Appendix 3.A Review and Replication Analysis English 2003

Maryland High School Assessment

Review and Replication Analysis

English 2003

February 2, 2004

Educational Testing Service

Appendix 3.A Review and Replication Analysis English 2003

MSDE asked ETS to investigate and replicate the 2003 analyses for the English High School test completed by their previous vendor, CTB/McGraw-Hill. An estimated 6% drop in students classified as proficient in 2003 compared to 2002 at the state level prompted this request. The purpose of this study was to 1) review the technical documentation and steps completed by CTB/McGraw-Hill and note any suggested modifications; 2) replicate the study completed by CTB/McGraw-Hill; and 3) determine if a change in linking design would have made any important difference in the percent of students identified as proficient.

Summary of the Process Completed by CTB/McGraw-Hill

Based on the technical documentation, the analyses completed were consistent with high stakes assessment programs and involved item analyses, calibration, and equating. Item-pattern scoring was completed using the resulting item parameters. The completion of the work was within normal standard with the exception of the linking study design and outcome.

January Administration

For the January administration, four forms were administered, forms A, B, C, and W. Forms A, B, and C were built to match the test blueprint and consisted of items administered in 2002, as well as items field tested in 2000 and 2001, along with an embedded field test section. These forms also shared a common anchor set of 36 selected-response items. Form W was an exact duplicate of the 2002 Form W; all items were administered and calibrated in May 2002. This form did not match the test blueprint but did consist of a mix of selected-response (SR), brief constructed-response (BCR) and extended constructed-response (ECR) items (see Table 3.A.1). Including the embedded field test section on Forms A-C, all administered forms had very similar test lengths although Form W had 4 to 5 more SR items than the other forms.

Table 3.A.1. Number and Type of Item by Form

Form	Item Type	SR	BCR	ECR
A	FT	15	1	-
	OP	50	2	1
B	FT	16	1	-
	OP	50	2	1
C	FT	15	1	-
	OP	50	2	1
W	OP	70	3	1

Note. FT= field test. OP= operational.

Appendix 3.A

There were no common items between Form W and Forms A-C (see Table 3.A.2). The operational scale was defined by the 2002 administration, and the equating and linking design for the January 2003 forms was based on a mixed common item, randomly equivalent groups design. An intact form from 2002 (Form W) was spiraled along with the three new forms (A-C). Forms A-C shared a common anchor set, however there were no common items between these forms and Form W. The linking study design was to complete a concurrent calibration of Forms A-C with Form W, then link all of the forms to the 2002 scale through Form W. That is, the new, 2003 parameter estimates for Form W would be linked to the 2002 parameter estimates using a Stocking and Lord procedure. The resulting transformation constants would then be applied to Forms A-C. With the exception of four items that were removed from the calibration and anchor sets due to poor item performance[3], following the 2002 administration, all SR items on Form W were identified as the anchor set to place the 2003 forms onto the 2002 scale. Note, "X" represents a block of items.

Table 3.A.2. Composition of January Forms Relative to Previous Administrations

	2000 – 2001 Field Test Administrations		2002 Administration (0perational Scale)		Embedded Field Test
	Unique Items	Common Items	Unique Items	Common Items	Unique Items
F/T Pool	X	X			
2002 Pool			X	X	
	Items Contributing to Student Scores 2003				
2003 W			X		
2003 A		X			X
2003 B[1]		X			X
2003 C		X			X

[1]Note. Form B also included 2 items from the 2002 administration.

The intended linking study design was dependent on the assumption that the forms would be completed by randomly equivalent groups of students. To obtain randomly equivalent samples, the four forms were packaged and spiraled within each classroom. That is, the first student would receive Form A, the second Form B, the third Form C, the fourth form W, the fifth, Form A, and so on. The exception was the accommodation package – forms administered to students requiring specific accommodations. For these students,

[3] Items were not calibrated following the 2002 administration due to poor classical statistics. These items had very low or negative point-biserial correlations.

only Form A was included in the package; however, it was expected that only a small percentage of students required special accommodations to complete the form.

After the forms were administered and scored, it was determined that the forms were not administered to randomly equivalent groups of students. Based on the Draft 2003 Technical Report (CTB/McGraw-Hill, December, 2003) the forms were not administered to randomly equivalent groups of students because:

1. Large print and Braille forms were available for Form A only, resulting in disproportionate numbers of accommodated students receiving these forms.

2. Special Education students tended to be over represented [sic] on the first couple of forms within each content area. It appears that administrators tended to use the first one or two forms in each package for a disproportionate number of students who required special accommodations.

> Because of the requirement that these students be included in the calibration and equating, it was not possible to sample down in order to achieve comparable groups across test forms (p. 37).

As a result, a modification was made to the intended linking design. The steps completed were summarized below:

1. All forms were calibrated together in a single calibration run, then, using Form W, the forms were equated to the 2002 scale via a Stocking and Lord procedure, and the parameters for all forms adjusted with the resulting equating constants.

   Because Forms A, B, and C shared anchor items, this step placed these 3 forms on the same scale. However, W did not share items with A, B, and C, so this procedure did not place W on the same scale as A, B, and C via anchor items. Random equivalence of samples also did not place W on the same scale as the three other forms,[4] so an additional step was needed.

2. A second linking step was completed. This involved equating Form C to Form W using a linear approximation to equipercentile equating procedure. To complete this step Form W was scored with the 2002 item parameters and Form C was scored with the 2003 item parameters. The resulting equating constants were then applied to the items in forms A, B, and C. The rationale for this step was that the test scores and demographic characteristics of the students completing Form W were very similar to Form C. Due to the

[4] It appears that CTB/McGraw-Hill's parameter estimation software, Pardux, is not designed to automatically align parameters from non-overlapping, randomly equivalent samples. An external procedure, such as the linear equipercentile procedure, is needed. Because this external step is needed, the Stocking and Lord procedure used in Step 1 was not necessary and was over-ridden by Step 2.

disproportionately high representation of ESL and Special Education students, Form A had substantially lower test scores than the other forms.

For scoring purposes, the transformed parameters from step 2 above were used for Forms A, B, and C. Form W was scored with the item parameters estimated in 2002.

Appendix 3.A

May 2003 Administration

For the May administration, 11 forms were administered: Forms D, E, F, G, H, J, K, L, M, N, and P. All forms were similar with regard to the distribution of item type and test length (see Table 3.A.3).

Table 3.A.3. Number and Type of Item by Form

Form	Item Type	SR	BCR	ECR
D	ANC	33	-	-
	OP	17	2	1
	FT	19	1	-
E	ANC	33	-	-
	OP	17	2	1
	FT	16	1	-
F	ANC	33	-	-
	OP	17	2	1
	FT	17	1	-
G	ANC	33	-	-
	OP	17	2	1
	FT	17	1	-
H	ANC	33	-	-
	OP	17	2	1
	FT	16	1	-
J	ANC	33	-	-
	OP	17	2	1
	FT	17	1	-
K	ANC	33	-	-
	OP	17	2	1
	FT	17	1	-
L	ANC	33	-	-
	OP	17	2	1
	FT	15	1	-
M	OP	50	2	1
	FT	17	1	-
N	OP	49	3	1
	FT	18	-	-
P	OP	50	2	1
	FT	16	1	-

Note. FT= field test. OP= operational. ANC= anchor.

Forms D through L were built to match the test blueprint and consisted of items administered in 2002, as well as items field tested in 2000 and 2001, along with an embedded field test section. These forms shared a common anchor set of 36 selected-

response items with Forms A-C. Form M contained 28 SR items that were also administered in one of the forms administered in 2002, as well as newly developed items. Forms N and P were identified as "block field test books" and included only newly developed items. Forms M, N, and P had no items in common with either Form W or Forms A-L. Table 3.A.4 illustrates the composition of the 2003 forms relative to previous administration and new development. Note, "X" represents a block of items.

Table 3.A.4. Composition of 2003 Operational Forms Relative to Previous Administrations

	2000 – 2001 Field Test Administrations		2002 Administration (Operational Scale)		Field Test Items
	Common Items	Unique Items	Common Items	Unique Items	Unique Items
2002			X		
Items Contributing to Student Scores in 2003					
2003 W			X		
2003 A	X	X			X
2003 B[1]	X			X	X
2003 C	X				X
2003 D	X	X		X	X
⋮	⋮			⋮	…
2003 L	X	X		X	X
2003 M				X	X X
2003 N					X X
2003 P					X X

[1]Note. Form B also included 2 items from the 2002 administration.

Linking the May forms to the operational scale involved the following steps:

1. All forms were concurrently calibrated. This produced item parameters for each form approximately[5] relative to a true theta scale with distribution Normal (0,1).

2. Forms D-L were linked to the operational [scale score] scale via the set of common items shared with Forms A-C from the January administration in a

[5] Again it does not appear that Pardux is designed to precisely align parameters from simultaneous calibrations of forms with no over-lapping anchor items. Steps 2 and 3 provided the necessary link across forms.

Stocking and Lord procedure, and the item parameters were adjusted with the resulting equating constants.

3. Forms M, N, and P were placed onto the operational scale by equating each form to Form L using a linear approximation to equipercentile equating.

Item pattern scoring was completed using the resulting transformed item parameters.

The final resulting scale score means and standard deviations for each test form were listed in Table 3.A.5 below (CTB/McGraw-Hill Technical Report, pp. 40-41). The mean scale scores ranged from 390.3 to 399.4. The mean score was lowest for the first form of the spiral in both the January (Form A) and the May (Form D) administrations. In both cases these forms were also administered to the largest number of students within each of the calibration samples. While large print and Braille forms administered in May were the same forms administered in January (Form A), students with other types of accommodations were administered one of the May test forms. Of note, students completing a make-up form were excluded from the calibration samples.

Table 3.A.5. CTB/McGraw-Hill Summary Statistics English 2003

Form	N[1]	Mean	SD
January			
A	2370	390.8	38.1
B	2090	396.4	34.6
C	2019	395.5	34.6
W	1986	395.5	34.4
May			
D	5831	390.3	39.8
E	4797	397.7	34.5
F	4806	398.0	34.8
G	4772	397.1	34.3
H	4775	397.5	35.5
J	4720	397.9	35.5
K	4673	399.4	34.4
L	4600	398.8	36.2
M	4596	398.7	36.7
N	4508	398.7	36.7
P	4483	398.9	35.9

[1]Note. Based on calibration samples.

A summary of the results from the January and May administrations compared to the 2002 results were presented in Table 3.A.6; the 2002 results were taken from the CTB/McGraw-Hill Technical Report (p. 43) and included all students that participated in each administration. The results in Table 3.A.6 include a large number of students taking a make-up form: 933 students completed a make-up form in January and 3,543 students completed a make-up form in January. The make-up forms in both administrations were

the same. Form A was administered in the first make-up week; Form B was administered in the second make-up week. These students generally performed much poorer relative to the calibration sample. For example, students completing Form A in the first make-up week of the January administration had a mean scale score of 353 (sd 61.5).

The mean score for the January 2003 administration was 8.8 points lower. However, there was less than one score point difference between the May 2002 and May 2003 administrations. Unlike in 2002, in 2003 the scores for the January administration were 5.1 points lower than the scores for the May administration. Also noted was the difference in the test score variation in May 2003 compared to all other administrations.

Table 3.A.6. CTB/McGraw-Hill Summary Statistics by Administration and Year

Administration	N	Mean	SD
January 2002	9,339	398.3	41.0
May 2002	52,172	395.4	47.0
January 2003	9,488	389.5	42.2
May 2003	56,426	394.6	39.5

Study Methodology

The main purpose of this study was to replicate the results obtained by CTB/McGraw-Hill and to identify any design revisions that may produce different results. To replicate the results, all analyses steps, as described in the technical documentation supplied by CTB/McGraw-Hill were completed. Items were calibrated using Multilog (Scientific Software International, Inc.). This software allows for the estimation of item parameters for both selected response (SR) and constructed response (CR) items. ETS proprietary software was used to complete the Stocking and Lord and the linear approximation to equipercentile equating procedures.

Results

The percent of students included in the calibration sample overall and by form were presented in Table 3.A.7 for the January administration and Table 3.A.8 for the May administration. Specific information for the calibration sample was not included in the CTB/McGraw-Hill Technical Report. Therefore, the data contained in this column consists of all students, including students completing a Braille form or a make-up form.

As observed by CTB/McGraw-Hill, Form A had the largest case count - 283 more students completed this form compared to Form B. Moreover, the first form in the May administration (Form D) also had the largest case count – 5827 compared to 4799 for Form E. Regardless of the differences in case counts, the forms were spiraled to similar proportions of students for all of the demographic variables except Special Education students. In January 16.7% of the Form A sample were identified as Special Education students, compared to 9% for Forms B and C. In May, the differences were even more

pronounced: 21.8% of the Form D calibration sample were identified as Special Education students, compared to only 8.1 to 9.1% of the sample for the other forms.

Table 3.A.7. Characteristics of Calibration Samples by Form for January 2003

	CTB [1]	Replication				
	Total N=9488	Total N=8436	Form A N=2364	Form B N=2084	Form C N=2014	Form W N=1974
Female	49.7	50.3	48.8	51.3	50.7	50.7
Male	49.7	49.3	50.8	48.1	49.0	48.9
Gender Not Specified	0.4	0.4	0.5	0.5	0.3	0.4
African American	28.8	27.9	28.5	27.9	28.7	26.4
American Indian	0.3	0.3	0.3	0.1	0.4	0.3
Asian	2.5	2.3	2.3	2.4	2.1	2.5
Hispanic	2.2	1. 6	1.9	1.3	1.4	1.7
White	65.0	66.9	66.2	66.9	66.7	68.1
Other Ethnicity	1.1	0.9	0.8	1.3	0.7	0.9
Accommodated	-[2]	1.6	3.1	1.2	1.2	0.7
Eng Lang Learner	-[2]	0.3	0.3	0.2	0.2	0.5
Special Education	-[2]	10.9	16.7	9.0	9.0	7.9

Note. [1]. Data reported in this column is based on all students completing the January administration.
[2]. Information not included in the CTB/McGraw-Hill Technical Report

Table 3.A.8. Characteristics of Calibration Samples by Form for May 2003

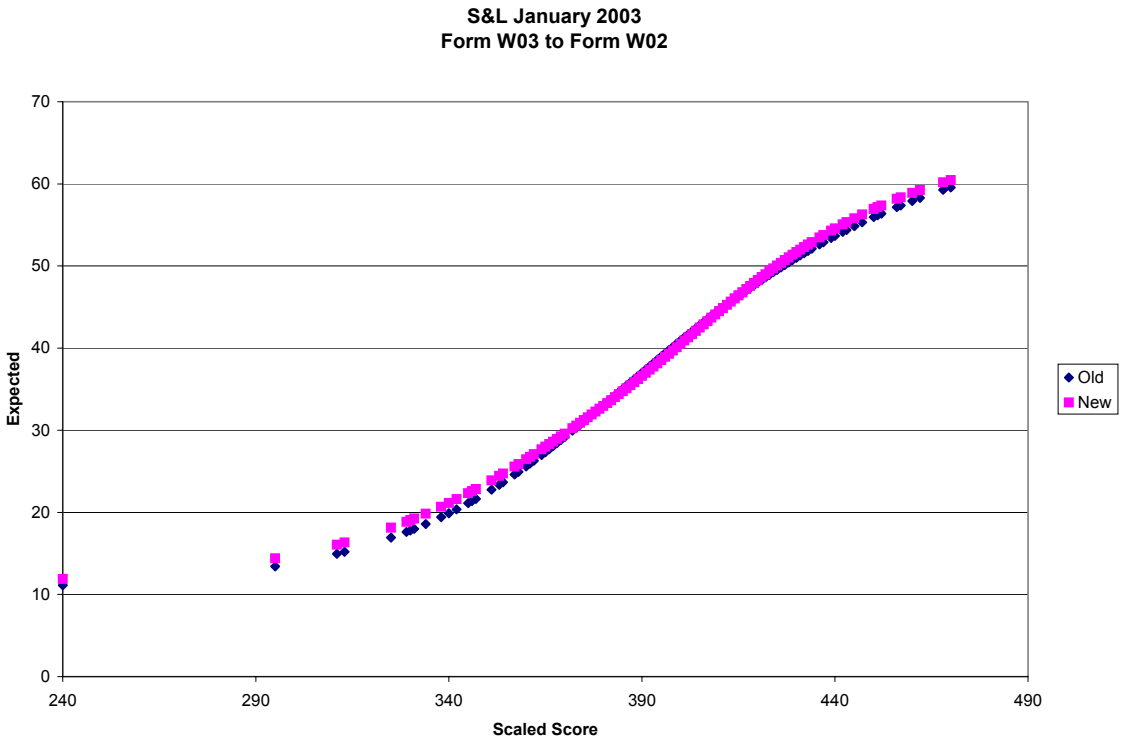
	CTB	Replication											
	Total N=56914	Total N=52549	Form D N=5827	Form E N=4799	Form F N=4807	Form G N=4773	Form H N=4770	Form J N=4712	Form K N=4672	Form L N=4599	Form M N=4600	Form N N=4507	Form P N=4483
Female	49.0	49.6	46.2	50.0	49.1	48.5	49.7	51.0	51.2	50.5	48.9	49.8	51.5
Male	50.1	50.0	53.1	49.6	50.5	51.1	50.0	48.4	48.5	49.2	50.6	49.8	48.0
Gender Unspecified	0.9	0.4	0.7	0.4	0.4	0.4	0.3	0.6	0.3	0.3	0.5	0.4	0.4
African American	35.7	35.0	36.2	35.2	34.8	35.8	35.5	34.6	35.2	34.8	34.5	34.5	34.0
American Indian	0.4	0.3	0.3	0.2	0.3	0.3	0.4	0.4	0.5	0.3	0.3	0.5	0.2
Asian	5.3	5.6	4.6	5.1	5.8	5.7	5.8	6.3	6.0	5.7	5.3	5.2	6.2
Hispanic	4.9	4.9	4.8	5.0	5.0	4.7	5.2	4.9	4.5	4.9	4.9	4.9	5.2
White	52.4	53.5	53.0	53.9	53.4	52.8	52.6	52.9	53.3	53.7	54.5	54.2	53.7
Other Ethnicity	1.2	0.7	1.0	0.7	0.7	0.7	0.6	0.8	0.5	0.6	0.6	0.7	0.6
Accom-modations	-[1]	1.6	3.6	1.1	1.5	1.6	1.3	1.4	1.4	1.2	1.2	1.3	1.5
Eng Lang Learner	-[1]	1.2	1.0	1.3	1.3	1.1	1.3	1.2	1.1	1.3	1.1	1.2	1.2
Special Education	-[1]	10.2	21.8	9.1	8.8	8.6	8.3	8.5	8.5	8.1	8.4	8.9	8.2

Note. [1] Data reported in this column is based on all students completing the January administration.
[2] Information not included in the CTB/McGraw-Hill Technical Report

January Results

As described earlier, a single calibration was completed for the January sample using Multilog. Forms A-C were then linked to the 2002 scale via the Form W item parameters in a Stocking and Lord procedure. As observed in Figure 3.A.1, differences in the test characteristic curves for Form W 2002 (old) and Form W 2003 (new) were noted at the lower end of the scale. This is related to differences between Pardux and Multilog in how the C-parameter is estimated. While many of the 2002 parameters had an estimated value of zero, non-zero estimates were obtained for the 2003 parameters using Multilog.

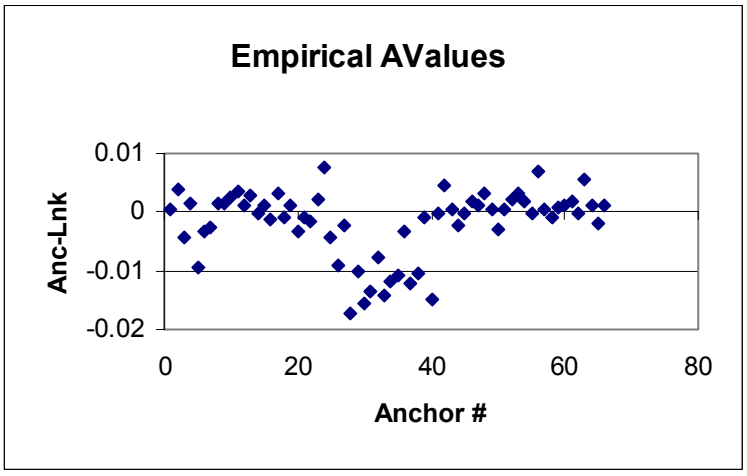
Figure 3.A.1.

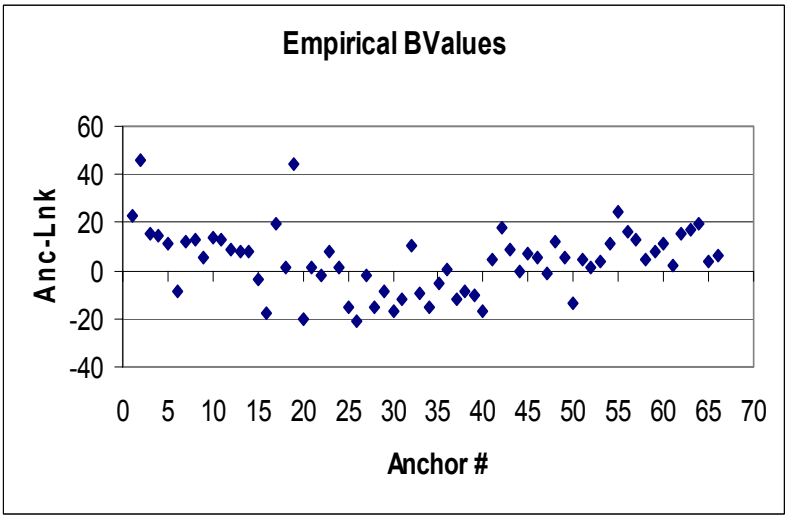


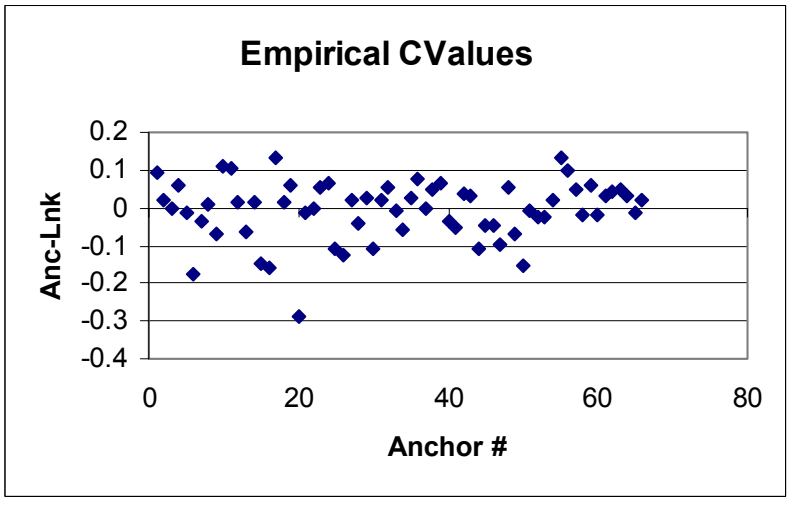
The differences in the 2003 A-, B- and C-parameters from our replication compared to 2002 were plotted in Figures 3.A.2 to 3.A.4.

Appendix 3.A

Figures 3.A.2 to 3.A.4. Differences in Item Parameter Values Compared to 2002.







Following CTB/McGraw-Hill's procedure, the Form W Stocking and Lord equating constants (slope=32.8196; intercept=393.2301) were then applied to all items in Forms A-C. Item-pattern scale scores were produced using the transformed parameters for Forms A-C and the 2002 parameters for Form W. Summary statistics are presented in Table 3.A.9.

Table 3.A.9. Descriptive Statistics January 2003 after Stocking and Lord

		CTB/McGraw-Hill		Replication		
Form	N	Mean	SD	N	Mean	SD
A	2370	380.8	45.8	2364	387.8	39.0
B	2090	387.6	41.7	2084	393.3	35.8
C	2019	386.5	41.6	2014	392.7	35.1
W	1986	395.5	34.4	1974	395.4	34.4

Following this transformation, a linear approximation to equipercentile equating was completed between Form C and Form W. The resulting transformation constants (slope=0.98302; intercept=10.5388) were then applied to Forms A, B, and C. Summary statistics are presented in Table 3.A.10. The additional transformation resulted in mean scores that were within one scale score point of the results reported in the Draft Technical Document (CTB/McGraw-Hill, December, 2003). In all cases, the replicated scores were higher, although the sample sizes were slightly different, which may account for the discrepancies.

Table 3.A.10. Descriptive Statistics January 2003 after Linear Equipercentile

		CTB/McGraw-Hill		Replication		
Form	N	Mean	SD		Mean	SD
A	2370	390.8	38.1	2364	391.8	38.3
B	2090	396.4	34.6	2084	397.2	35.2
C	2019	395.5	34.6	2014	396.5	34.5
W	1986	395.5	34.4	1974	395.4	34.4

After reviewing the design used by CTB/McGraw-Hill and noting that an extra step (Form W Stocking and Lord) had been used (see Footnote 2), we determined that the forms could have been placed onto the operational scale using only the linear approximation to equipercentile equating. As part of this study, we compare the results of the two-step linking to a single-step linking design. Not unexpectedly, the results were very similar (see Table 3.A.11).

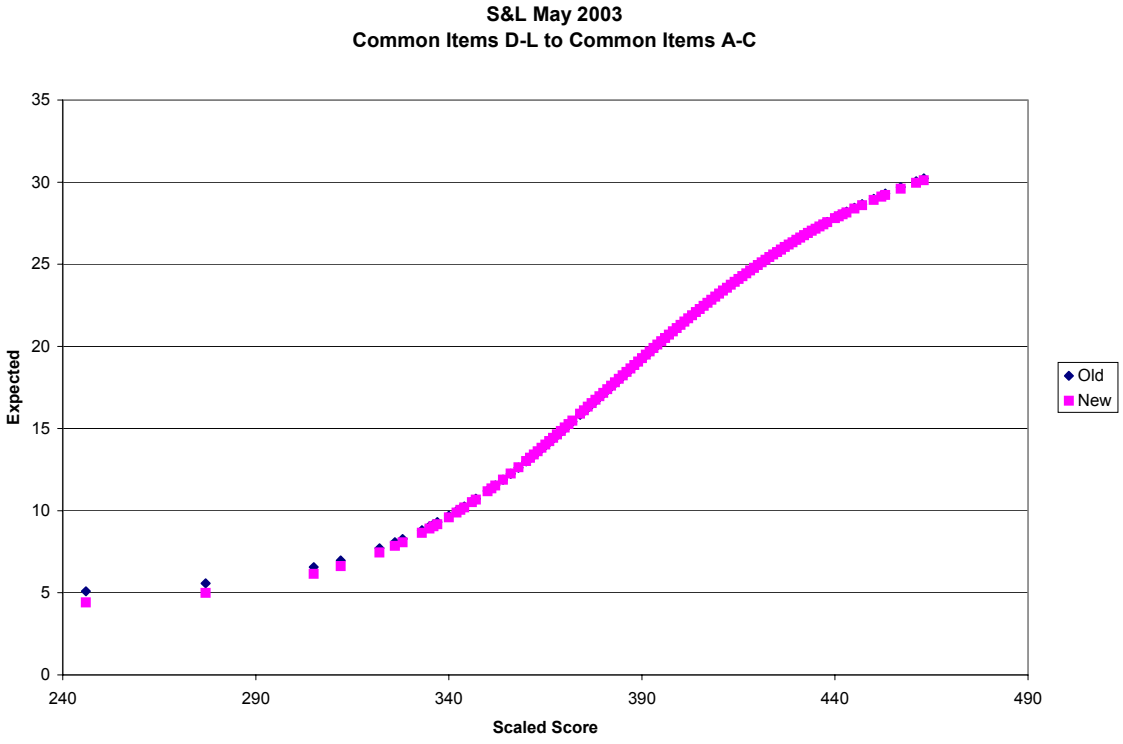
Table 3.A.11. Descriptive Statistics January 2003 Omitting Form W S&L Link

		CTB/McGraw-Hill		Replication Omitting Form W S&L Link		
Form	N	Mean	SD	N	Mean	SD
A	2370	390.8	38.1	2364	391.4	39.2
B	2090	396.4	34.6	2084	396.9	35.9
C	2019	395.5	34.6	2014	396.5	34.4
W	1986	395.5	34.4	1974	395.4	34.4

May Results

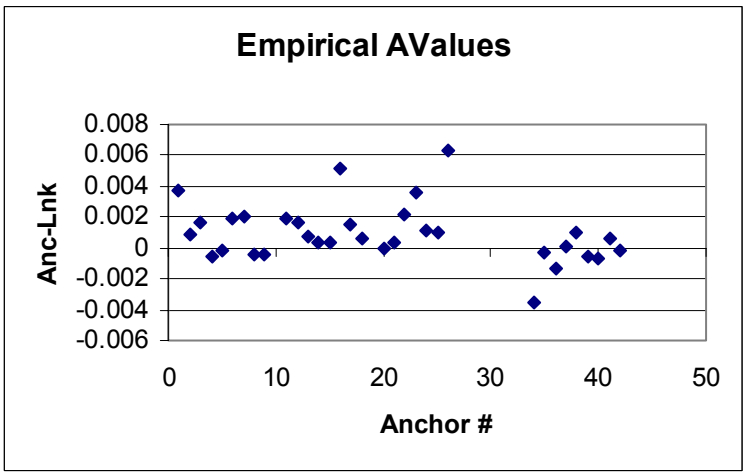
Following CTB/McGraw-Hill's procedure, the May 2003 forms were concurrently calibrated using Multilog. Forms D-L were placed onto the operational scale through the common item set shared between forms A-C (old) and Forms D-L (new) via a Stocking and Lord linking procedure. As observed in Figure 3.A.5, there were almost no differences in the test characteristic curves.

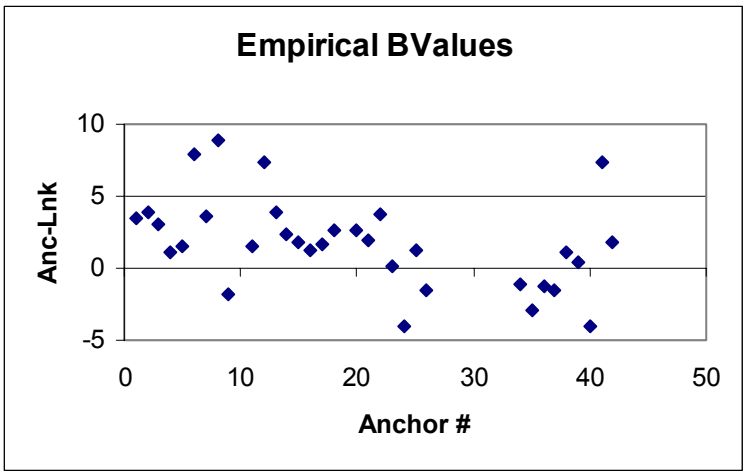
Figure 3.A.5.

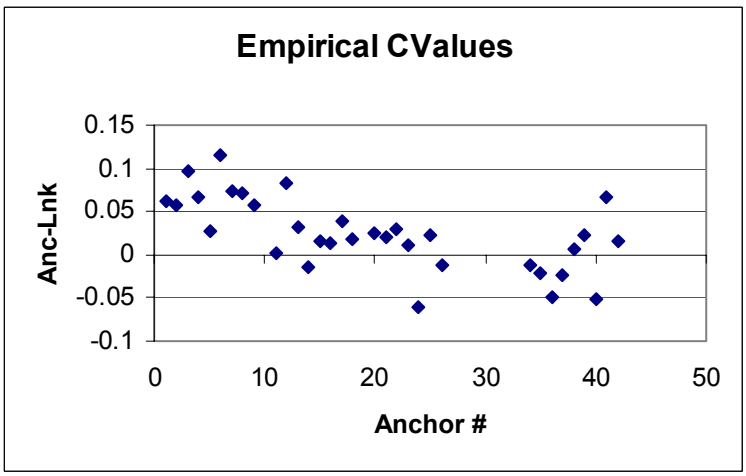


The differences in the January 2003 A-, B- and C-parameters compared to May 2003 are plotted in Figures 3.A.6 to 3.A.8.

Figures 3.A.6 – 3.A.8. Differences in Anchor Item Parameter Values: Forms A-C compared to Forms D-L.







The resulting equating constants (slope=33.13; intercept=399.5) were then applied to items in Forms D-P and item-pattern scale scores produced. Summary statistics are presented in Table 3.A.12. As observed in the January 2003 forms, the resulting means and standard deviations were very similar to the results reported in the Draft Technical Document (CTB/McGraw-Hill, December, 2003). In Forms D-L, the mean scores were approximately one scale score point higher.

Table 3.A.12. Summary Statistics May 2003 After Stocking and Lord

		CTB/McGraw-Hill		Replication		
Form	N	Mean	SD	N	Mean	SD
D	5831	390.3	39.8	5827	391.3	39.4
E	4797	397.7	34.5	4799	398.7	35.0
F	4806	398.0	34.8	4807	399.0	35.3
G	4772	397.1	34.3	4773	398.1	34.6
H	4775	397.5	35.5	4770	398.6	35.7
J	4720	397.9	35.5	4712	399.0	35.7
K	4673	399.4	34.4	4672	400.5	34.3
L	4600	398.8	36.2	4599	400.0	36.4
M	4596	388.1	38.7	4600	397.0	36.9
N	4508	393.0	36.9	4507	390.9	38.0
P	4483	395.3	37.7	4483	392.2	39.0

Because Forms M, N, and P shared no common items with Forms A-L or W, these forms were placed onto the operational scale using a linear approximation to equipercentile equating. Like the January analyses, the Stocking and Lord transformation constants were applied prior to completing the linear equipercentile equating. The descriptive statistics associated with these forms are presented in Table 3.A.13. The mean scores for these forms were very similar to Form L and slightly higher than the results obtained by CTB/McGraw-Hill.

Table 3.A.13. Descriptive Statistics January 2003 after Linear Equipercentile

	CTB/McGraw-Hill			Replication		
Form	N	Mean	SD	N	Mean	SD
M	4596	398.7	36.7	4600	401.3	34.9
N	4508	398.7	36.7	4507	402.6	34.4
P	4483	398.9	35.9	4483	402.8	33.9

Conclusions & Implications

Based on the results of this study, we found no evidence of a systematic error or problem with the calibrations and linking studies completed by CTB/McGraw-Hill. Using independent software, we were able to replicate the results. Small differences were noted in the parameter estimates, transformation constants, and mean scores; however, this is to be expected due to variations associated with inclusion/exclusion criteria for the calibration sample, and differences in the calibration software.

Several observations can be made. First, unless there were strict administration controls, it is very difficult to ensure that forms were be spiraled to randomly equivalent groups. For a variety of reasons, the spiral may have failed (e.g., seating assignments, re-ordering of forms by test administrators, etc.). In this study, the groups were very similar on all demographic variables except students classified as Special Education. A disproportionate number of these students were administered the first form in each administration. While the first of the May forms included relatively more special education students than the first January form, this did not affect the May equating. This is because the May forms were concurrently calibrated and linked using a common anchor set and the equating did not depend on the assumption of randomly equivalent groups[6]. If Forms A-C did not share a common item set, these forms could not have been placed onto the operational scale. Therefore, when randomly equivalent groups cannot be assured, it is prudent to always include common items across forms that can serve as an anchor set.

Second, the Stocking and Lord linking for Form W completed prior to the linear approximation to equipercentile equating was unnecessary. Completing two linking procedures only complicates the design. Essentially, this procedure would have produced similar final results to a single-step procedure. It appeared that the extra step was conducted by CTB/McGraw-Hill for January because it was not until after the Form W Stocking and Lord procedure was implemented that it was seen that this procedure was not sufficient. It is unclear why the two-step procedure was also implemented for the May forms.

Third, with a single cut-score near the middle of a score distribution, a relatively small difference in student scale scores can result in noticeable differences in percents of proficient students. Legitimate equating procedures can produce small variations in scale scores, which can make a noticeable difference in performance classifications.

[6] May forms without common items were linked using linear approximation to equipercentile equating.