these items and that the item statistics and model fit was acceptable, we agree that these items should be included in the equipercentile equating.

In addition to inspecting the model fit for the flagged items, we examined the model fit for all items. The model exhibited excellent to acceptable fit for most of the items. However, there were a few items that exhibited small to moderate magnitudes of misfit. For example, items 21, 40 and 52 exhibited a small magnitude misfit (see Figures V-1 to V-3 for the observed and expected ICC plots). However, including these items in the equipercentile equating will have a negligible effect and removing the items may have a detrimental effect on reliability. Therefore, given all of the analyses to this point, we agree that all of the core items should be included in the equipercentile equating.

Figure V-1. *Model fit plot for item (SEQ) 21.*

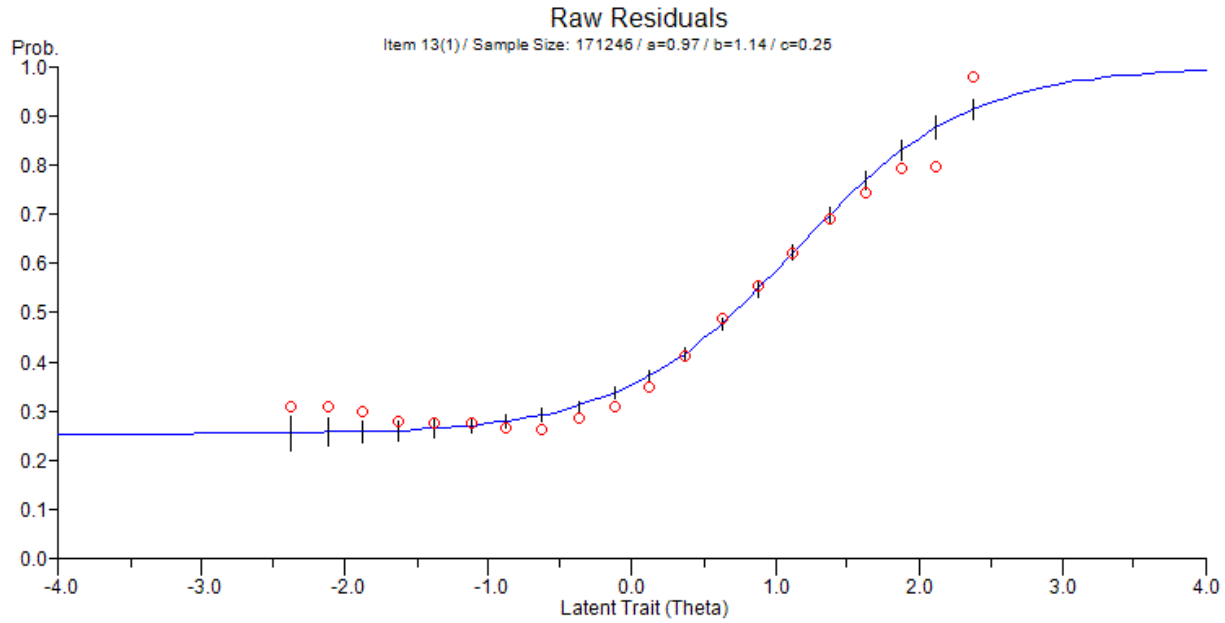


Figure V-2. *Model fit plot for item (SEQ) 40.*

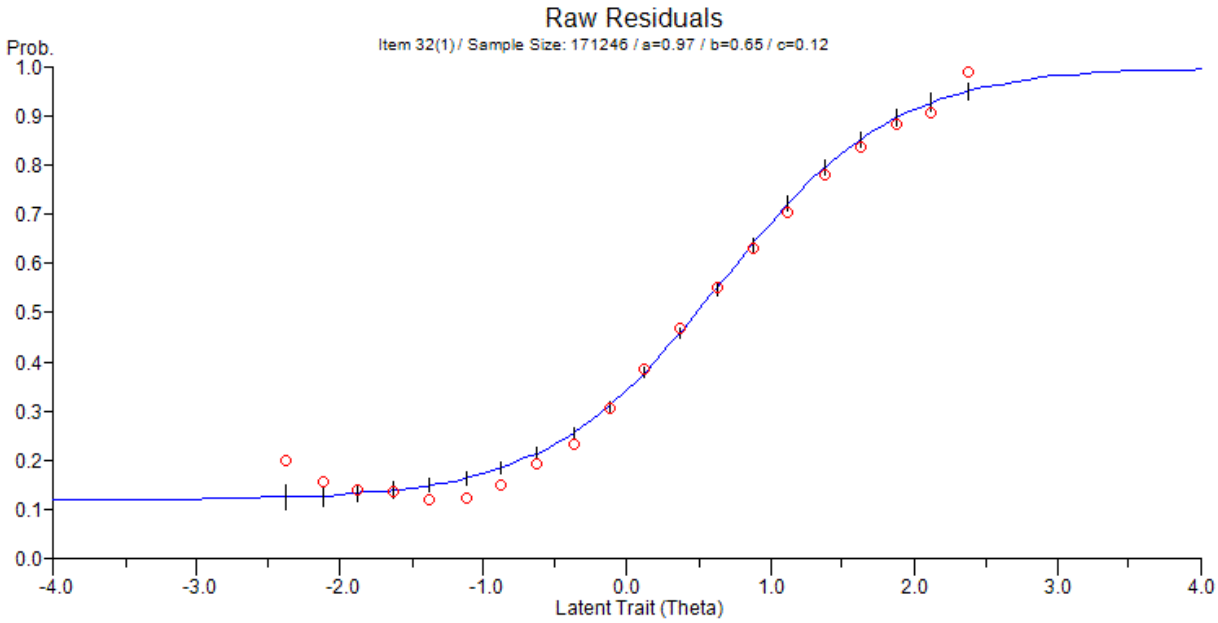
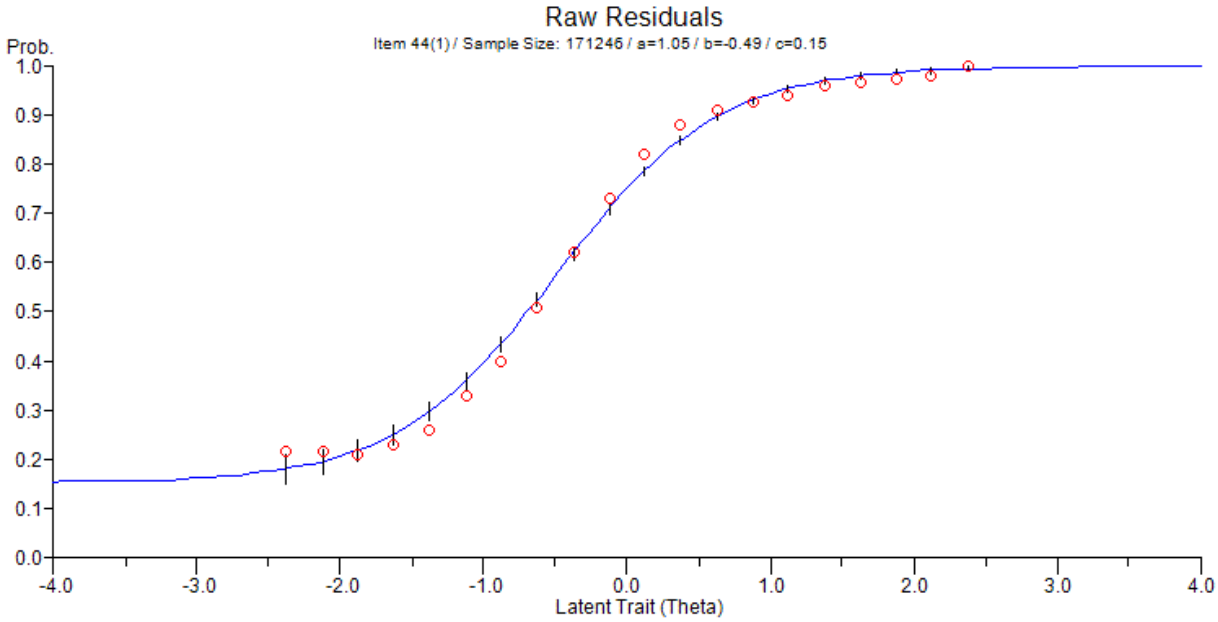


Figure V-3. *Model fit plot for item (SEQ) 52.*



Scale Scores and Equipercentile Equating

Before estimating the scale scores for each student, the new item parameter estimates (a_{New}, b_{New} and c_{New}) were transformed onto the FCAT scale as follows:

$$a_{SS} = \frac{a_{New}}{50},$$

$$b_{SS} = b_{New} * 50 + 300$$

and

$$c_{SS} = c_{New}.$$

The newly transformed item parameters (a_{SS}, b_{SS}, and c_{SS}) were nearly identical to those reported by Pearson (barring rounding error) and were used to estimate the students' IRT ability estimate. The computer program IRT Score Estimation (Chien, Hsu, & Shin, 2011) was used to perform maximum likelihood estimation. Table V-3 reports the descriptive statistics for the raw scores and scale scores.

Table V-3

Descriptive Statistics for Proficiency Estimates

	Raw Score	IRT Scale Score
Mean	28.07	300.40
Standard Deviation	8.64	56.14
Kurtosis	-0.82	0.80
Skewness	-0.14	-0.05

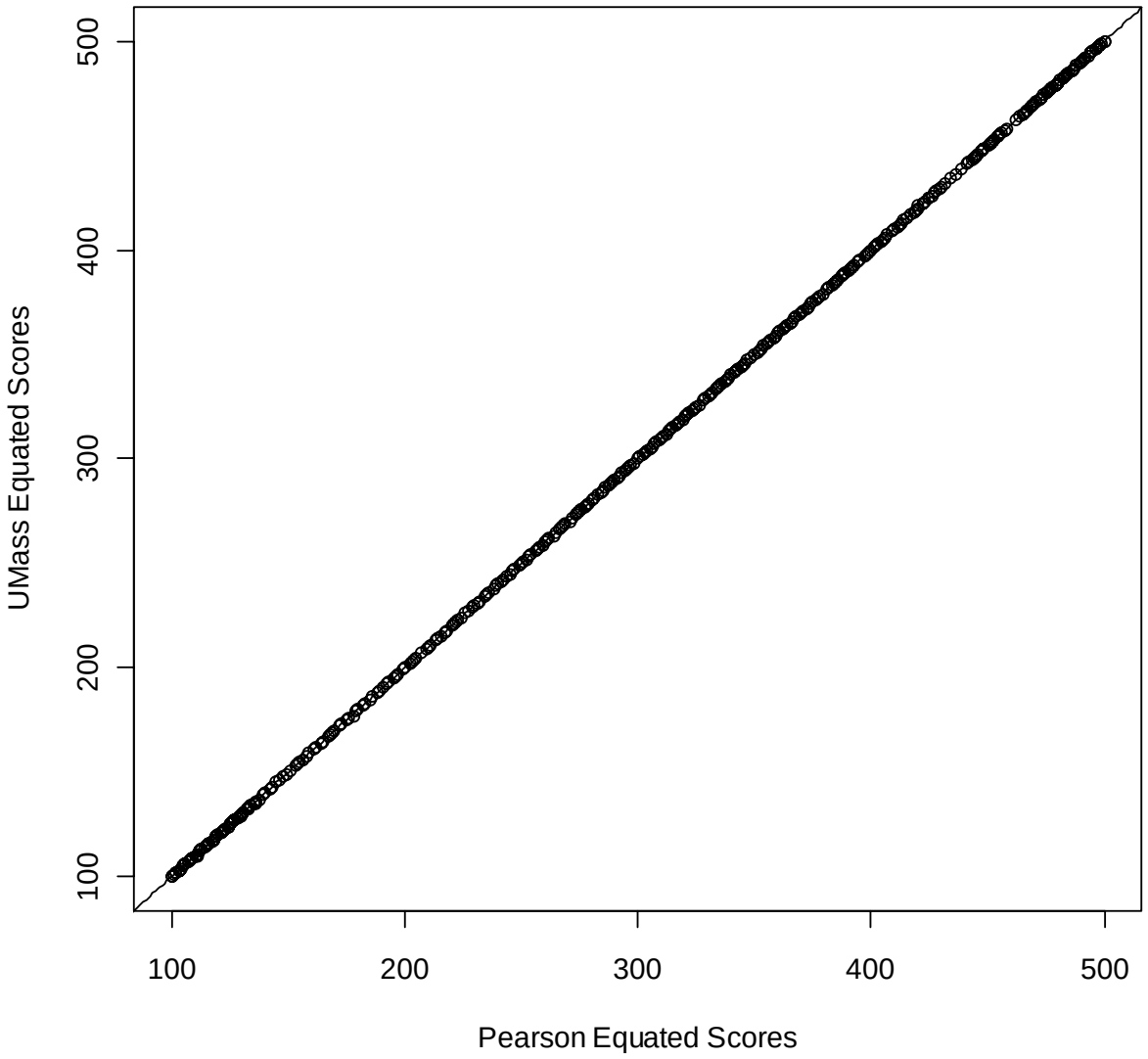
The equipercentile equating was conducted to adjust the 2011 scale score distribution such that it was equivalent to the 2010 distribution using the computer program RAGE-RGEQUATE (Zeng, Kolen, Hanson, Cui & Chien, 2005). The frequency distribution for the scale scores was created and used with the scale score frequency distribution from 2010 to create a conversion table linking the 2011 scale scores to the 2010 scale scores for the Grade 10

Reading test. The operational postsmooth limit was used in the equating. The postsmooth lower limit was based on the following equation: (percent of examinees who received a 100 scale score + 0.5)/100. The postsmooth upper limit was based on the following equation: (percent of examinees who received a 500 scale score + 0.5)/100. In this case, the lower and upper limit was 0.011[4] and 0.009, respectively. Therefore, the equipercentile was run twice - once for each limit. The final conversion table was a combination of the two equatings - equated scale scores for 100 to 300 were based on the postsmooth lower limit value and the equated scale scores for 300 to 500 were based on the postsmooth upper limit.

The final conversion table was nearly identical to that constructed by Pearson. Our equated scores correlated 0.9999 with Pearson's equated scale scores and differed by no more than 1 point for any equated score. The proportion of exact agreement exceeded 99%. To illustrate the high level of agreement, Figure V-4 compares the Pearson and SPS equated scale scores from the conversion table to the identity line which represents exact agreement. If our scale scores are in agreement, then the points should fall directly on the identity line. The solid line represents the identity line and runs perfectly through the majority of the equated scores. Although a few scale scores differed by one point, this was likely due to rounding error in the item parameter estimates used in the IRT scoring program and in rounding the scale scores. Because the conversion table was nearly identical, we agree with Pearson's final conversion table.

[4] The lower postsmooth limit was based on the 2010 frequency distribution because the percent of students with a scale score of 100 was higher for 2010 compared to 2011.

Figure V-4. Plot comparing Pearson and SPS equated scale scores.



After creating the conversion table, we transformed the 2011 scale scores onto the 2010 scale. The descriptive statistics for the equated scores, shown in Table V-4, were identical (to the second decimal place) to those reported by Pearson.

Table V-4

Descriptive Statistics for Equated Proficiency Estimates

	Pearson	SPS
Mean	309.60	309.60
Standard Deviation	61.40	61.41
Kurtosis	0.80	0.80
Skewness	-0.29	-0.29

Conclusion

In summary, we were able to successfully replicate Pearson's operational procedures and results including creating the calibration sample given the exclusion/inclusion rules, scoring the raw item responses, verifying the quality of the items (item statistics and model fit), reproducing identical item parameter estimates, and the (nearly) identical conversion table and scale score distribution on the FCAT 2.0 Grade 10 Reading Assessment. Given this successful replication, we feel confident that the operational procedures were conducted correctly.

Appendix V-A

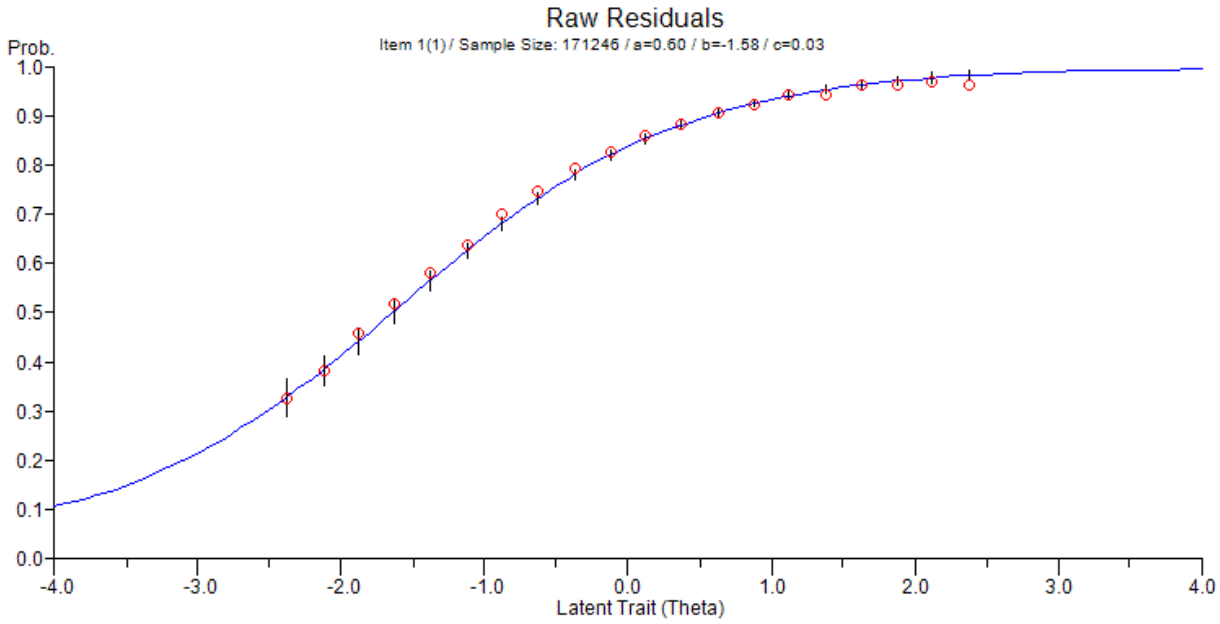
Sample MULTILOG Code for Grade 10 Reading

```
FCAT 2.0 2011
Grade 10 Reading
>PRO RAN,IND,NIT=45,NEX=171246,NGR=1,DAT='G10Reading.DAT';
>TES ITE=(1(1)45),L3;
>PRIORS ITE=(1(1)45),CJ,PAR=(-1.4,1.0);
>EST NCY=120;
>SAVE;
>END;
2
01
111111111111111111111111111111111111111111111
N
(45A1)
```

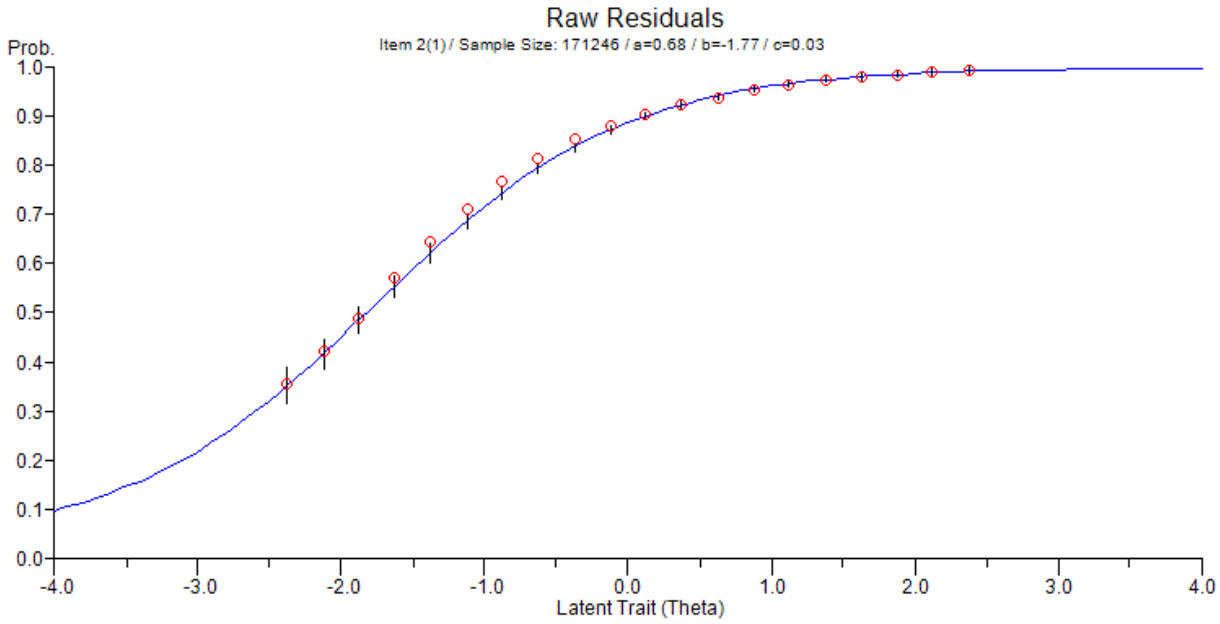
Appendix V-B

Model Fit Plots for Items Flagged During Item Calibration Inspection

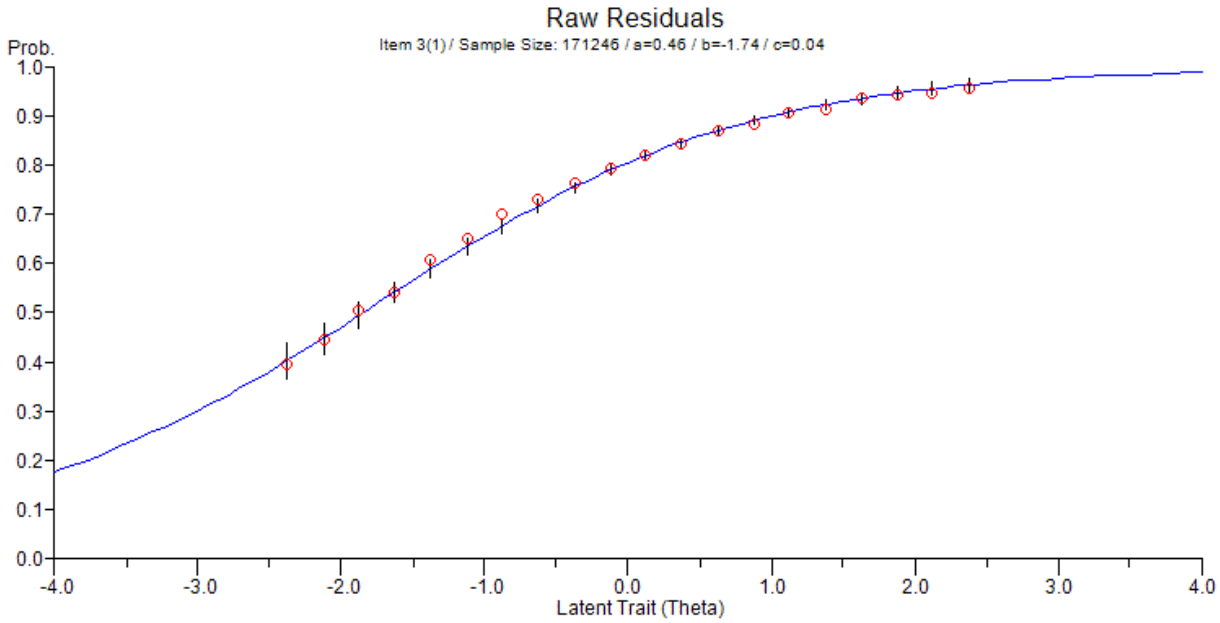
Item (SEQ) 1



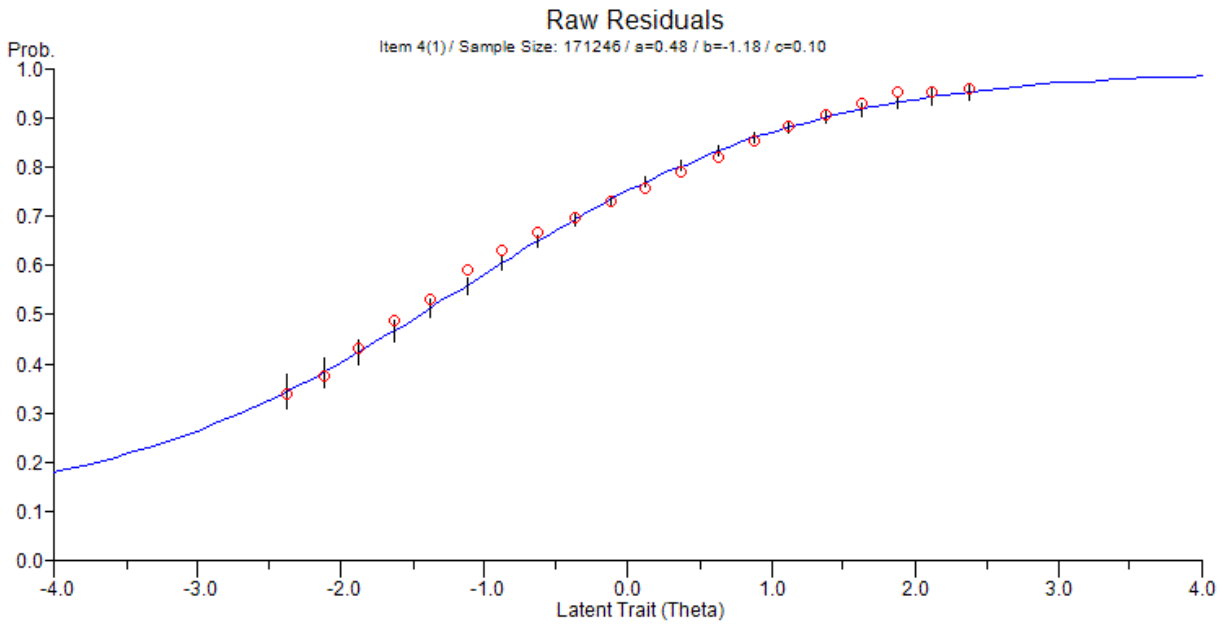
Item (SEQ) 2



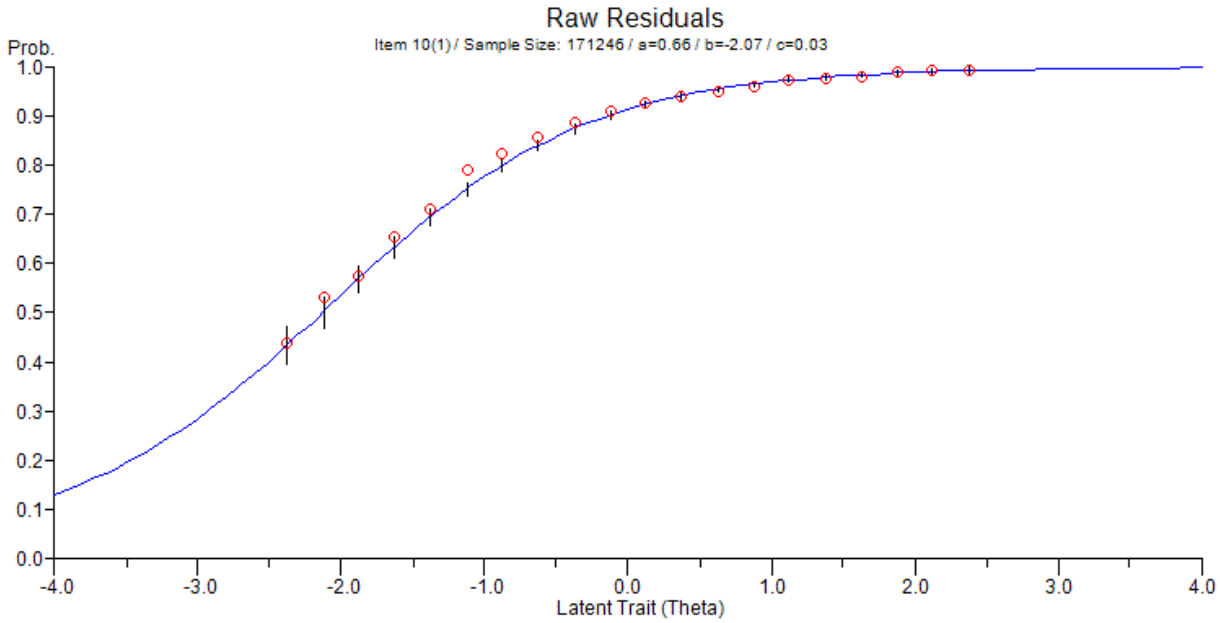
Item (SEQ) 3



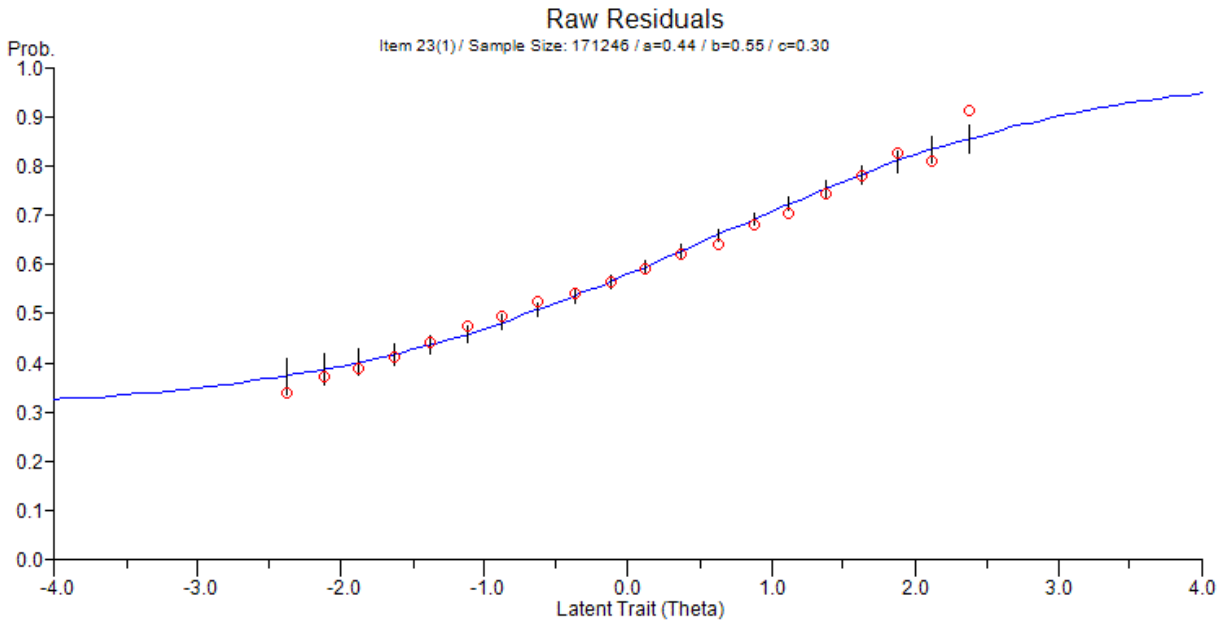
Item (SEQ) 4



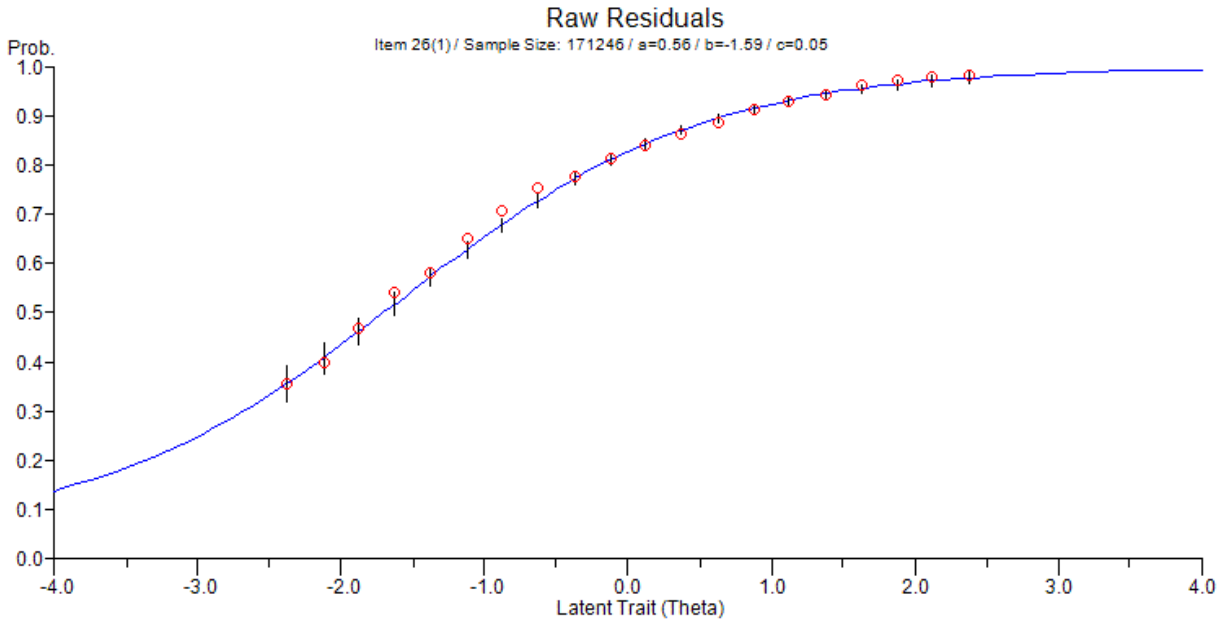
Item (SEQ) 18



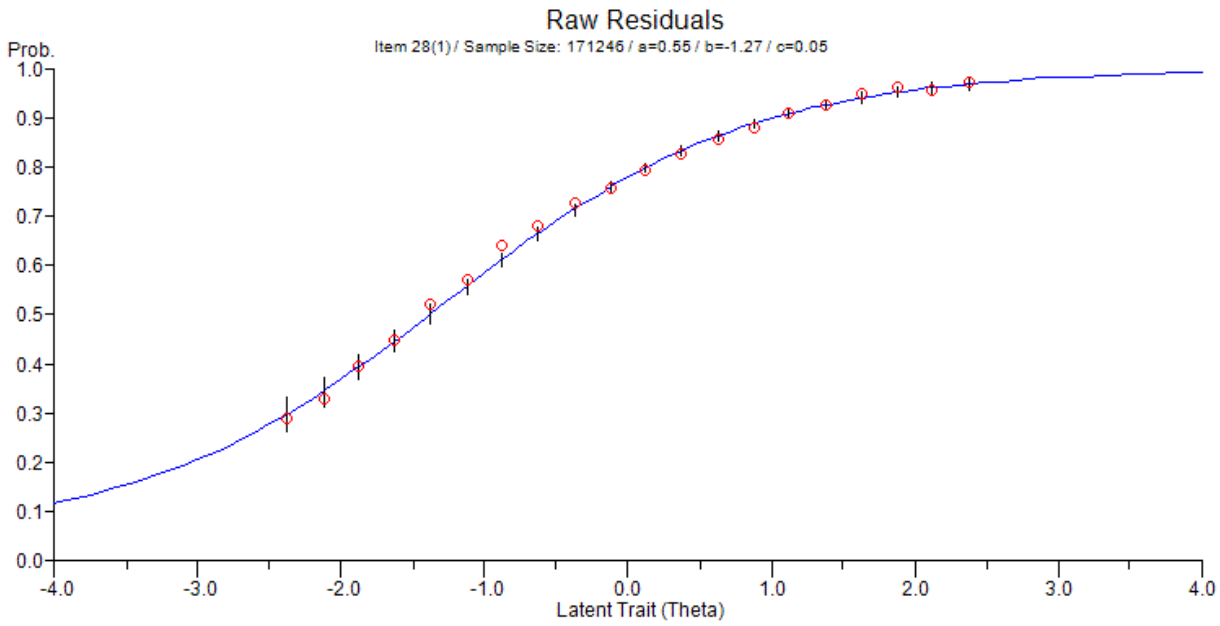
Item (SEQ) 31



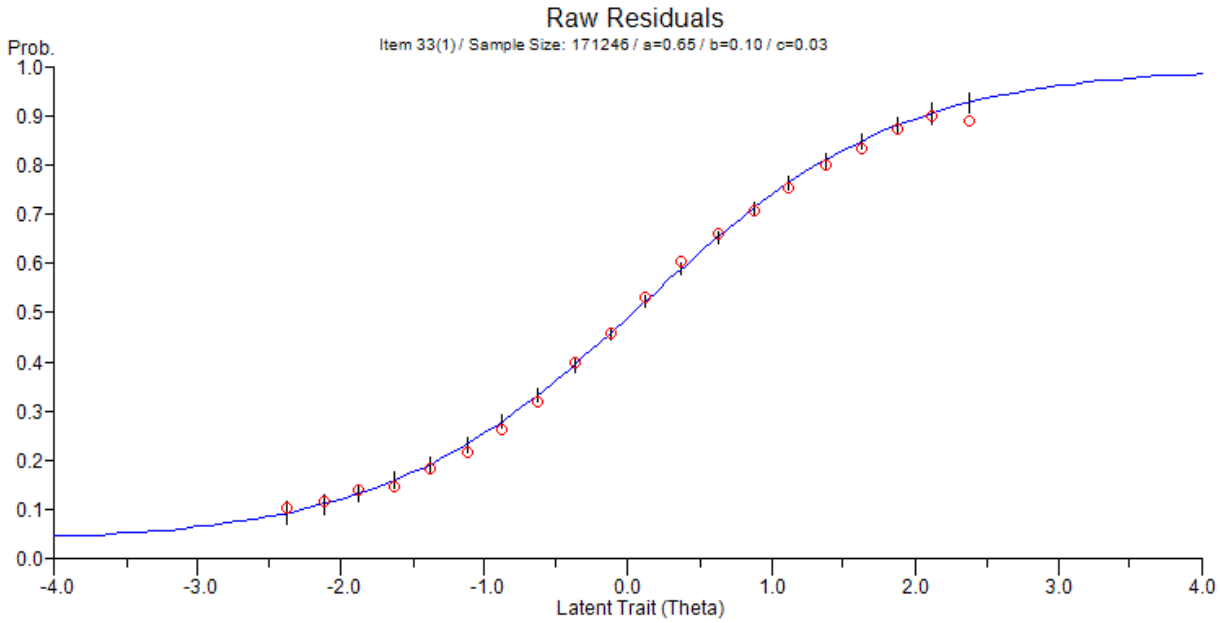
Item (SEQ) 34



Item (SEQ) 36



Item (SEQ) 41


Raw Residuals
Item 33(1) / Sample Size: 171246 / a=0.65 / b=0.10 / c=0.03
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

VI. Replication of the 2011 Algebra I End-of-Course Assessment

In this chapter, we summarize and discuss our analyses and results pertaining to the 2011 Algebra I End-of-Course assessment. As with the other exams, the chapter discusses receipt and cleaning of the data, descriptive statistics, IRT calibration and fit analyses, equating analyses, and a comparison of our results with those of Pearson.

Calibration Sample and Demographic Variables

The initial calibration file we downloaded from the sFTP site contained 200,201 students. We first excluded students who did not have a school type of 10, 11, 14, or 99 and those who used large print or Braille. From the excluded sample, we selected only students with reportable scores (Score Flag = 1 in the data file) and those who had not earned a previous algebra credit (CreditALG = 0) and were currently taking one algebra course (EnrolALG = 1, 2, 3, 4, or 5). Using the exclusion rules, the calibration sample contained 180,915 students. However, Pearson reported a calibration sample of 180,914 students. After further analyses, we discovered that the additional student in our sample took the Algebra test on paper. Therefore, that student was excluded from the calibration sample, producing a sample size of 180,914, which was identical to Pearson's report. The demographics for the calibration sample, shown in Table VI-1, were also identical to that reported by Pearson.

Table VI-1

Demographics for Algebra I Calibration Sample

		Pearson	SPS
Gender	Female	89,839	89,839
	Male	91,075	91,075
Ethnicity	Asian	4,701	4,701
	Black	40,551	40,551
	Hispanic	51,343	51,343
	American Indian/Alaskan	707	707
	Multiracial	4,706	4,706
	Native Hawaiian/Pacific	166	166
	Unknown	151	151
	White	78,589	78,589

Rescoring Item Responses and Flagging Items via Classical Item Statistics

Once the calibration sample was created, we next rescored the raw core item responses and compared them to the scored item responses provided in the data file. There were 40 forms, three of which were used for calibration (forms 100, 200, and 300). The forms were comprised of multiple-choice and fill-in response items, all dichotomously scored. The raw item responses were rescored using the answer key provided in the test map file. A correct response was given a 1 while an incorrect response was given a 0. Each rescored core item was identical to the scored items in the file provided by Pearson.

We next used classical item statistics to flag potentially problematic core and anchor items on the calibration forms (100, 200, and 300). Items with the following characteristics (determined by Pearson) were flagged:

- Classical item discrimination ($r_{pbi\text{-}c}$) was less than 0.2,
- Classical item difficulty (*p*-value) was greater than 0.9 or less than 0.15,
- An incorrect option was selected by more than 40% of the sample,

- The *p*-value on any one form differed from the overall *p*-value by more than $|0.08|$.

Using the above criteria, several core and anchor items on the calibration forms were flagged. Table VI-2 reports the flagged items for each form and the reason for being flagged.

Table VI-2

Flagged Items per Form

Form 100		Form 200		Form 300	
Item (SEQ)	Reason	Item (SEQ)	Reason	Item (SEQ)	Reason
4	$r_{pbi} = 0.07$	10	$r_{pbi} = 0.03$	9	$r_{pbi} = 0.04$
7	$r_{pbi} = 0.19$	11	$r_{pbi} = 0.10$	10	$r_{pbi} = 0.19$
	Option "B" = 0.47				
8	Option "B" = 0.44	25	$p = 0.13$	26	$p = 0.14$
9	$r_{pbi} = 0.09$	28	Option "B" = 0.42	28	Option "C" = 0.40
11	$p = 0.14$	30	$r_{pbi} = 0.17$	29	Option "B" = 0.52
38	$r_{pbi} = 0.17$	38	$p = 0.13$	36	$r_{pbi} = 0.19$
39	$p = 0.08$	39	$p = 0.11$	38	$r_{pbi} = 0.17$
40	$r_{pbi} = 0.17$	40	$r_{pbi} = 0.11$	40	$r_{pbi} = 0.19$
					$p = 0.13$
42	$r_{pbi} = 0.19$	45	$r_{pbi} = 0.19$	**41**	$\mathbf{r_{pbi} = -0.16}$
			$p = 0.01$		$\mathbf{p = 0.12}$
45	$r_{pbi} = 0.17$			42	$r_{pbi} = 0.01$
	$p = 0.01$				$p = 0.08$
53	$\mathbf{r_{pbi} = -0.001}$			45	$p = 0.01$
56	$p = 0.14$			56	$p = 0.15$
59	$r_{pbi} = 0.08$			59	$r_{pbi} = 0.16$
				61	$r_{pbi} = 0.17$

r_{pbi} = corrected point-biserial correlation.
p = p-value.

After examining each of the flagged items carefully, it was apparent that only two of the items (bolded in Table VI-2) were problematic, especially with respect to the item parameter estimation. We will address each of these two items in the item parameter calibration section, in particular, examining the reasonableness of the item parameter estimates and model fit.

Item Parameter Calibration

An initial item parameter calibration was conducted using the computer program MULTILOG[5] (Thissen, 2003) for each calibration form separately to examine the quality of the item parameter estimates and model fit. The three-parameter logistic model (3PLM) was used for the multiple-choice items and the two-parameter logistic model (2PLM) was used for the fill-in response items. A prior for the *c* parameter, which is based on the normal distribution with a mean of -1.4 and a standard deviation of 1 on the logit metric, was implemented only for the *c* parameter in the 3PLM. The item parameter estimates provide by MULTILOG were transformed onto the logistic metric so that we could compare them to the estimates reported by Pearson. The item parameter estimates were transformed as follows:

- For 2PLM

$$a_{\text{New}} = \frac{a_{\text{MLG}}}{1.7}$$

$$b_{\text{New}} = b_{\text{MLG}}$$

- For 3PLM

$$a_{\text{New}} = \frac{a_{\text{MLG}}}{1.7},$$

$$b_{\text{New}} = \frac{-b_{\text{MLG}}}{a_{\text{MLG}}}$$

and

$$c_{\text{New}} = \frac{\exp\left[c_{\text{MLG}}\right]}{1+\exp\left[c_{\text{MLG}}\right]}.$$

[5] We used a version of MULTILOG provided by Pearson because the commercial version could not accommodate sample sizes greater than 99,999.

Although MULTILOG successfully converged for each form, the two previously identified problematic items produced unreasonable item parameter estimates, mainly due to negative *a*-parameter estimates (Form 100-Item (SEQ) 53: a = **-0.69**, b = -3.73, and c = 0.28; Form 300-Item (SEQ) 41: a = **-0.64**, b = -2.74, and c = 0.05). The other items had reasonable item parameter estimates. The FDOE, Pearson, HumRRO and SPS teams discussed several strategies for dealing with the negative *a*-parameter estimates. First, the *a*-parameter estimate for the two aforementioned items was fixed to 0.10. Unfortunately, this produced unreasonable *b*-parameter estimates (b_{53} = 35.46, b_{41} = 57.25) and the model fit was unacceptable. A second strategy that SPS tried was to fix the *c*-parameter estimate - for item 53, the *c* parameter was fixed to 0.25, and for item 41, the *c*-parameter was fixed to 0.25. Although fixing the *c* parameter produced reasonable item parameter estimates for item 53 (a = 1.52, b = 2.44) and acceptable model fit, the solution was not satisfactory for item 41 in that the model fit was very poor. A third strategy examined deleting the two items (a third item from the same content area in Form 200 was deleted as well so that the reporting content areas would have the same number of items across calibration forms - this was deemed important for score reporting). The rationale for deleting the items was that the poor fit and negative discrimination was an indication that the item was not measuring the same construct as the other items on the test. Thus, including the items in the calibration may introduce unwanted test dimensionality, which may cause problems for establishing a unidimensional base scale. We agreed with the decision to exclude these items from the calibration.

Items were also flagged for detailed inspection using the following criteria: $a < 0.5$, $2.0 < b < -2.0$, or, for multiple-choice items, $c < 0.05$. Table VI-3 reports the flagged items per calibration form and the reason for being flagged.

Table VI-3

Items Flagged Given Above Criteria

Form	Item (SEQ)	Reason	*a*	*b*	*c*	Model Fit
100	1	$a < 0.5$ $c < 0.05$	0.47	-1.59	0.03	Small Misfit
100	9	$b > 2$	1.53	2.10	0.35	Small Misfit
100	59	$b > 2$	1.60	2.16	0.29	Small Misfit
100	38	$a < 0.5$ $b > 2$	0.49	2.74		Small Misfit
200	29	$c < 0.05$	0.71	-0.61	0.01	Good
200	39	$b > 2$	1.03	2.45		Good
200	45	$b > 2$	2.13	3.01		Good
300	29	$c < 0.05$	0.79	-0.58	0.01	Small Misfit
300	42	$b > 2$	1.74	2.99	0.08	Acceptable
300	38	$b > 2$	0.48	2.85		Small Misfit
300	45	$b > 2$	2.19	2.88		Good

The item parameter calibration and model fit of the flagged items were further inspected to determine if they should be excluded from additional analyses. Model fit was examined via an inspection of raw residuals around the item characteristic curve (ICC) that is defined by the item parameter estimates. The computer program ResidPlots (Liang, Han, & Hambleton, 2008) was used to examine model fit. Reasonable or acceptable model fit occurs when the majority of the observed proportions are randomly distributed around the ICC, with very few observed points falling far from the ICC. For each flagged item, the item parameter estimation was acceptable and the model fit was, at worst, slightly poor for a few of the items (Appendix VI-B contains model fit plots for each flagged item). However, the model fit for the items was still good-enough to be included in the establishing the base scale for the Algebra I assessment. Therefore, given that there were no key check issues with these items and that the item statistics and model fit was good-enough, we agree that these items should be included in establishing the base scale.

In addition to inspecting the model fit for the flagged items, we examined the model fit for all items. The model exhibited excellent to acceptable fit for most of the items. However, there were two items that exhibited small to moderate magnitudes of misfit, both displaying a similar pattern of non-monotonicity at the low to middle portion of the θ scale (see Figures VI-1 and VI-2). However, item (SEQ) 40 was dropped from the test due to a problem with the item format for an unknown proportion of the sample (see explanation in following paragraph). Although including item (SEQ) 11 in the calibration will not have an effect on the base scale, it may be important to consider whether it should be used in the pre-equated forms that will be administered in the following school year. Items for which the model does not fit well and that exhibit non-monotonicity can pose a problem for establishing a stable scale score over time. For example, the item parameter values may not be invariant and, thus, resemble item parameter drift. Unfortunately, anchor stability is not evaluated in pre-equated forms. Nonetheless, given all of the analyses to this point, we agree that all of the core and anchor items in the calibration forms, excluding those three items deleted due to poor statistics, should be included in establishing the base scale.

Figure VI-1. Model fit plot for item (SEQ) 40.

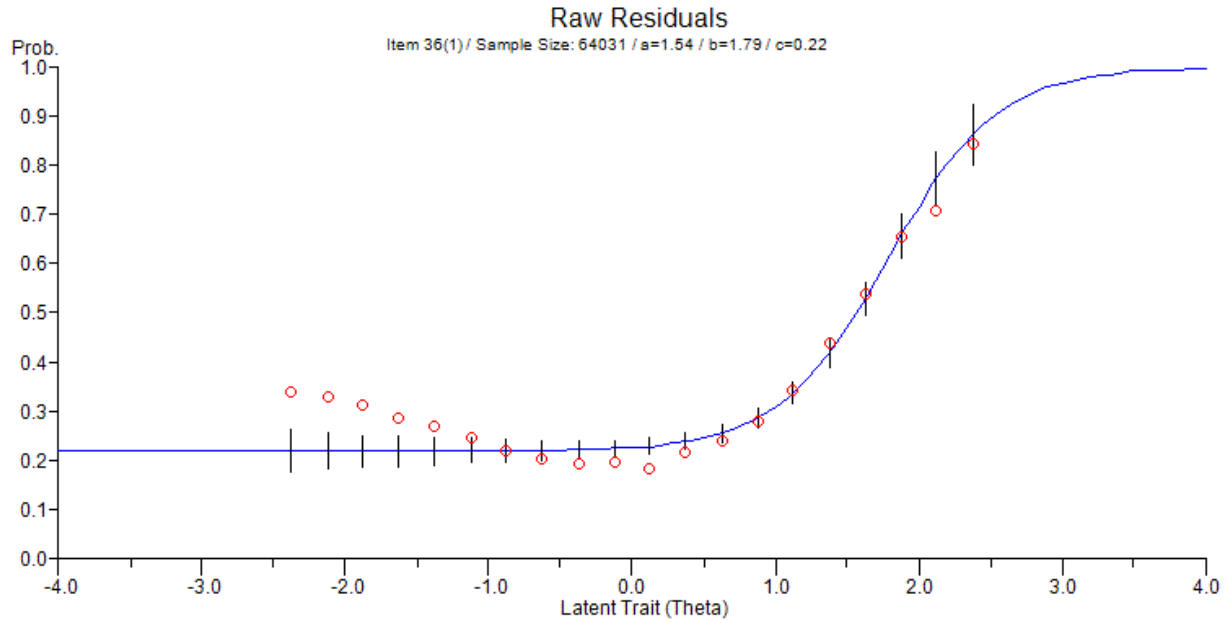
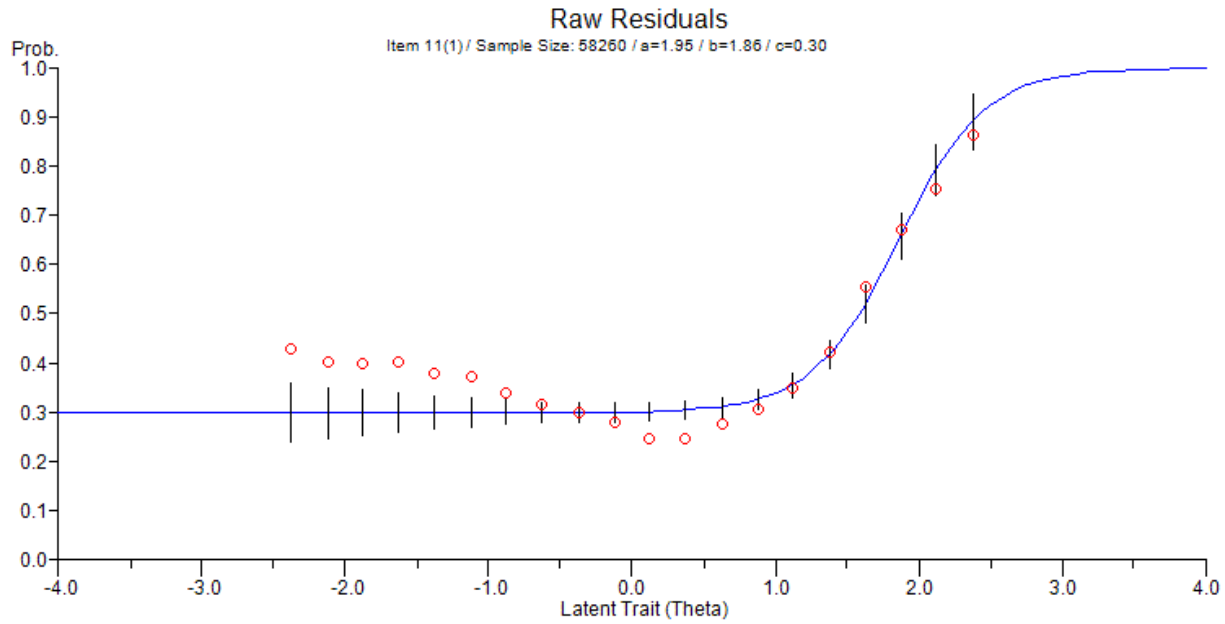


Figure VI-2. Model fit plot for item (SEQ) 11



In addition to the three items previously deleted, one item from each calibration form was deleted due to an item format issue. These three items used Venn diagrams and required the student to observe the shaded area of the diagram. Unfortunately, for an unknown proportion of the sample, the shaded area in the Venn diagram was not present. To address this issue, it was decided to exclude these items from the calibration. We find this solution acceptable (although including these items in the calibration would not influence the base scale - but they should not be included in building pre-equated forms due to the potential for inaccurate item parameter estimates).

Once the final set of items was established, we calibrated the item parameters using the computer program MULTILOG (Thissen, 2003) via concurrent calibration on the calibration sample and forms. Sample MULTILOG code is provided in Appendix VI-A. All of the final transformed item parameter estimates and their corresponding standard errors were reasonable values. Furthermore, the item parameter estimates were nearly identical to those reported by Pearson. The *a*-, *b*-, and *c*-parameter estimates correlated with those reported by Pearson about 0.999. Furthermore, any difference in item parameter estimates were observed in the second and third decimal place. Therefore, we agree with the final set of item parameter estimates that will be used to establish the base scale for the Algebra I EOC assessment.

Proficiency Score Estimation

The final item parameter estimates for the calibration forms were used to estimate the IRT proficiency parameter. The computer program IRT Score Estimation (Chien, Hsu, & Shin, 2011) was used to perform maximum likelihood estimation. The IRT proficiency estimates were converted to *T*-scores as follows: $\hat{\theta}_T = \hat{\theta} * 10 + 50$. The mean and standard deviation for all examinees was nearly identical to those reported by Pearson (see Table VI-4).

Table VI-4

Descriptive Statistics for Proficiency Estimates (*T*-scores)

	Pearson	SPS
Mean	49.41	49.42
Standard Deviation	11.50	11.51

Conclusion

In summary, we were able to successfully replicate Pearson's operational procedures and results including creating the calibration sample given the exclusion/inclusion rules, scoring the raw item responses, verifying the quality of the items (item statistics and model fit), reproducing (nearly) identical item parameter estimates, and the (nearly) identical *T*-score distribution descriptive statistics for the 2011 Algebra I EOC Assessment. Given this successful replication, we feel confident that the operational procedures were conducted correctly.

Appendix VI-A

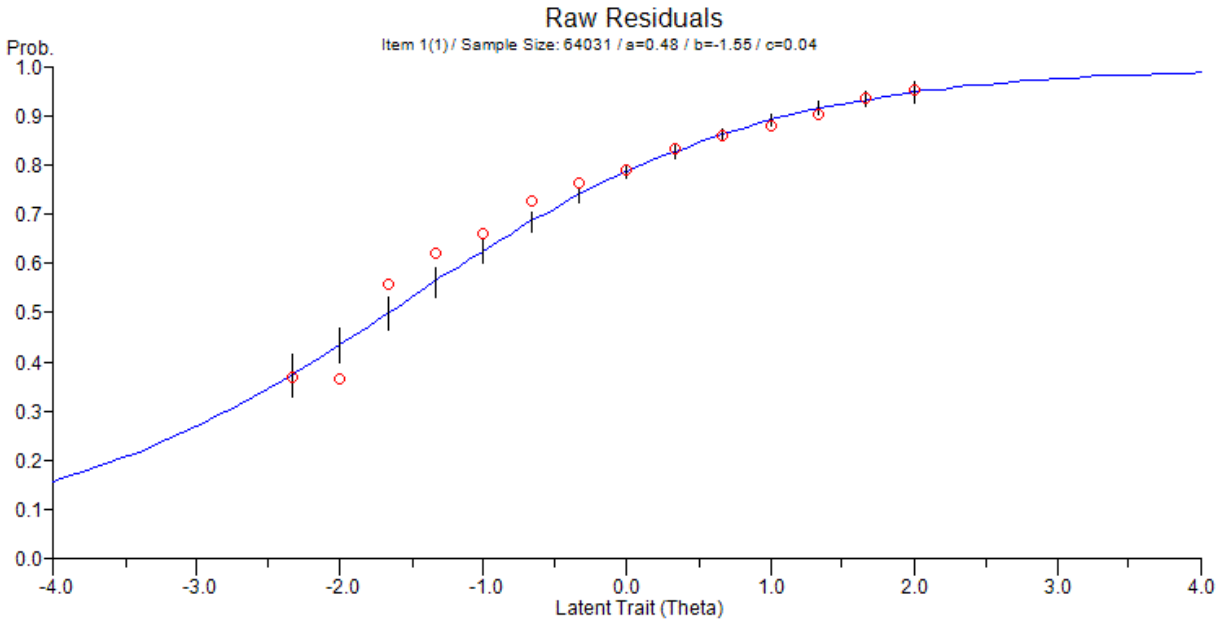
Sample MULTILOG Code for Algebra I EOC

```
FCAT 2011
Algebra I EOC
>PRO RAN,IND,NIT=162,NEX=180914,NGR=3,DAT='Algebra.DAT';
>TES ITE=(1(1)32,53(1)86,107(1)140),L3;
>TES ITE=(33(1)52,87(1)106,141(1)160),L2;
>EQU AJ,BJ,CJ ITE=(61,62,63,64,66,67,69)
WIT=(9,10,11,12,13,14,15);
>EQU AJ,BJ,CJ ITE=(117,118,119,120,121,122,123)
WIT=(9,10,11,12,13,14,15);
>EQU AJ,BJ,CJ ITE=(74,75,80,81,82,83) WIT=(20,21,26,27,28,29);
>EQU AJ,BJ,CJ ITE=(128,129,134,135,136,137)
WIT=(20,21,26,27,28,29);
>EQU AJ,BJ ITE=(93,94,95,102,103,104,105)
WIT=(39,40,41,48,49,50,51);
>EQU AJ,BJ ITE=(147,148,149,156,157,158,159)
WIT=(39,40,41,48,49,50,51);
>PRIORS ITE=(1(1)34,55(1)88,109(1)142),CJ,PAR=(-1.4,1.0);
>EST NCY=300;
>SAVE;
>END;
3
019
11111111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111111
Y
9
(I1,162A1)
```

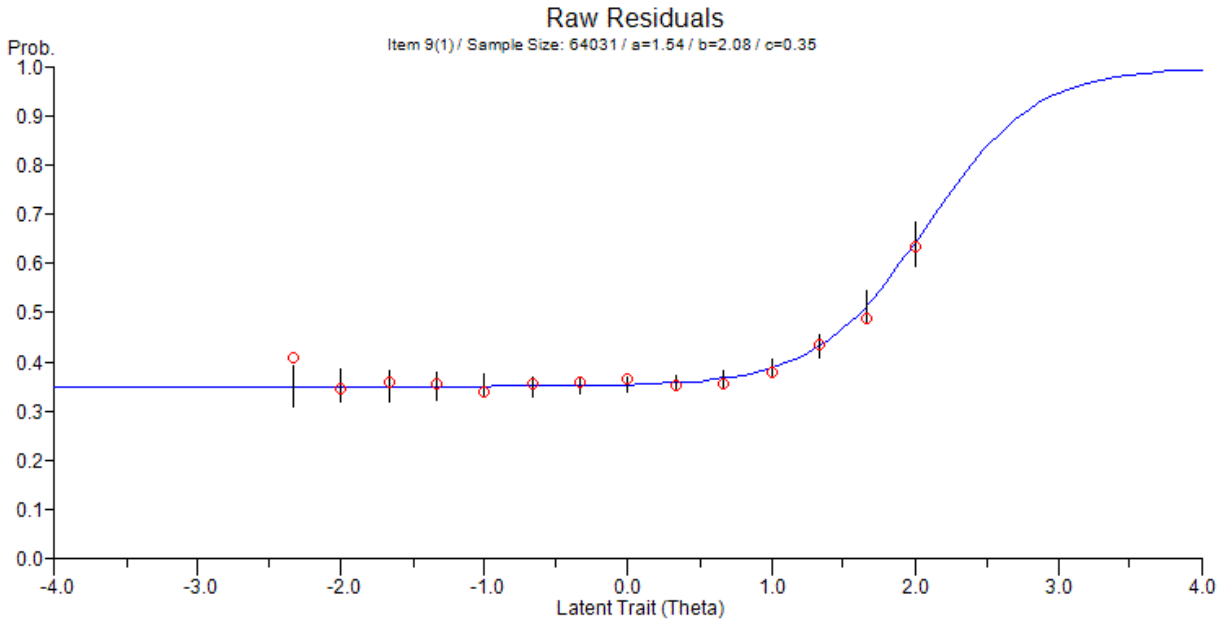
Appendix VI-B

Model Fit Plots for Items Flagged During Item Calibration Inspection

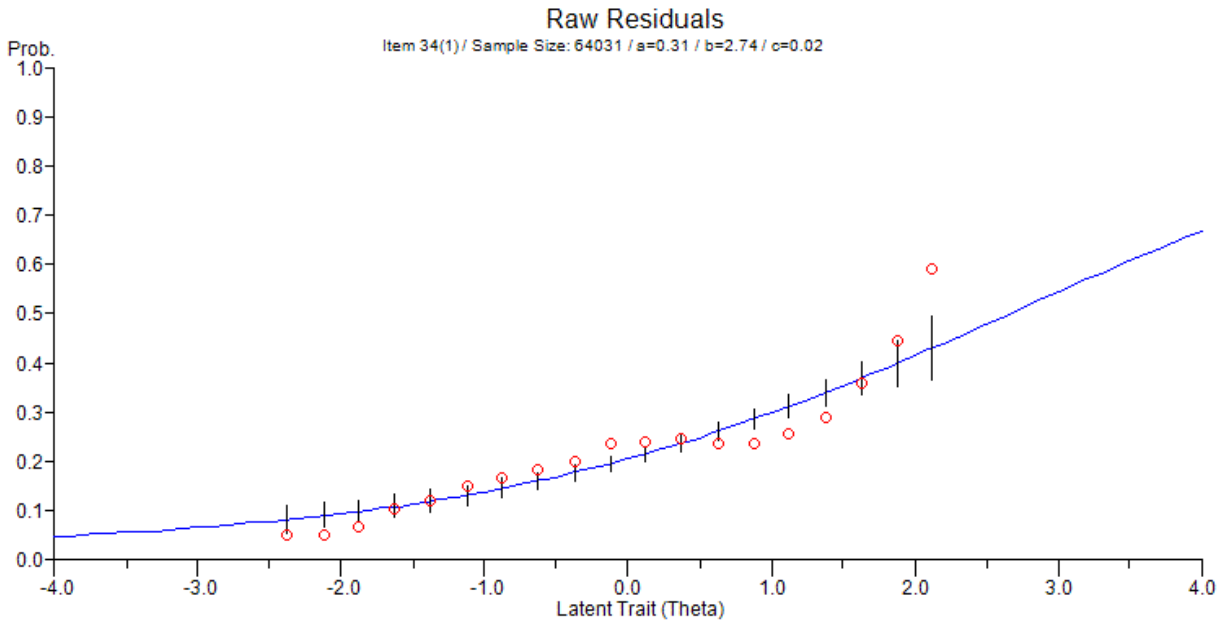
Form 100: Item (SEQ) 1



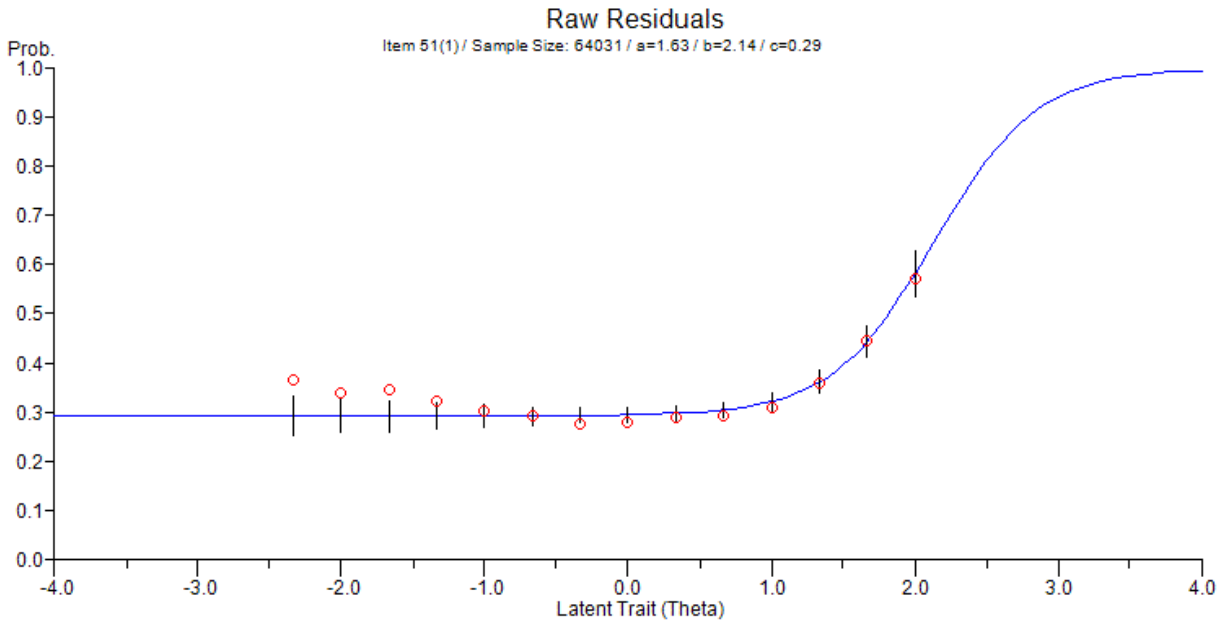
Form 100: Item (SEQ) 9



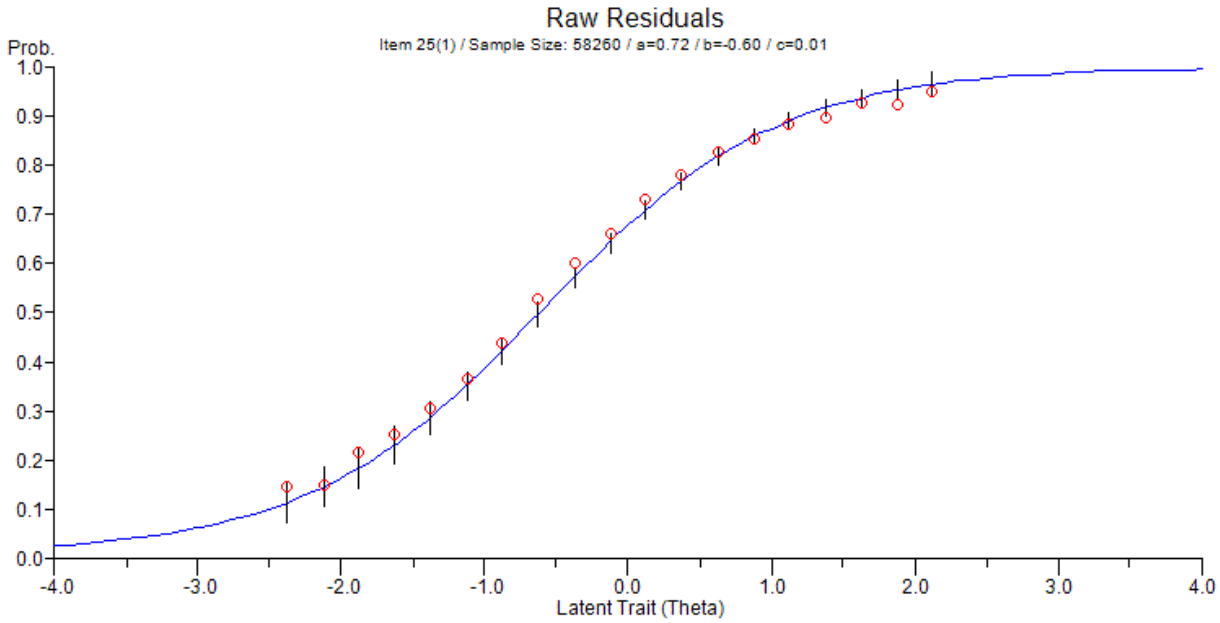
Form 100: Item (SEQ) 38



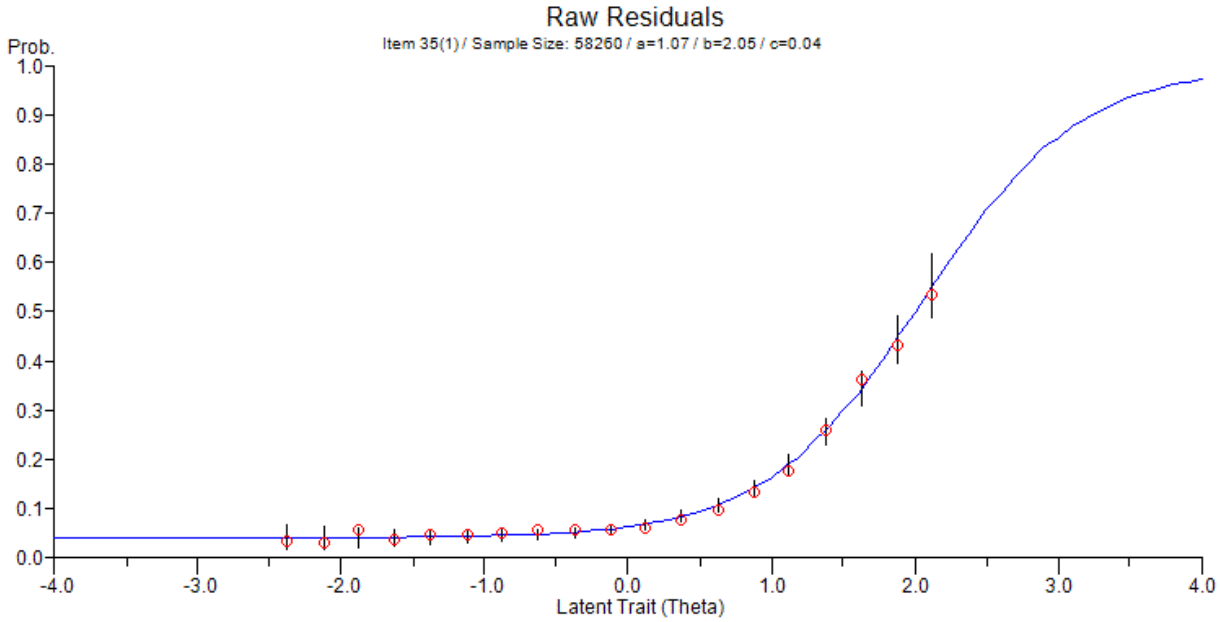
Form 100: Item (SEQ) 59



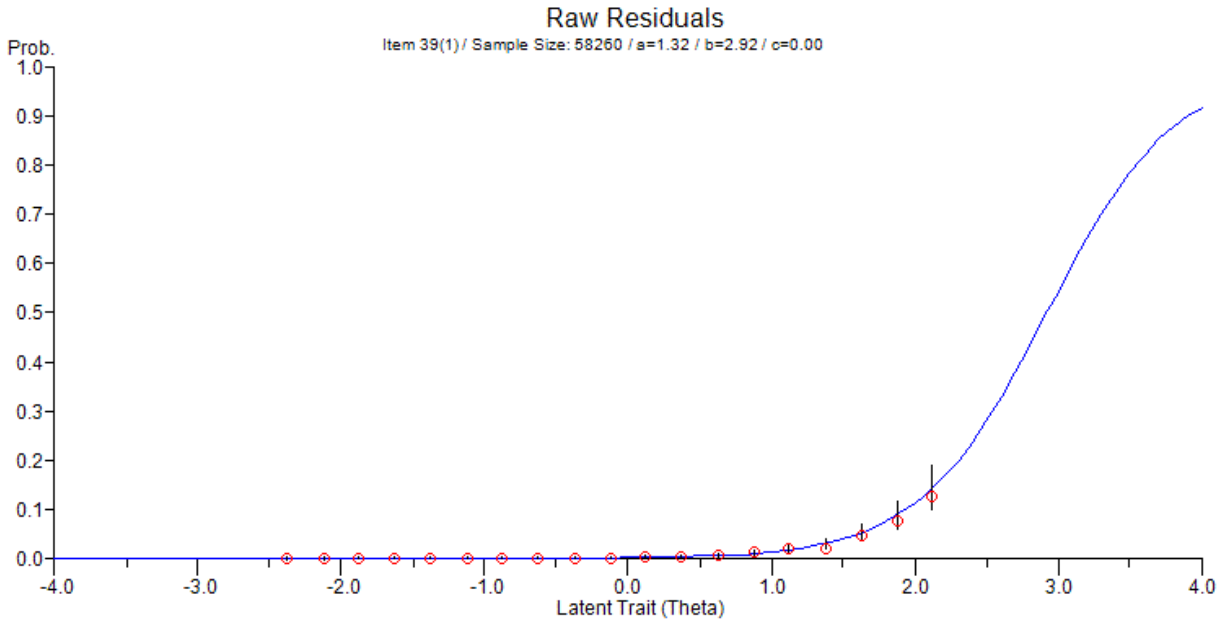
Form 200: Item (SEQ) 29



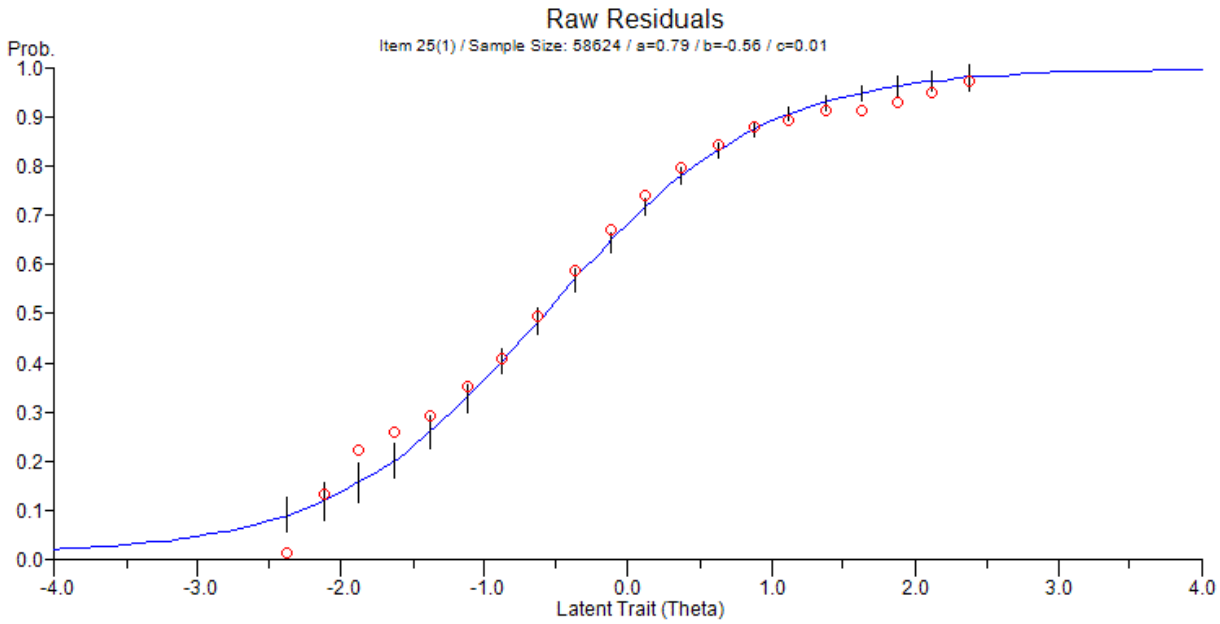
Form 200: Item (SEQ) 39



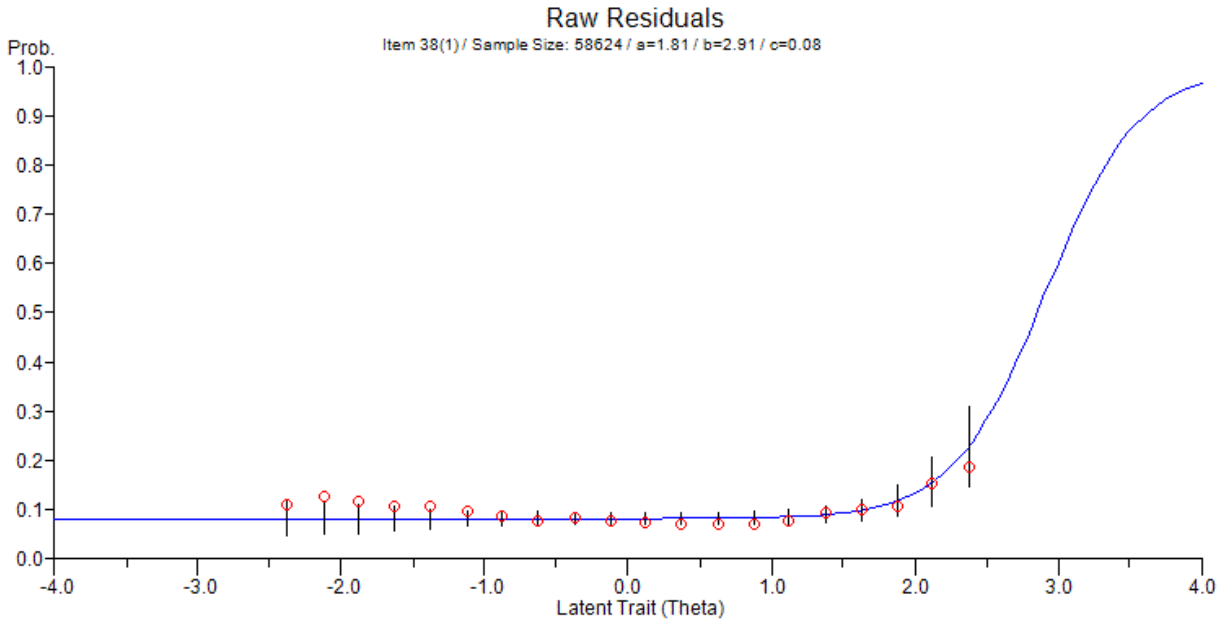
Form 200: Item (SEQ) 45



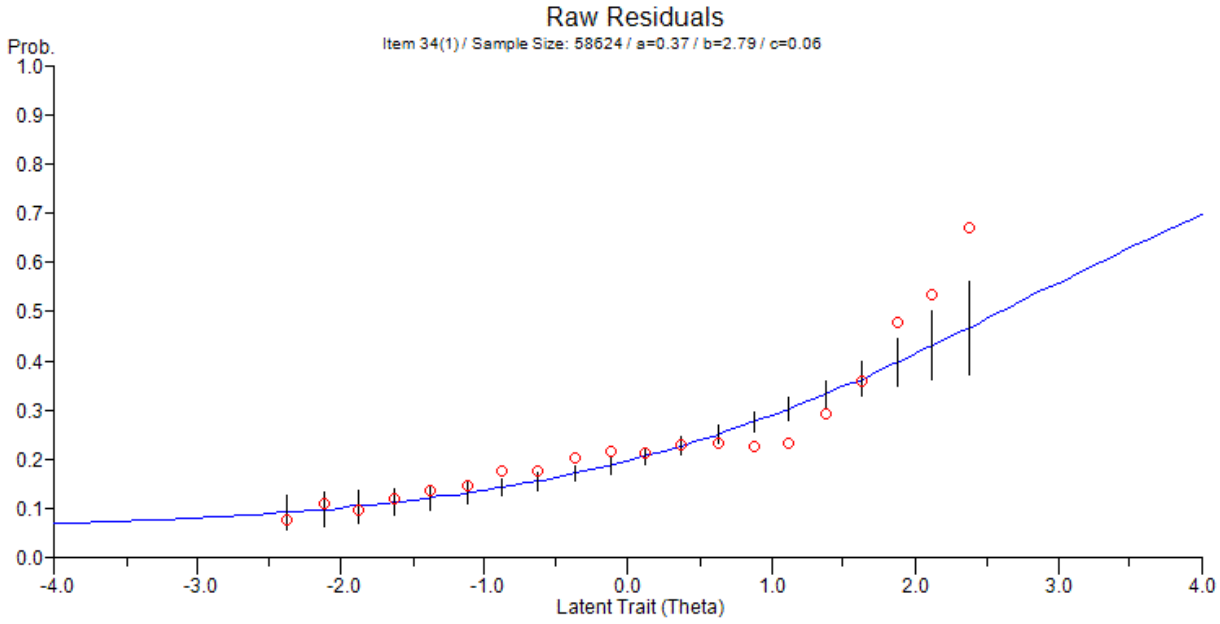
Form 300: Item (SEQ) 29



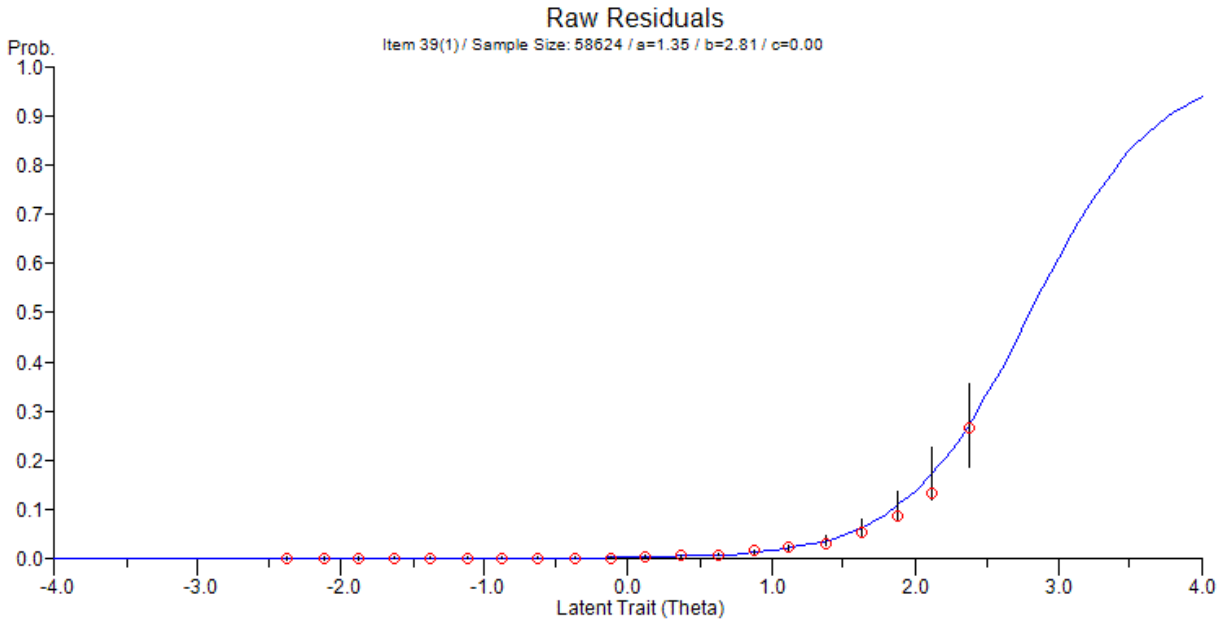
Form 300: Item (SEQ) 42



Form 300: Item (SEQ) 38



Form 300: Item (SEQ) 45



References

Chien, M. , Hsu, Y., & Shin, D. (2006). *IRT Score Estimation Program [computer program]*. Iowa City, IA: Pearson.

Kim, S., & Kolen, M. J. (2004). *STUIRT* [Computer software]. Iowa City, IA: Iowa Testing Programs, The University of Iowa.

Liang, T., Han, K. T, & Hambleton, R. K. (2008). ResidPlots-2:computer software for IRT graphical residual analyses, Version 2.0 [Computer Software]. Amherst, MA: University of Massachusetts, Center for Educational Assessment.

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

Thissen, D. (2003). *MULTILOG 7: Multiple categorical item analysis and test scoring using item response theory [computer program]*. Chicago, IL: Scientific Software.

Zeng, L., Kolen, Ml. J., Hanson, B. A., Cui, Z., & Chien, Y. (2005). *RAGE-RGEQUATE [computer program]*. Iowa City, IA: The University of Iowa.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG 3.0 [computer software]. Lincolnwood, IL: Scientific Software International.