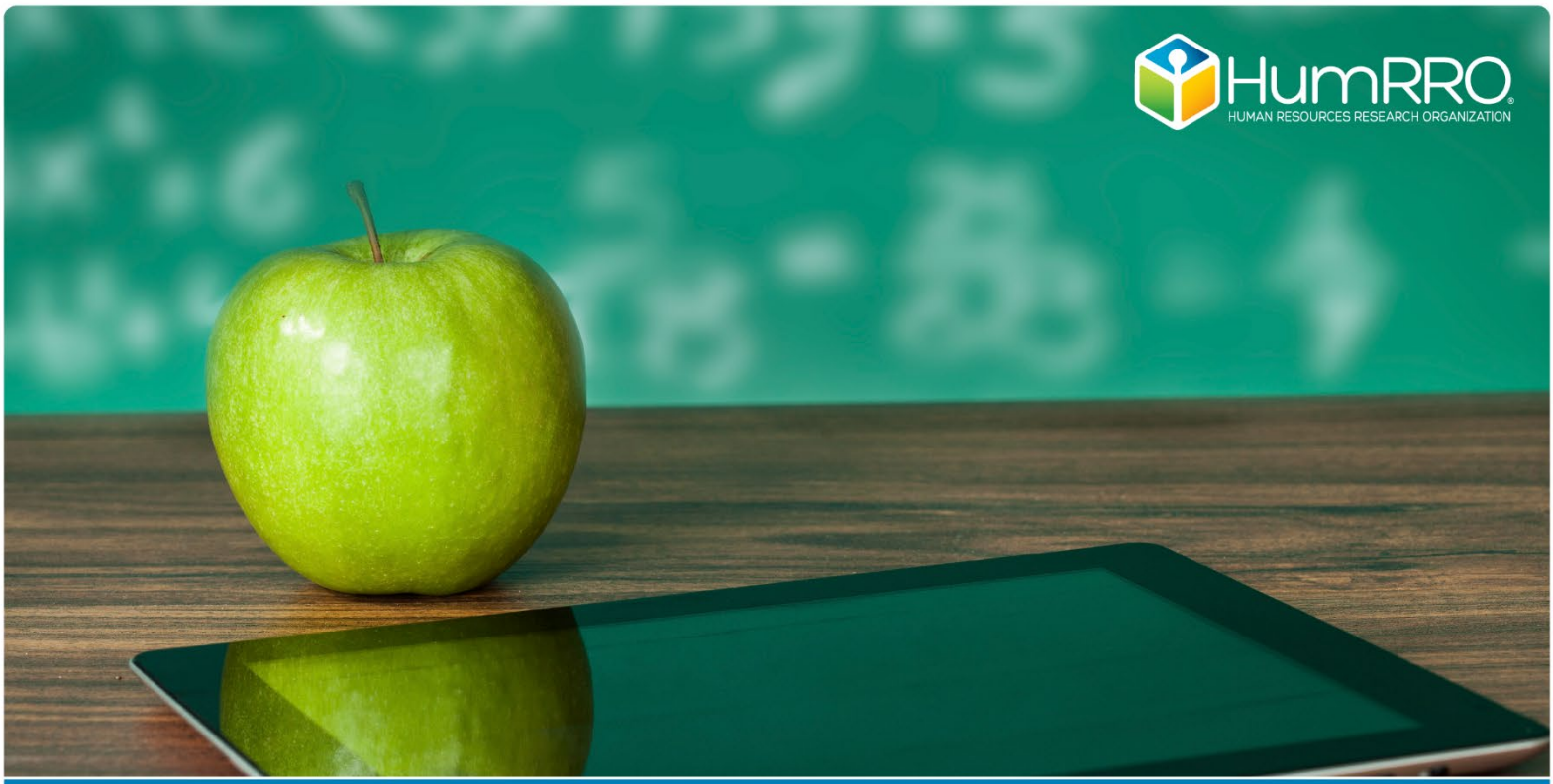

HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

2018 No. 048

PARCC ELA Pre-Equating Study

Prepared for:	New Meridian Corporation 5800 Trailridge Dr. Austin, TX 78731
Prepared under:	New Meridian Corporation 5800 Trailridge Dr. Austin, TX 78731
Authors:	Erin S. Banjanovic Lucy Chen Arthur Thacker
Date:	September 30th, 2018

Executive Summary

One of the key features of the PARCC assessments is the ability to compare student performance from one year to the next. Ensuring year-to-year comparability of test scores is accomplished through a process called equating. The PARCC assessments were originally designed to use post-equating to derive student scores. This process uses student data from the *current year* to estimate information about item and test function (using item response theory), equate the current test to previous tests, and generate student scores. This approach provides the best estimates of current student ability, but it occurs after test administration and takes considerable time. An alternative approach is known as pre-equating. Pre-equating uses estimates of item and test function from *earlier test administrations* to generate student scores. It allows scoring tables to be constructed prior to test administration and permits automatic score reporting. However, pre-equating does not take into account any changes to item or test function that could occur between test administrations (e.g., changes in curriculum, item position on the test, student motivation); which could impact student scores.

The Human Resources Research Organization (HumRRO) was retained by the New Meridian Corporation to investigate the impact of transitioning the PARCC ELA assessments to a pre-equated test design. A primary concern with pre-equating these assessments has been student motivation on field test items.[1] The current study attempted to address this issue by removing students who may have motivation issues from field test analyses. The study then examined the accuracy of the pre-equated solution through comparisons to the post-equated item parameters and score tables and through comparisons to results from a study that led to the pre-equating of the PARCC Math assessments.

Key findings from this study are presented below.

- The pre-equated item parameters were slightly more discriminating, but no more or less difficult than the post-equated parameters. The lack of consistent differences in item difficult suggest the exclusion rules addressed the student motivation issues.

- A large number of items (21.4%-52.9% across the grades) were flagged for item drift when comparing the pre- and post-equated parameter estimates. These proportions were larger than those commonly observed during post-equating.

- Some students were assigned different scale scores and performance levels based on the pre- vs post-equated scoring tables created from the reduced forms (only including the subset of items field tested in 2017). However, the score differences ranged from no effect to a small magnitude effect (Cohen's $d \leq 0.22$) for 95.4% to 100% of the students across the grades. When the score tables were created from the full forms, 99.3% to 100% of students experienced score differences of no effect to a small magnitude effect.

- Item drift was found to drive differences in test function for some grades. The magnitude of the item drift varied considerably across the items flagged for drift and large amounts of drift contributed to differences in student scores for some test forms, but not all test forms.

- Pre-equated scores in Math grade 7 (the only grade with results available for comparison) were generally within 2 score points of the equivalent post-equated scores; and pre-equated ELA scores were generally within 5 score points. ELA exhibited slightly larger differences than Math, but they were reasonably close.

Overall, the item drift results are the most concerning when considering a transition to pre-equating; but they are not surprising. We know item drift occurs and it impacts both item and test function. In a pre-equated context, drift will also impact student scores. Thus, the current study aimed to examine the level of drift that occurs and the impact on student scores. Although a large number of items were flagged for drift, the majority of pre- and post-equated test characteristic curves (TCCs) showed minimal differences. Furthermore, the difference between the pre-equated and post-equated scores for almost all students were of a small magnitude (within 0.22 scale score standard deviations) or were not different at all. We know post-equating provides the *best* estimates of current student ability, but the pre-equated estimates were reasonably close to post-equated scores. Together, these findings suggest item drift from pre-equated field test parameters does not have a detrimental impact on PARCC ELA student scores and that the pre-equated scores are comparable to post-equated scores.

If the PARCC ELA assessments are transitioned to a pre-equated scoring design, item drift analyses will play an important role in maintaining the item bank. Item drift of all operational items will need to be examined after test administration and items flagged for drift should be reviewed by a content specialist to evaluate whether there may be a reason for item drift. Recalibration of the item parameters should also be considered using student data from the operational administration. The drifted item parameters would not be updated for students who were administered the item, but they could be updated for students who will see the item in the future.

PARCC ELA Pre-Equating Study

Table of Contents

List of Tables

Table of Contents (Continued)

List of Figures

PARCC ELA Pre-Equating Study

Introduction and Purpose

The Partnership for Assessment of Readiness for College and Careers (PARCC) assessment system is a summative testing program offered in Math and English Language Arts/Literacy (ELA), for grades 3-8 and high school. The assessment system was created in 2015 by a consortium of states as an innovative, new-era assessment that is aligned to the Common Core State Standards (CCSS).

One of the key features of the PARCC assessments is the ability to compare student performance from one year to the next. This allows states to determine, for example, if fourth grade students' math scores are improving or declining at the state, district or school level and it allows states to implement education accountability policies. Ensuring year-to-year comparability of test scores is accomplished through a process called equating. The PARCC assessments were originally designed to use a psychometric process known as post-equating to derive student scores. This process uses student data from the *current year* to estimate information about item and test function (using item response theory), equate the current test to previous tests by comparing student performance on a repeated set of test items, and generate student scores. This approach provides the best estimates of current student ability, but it takes considerable time to generate student scores since the scoring activities must occur after test administration. An alternative approach to scoring is known as pre-equating. Pre-equating uses estimates of item and test function from *earlier test administrations* to generate student scores. It allows scoring tables to be constructed prior to test administration and permits automatic score reporting. However, it does not take into account any changes to item or test function that could occur between test administrations. For example, changes in curriculum, item position on the test, student motivation, testing conditions, or the general ability of the group taking the assessment could all impact item and test function, and subsequently, student scores.

Timely reporting of student scores is becoming an increasing priority for state assessment programs, leading to an increase in the demand for pre-equated assessment systems. In the summer of 2016, Pearson conducted a study to examine the impact of transitioning the Math assessment to a pre-equated test design. The results from the study supported the transition to pre-equating and the Math tests were transitioned during the spring 2017 administration. Pearson also conducted some preliminary investigations of the ELA assessments and concluded they would be more difficult to transition because the field test items might be impacted by student motivation issues. The ELA assessments have undergone some test design changes to try to address this issue, but these changes are not yet fully implemented.

In the spring of 2018, the Human Resources Research Organization (HumRRO) was retained by the New Meridian Corporation to conduct a study to examine the impact of transitioning to pre-equating on the PARCC ELA assessment and students. This study aimed to address the motivation issues by removing unmotivated students from the field tests analyses. The study then investigated the impact of pre-equating by comparing the item parameters and score tables produced under a pre-equated design to those produced under a post-equated design. The current report details the methods, results, and conclusions from this study.

Methods

Data from the 2017 and 2018 ELA operational test administrations were used in this study. The procedures used to estimate the pre- and post-equated estimates are described below. This is

followed by a description of the analyses used to compare the pre- vs post-equated parameters and scores.

Post-Equated Parameter Estimates

The post-equated parameter estimates used in the current study are from the 2018 post-equating analyses of the operational items. The analyses occurred in May and June 2018 and used data from an early return of the spring 2018 student results (thus the student data included a representative sample[1] of the final spring 2018 student administration). Pearson served as the primary analyst and HumRRO provided third-party replication of the analyses. The ELA items were calibrated using a generalized partial credit model and a single-calibration approach (known as the Moulder method) was used to address issues of item dependence among the two traits of prose constructed response (PCR) items.[2] We used a common-item equating design, where the anchor item (common item) set included items that were administered operationally in 2018 and had also been administered operationally during a prior test administration. We used Robust- z and weighted root mean squared difference (WRMSD) to identify items that drifted and needed to be removed from the anchor item set. The final anchor item set was then used to estimate Stocking-Lord equating constants and place the items on the operational scale. A summary of the students included in the analyses, by state and grade, is presented in Table 1. A summary of the number of operational items, anchor items, drifted items, and final items included in equating are presented in Table 2. Please note, not all items flagged for drift are removed from equating. If more than 20% of the anchor items are flagged for drift, then a set of criteria are followed to force items with less drift back into the final anchor set until the proportion of removed items is approximately 20%. The two instances where this occurred in the spring of 2018 are presented in bold.

Pre-Equated Parameter Estimates

The pre-equated parameter estimates used in the current study are from the spring 2017 field test analyses. The analyses were originally conducted by Pearson during the summer of 2017 and used the full sample of students who were administered the assessment (not a sample, like in the operational analyses). Similar to the operational analyses, the items were calibrated using a generalized partial credit model; but because fewer students took the field test items, the field test PCR items were calibrated in two separate calibrations. The first calibration included all of the non-PCR items and the Reading Comprehension and Written Expression (RCWE) trait of the PCR items. The second calibration included all of the non-PCR items and the Written Knowledge and Language (WKL) trait of the PCR items. The parameters for the non-PCR items and the RCWE trait of the PCR items were taken from the first calibration and the parameters for the WKL trait of the PCR items were taken from the second. If any item had an insufficient number of observations to calibrate a particular item score category, Pearson imputed student data for the missing item score category to calibrate the item as intended.[3] This practice was only implemented during the field test analyses due to the reduced sample size. Common-person equating was then used to put the field test items onto the operational scale. Stocking-Lord equating constants were estimated from the operational items and then applied to the field

[1] Due to time constraints, post-equating must begin before the testing window has closed. Thus, Pearson monitors the representativeness of the submitted student assessment records and provides files for post-equating once pre-determined criteria have been met.
[2] A special study was conducted by ETS in 2015 to investigate calibration of PCR items. The results led to the use of the Moulder method for the PARCC ELA assessment.
[3] For example, for an item scored on a scale from 0 to 3, if a score of 3 was not observed across the student data, some imputed student records would be added to the student data that included a score of 3 for the item.

test items. The field test parameters were used solely for the purpose of form construction and then operational estimates for the items were derived during post-equating.

Table 1. Students Included in the 2018 OP Analyses, by Grade and State

	State or Territory							
Grade	BI	DC	DD	IL	MD	NJ	NM	Total
Count								
03	323	1,581	0	94,820	34,427	55,499	17,657	204,307
04	374	5,393	0	97,727	53,695	76,401	24,553	258,143
05	365	2,023	0	124,340	39,500	57,446	19,312	242,986
06	367	605	1,696	116,319	36,417	51,825	9,413	216,642
07	333	1,142	2,009	126,855	37,358	56,875	11,534	236,106
08	354	1,662	2,671	129,561	17,500	27,695	19,420	198,863
09	56	3,155	0	8	1,798	77,962	22,403	105,382
10	48	3,757	2,701	11	33,642	75,764	22,150	138,073
11	181	47	0	0	8,533	55,725	20,852	85,338
Percent								
03	0.16%	0.77%	0.00%	46.41%	16.85%	27.16%	8.64%	--
04	0.14%	2.09%	0.00%	37.86%	20.80%	29.60%	9.51%	--
05	0.15%	0.83%	0.00%	51.17%	16.26%	23.64%	7.95%	--
06	0.17%	0.28%	0.78%	53.69%	16.81%	23.92%	4.34%	--
07	0.14%	0.48%	0.85%	53.73%	15.82%	24.09%	4.89%	--
08	0.18%	0.84%	1.34%	65.15%	8.80%	13.93%	9.77%	--
09	0.05%	2.99%	0.00%	0.01%	1.71%	73.98%	21.26%	--
10	0.03%	2.72%	1.96%	0.01%	24.37%	54.87%	16.04%	--
11	0.21%	0.06%	0.00%	0.00%	10.00%	65.30%	24.43%	--

Note. BI = Bureau of Indian Education; DC = District of Columbia; DD = Department of Defense Education Agency; IL = Illinois; MD = Maryland; NJ = New Jersey; and NM = New Mexico.

Table 2. Items Included in the 2018 OP Analyses, by Grade

	Number of:			
Grade	Operational Items	Anchor Items	Anchor Items Flagged for Drift	Anchor Items Removed from Equating
03	47	21	4	4
04	59	18	3	3
05	56	22	0	0
06	60	21	2	2
07	61	21	**5**	**3**
08	58	20	3	3
09	63	24	4	4
10	60	20	2	2
11	58	22	**9**	**5**

Note. Bolded counts represent instances where anchor items that flagged for drift had to be forced back into equating to meet the criteria of dropping no more than 20% of the anchor item set.

HumRRO recalibrated the spring 2017 ELA field test items to remove the impact of unmotivated students. Overall, HumRRO used the same procedures employed by Pearson for the analyses, with two major exceptions. First, HumRRO included three additional exclusion criteria to remove students who were unmotivated to take the field test unit. The three rules, along with the number of students removed in each grade, are summarized in Table 3.

Table 3. Students Removed from 2017 FT Calibration Sample

	Students Removed Due to:				
Grade	Rules in the Specifications (e.g., incomplete test, non-English test)	Motivation Rule 1: Students who were administered the FT section but did not respond	Motivation Rule 2: Students who respond to hand-scored items on the OP sections but only machine-scored items on the FT section	Motivation Rule 3: Students with scores below chance on the FT section and scores 1 SD above chance on the OP sections [A]	Final Sample
03	11,563	416	63	2,948	350,101
04	12,718	515	53	2,079	360,559
05	13,728	463	40	1,574	386,136
06	13,967	3,063	104	861	374,889
07	13,571	1,996	150	3,349	375,521
08	13,833	2,789	327	1,632	371,605
09	7,317	1,899	1,143	1,302	177,318
10	8,006	747	1,777	1,438	162,850
11	7,925	496	2,580	893	100,233

Note. OP = Operational and FT = field test.
[a] Only multiple-choice items were included in the estimates of operational and field test scores. Chance was identified as 25%.

The additional exclusion criteria led to the field test sample sizes summarized in Table 4. Please note, the size of the field test sample was influenced by item type. The majority of field test items are multiple-choice or technology-enhanced items and are machine-scored; but the rest of the items are open-ended (the PCR items) and must be hand-scored. Hand-scoring is only done for a subset of the field test sample, thereby further reducing the sample size for the calibration of open-ended items. Overall, these samples were still sufficiently large for calibration using a generalized partial credit model.

Table 4. Sample Size for 2017 FT Calibration: FT Items

	Machine-Scored Items				Hand-Scored Items			
		Sample N				Sample N		
Grade	Item N	Mean	Min	Max	Item N	Mean	Min	Max
03	98	4,585	3,906	8,781	24	1,393	1,214	1,464
04	94	5,203	4,176	9,328	24	1,410	1,353	1,460
05	98	5,173	4,528	9,559	24	1,431	1,351	1,464
06	117	5,358	4,528	9,320	24	1,463	1,454	1,475
07	114	5,014	3,954	8,962	24	1,391	1,294	1,480
08	117	4,766	3,949	8,328	24	1,427	1,378	1,455
09	177	2,089	1,693	3,821	36	1,514	1,438	1,571
10	177	2,319	1,827	4,195	36	1,429	1,271	1,471
11	182	1,415	1,187	2,660	36	1,259	1,175	1,358

The second deviation from Pearson's field test procedures concerned instances where an item score category was not sufficiently observed to calibrate a corresponding step parameter.[4] Table 5 summarizes the instances where this issue occurred. In these instances, HumRRO collapsed the score category with the next lowest category to calibrate the item. This differed

[4] As mentioned previously, this issue has only arisen during field test analyses. This is largely due to the reduced sample sizes for the hand-scored, PCR items. In 2018, Pearson increased the number of hand-scored field test items from approximately 1,500 to 3,000, so this may not be an issue in subsequent field test calibrations.

from Pearson's field test procedure of imputing student data in these instances. Imputing data to calibrate the items as intended is acceptable in a post-equating context where the estimates are only used to construct test forms and not used to score students. However, this practice is not recommended in a pre-equating context where student scores could be derived based on parameters from imputed data. None of the items identified in Table 5 were used operationally in 2018; thus, the calibration issues for these items did not impact the current study results.

Table 5. FT Items with Collapsed Score Categories

Grade	Item UIN	Valid Scores	Score Point Not Observed
03	Item_001	0-3	3
	Item_002	0-3	3
05	Item_003	0-4	4
	Item_004	0-4	4
06	Item_005	0-4	4
	Item_006	0-4	4
07	Item_007	0-4	4
09	Item_008	0-4	4

Note. Item UIN has been modified for use in this report to mask proprietary information.

The final student sample included in the recalibration of the 2017 field test items is presented in Table 6. The results are presented by grade and state. Please note, a small sample of students from the Bureau of Indian Education (less than 350 per grade) were administered the ELA assessment in 2017 but are excluded from the recalibration. Data for this group of students was not delivered in time for inclusion in the current study. However, we do not believe the study results would change in any way if these students were included because they represented a very small proportion of the students and none were administered field test items.

Pre- vs Post-Equated Analyses

The pre- and post-equated parameters were used to investigate the impact of a transition from post-equating to pre-equating. Three primary sets of analyses were conducted to examine differences in item parameters and student scores and to compare these results to the prior study that investigated a similar transition for the PARCC Math tests. Each of the analyses are described in further detail below.

The first set of analyses focused on comparing item parameters. Only the subset of common items that were field tested in 2017 and then administered operationally in 2018 were used in these analyses. This was done to focus on the differences between the current pre-equating procedure (excluding unmotivated students) and post-equating. The differences in *a* and *b* parameters were examined to provide insight on the overall magnitude of changes in the parameters. Next, two estimates of item drift (robust-z and weighted root mean squared difference) were reviewed to identify items with significant changes in item function between the pre- and post-equated estimates.

Table 6. Students Included in the 2017 FT Analyses, by Grade and State

	State or Territory							
Grade	CO	DC	IL	MD	NJ	NM	RI	Total
Count								
03	58,252	5,795	103,216	57,451	92,153	23,843	9,391	350,101
04	59,223	5,809	105,032	63,550	92,956	24,215	9,774	360,559

05	59,310	5,145	132,520	63,580	91,501	23,916	10,164	386,136
06	55,648	4,551	133,107	56,952	90,985	23,752	9,894	374,889
07	54,744	4,223	135,497	55,720	92,161	22,886	10,290	375,521
08	52,094	4,190	131,687	58,041	92,866	22,642	10,085	371,605
09	45,892	3,337	17	1,828	93,052	23,083	10,109	177,318
10	0	4,121	14	53,931	82,668	22,116	0	162,850
11	7	63	0	14,850	65,199	20,114	0	100,233
Percent								
03	16.64%	1.66%	29.48%	16.41%	26.32%	6.81%	2.68%	--
04	16.43%	1.61%	29.13%	17.63%	25.78%	6.72%	2.71%	--
05	15.36%	1.33%	34.32%	16.47%	23.70%	6.19%	2.63%	--
06	14.84%	1.21%	35.51%	15.19%	24.27%	6.34%	2.64%	--
07	14.58%	1.12%	36.08%	14.84%	24.54%	6.09%	2.74%	--
08	14.02%	1.13%	35.44%	15.62%	24.99%	6.09%	2.71%	--
09	25.88%	1.88%	0.01%	1.03%	52.48%	13.02%	5.70%	--
10	0.00%	2.53%	0.01%	33.12%	50.76%	13.58%	0.00%	--
11	0.01%	0.06%	0.00%	14.82%	65.05%	20.07%	0.00%	--

Note. CO = Colorado; DC = District of Columbia; IL = Illinois; MD = Maryland; NJ = New Jersey; NM = New Mexico; and RI = Rhode Island.

The second set of analyses focused on comparing student scores based on reduced summative score conversion tables. The reduced summative score conversion tables were built from the subset of common items that were field tested in 2017 and then administered operationally in 2018. The pre-equated parameters (from the 2017 field test analyses) were used to create the pre-equated score conversion tables and the post-equated parameters (from the 2018 operational analyses) were used to create the post-equated score conversion tables. Creating scoring tables using the subset of common items permitted direct evaluation of the current pre-equating procedure (excluding unmotivated students) and did not confound the pre-equating results through the inclusion of potentially more stable item parameters derived from post-equating of a prior administration. The score conversion tables were then applied to the sample of students who were administered the ELA assessments in the spring of 2018, met the criteria for inclusion in the calibration analyses, and were administered a form that contained items that were field tested in 2017. Differences in pre- vs post-equated student proficiency level and scale score were examined to provide insight into the impact of pre-equating on estimates of student ability. Comparisons between the pre- and post-equated test characteristic curves (TCCs) were also made to understand whether differences in student ability estimates occurred at particular places along the ability scale.

Finally, the third set of analyses compared student scores based on the full summative score conversion tables. The pre-equated parameter estimates were taken from either the 2017 field test analyses or from the item bank and the post-equated parameters were estimated during the 2018 operational analyses. The inclusion of banked item-parameters (from post-equating of previous operational administrations) introduced greater stability into the item set and allowed comparisons of student scores based on intact operational forms that met the test blueprint. Similar to the previous analyses, two sets of conversion tables were created based on the item parameter sets and the tables were applied to a sample of students who were administered the assessment in the spring of 2018. Several comparisons were made between pre- and post-equated scores. Additionally, comparisons were made to results from the Math pre-equating study that was conducted by Pearson.

Results

The results for the three sets of analyses conducted to investigate the impact of transitioning the PARCC ELA assessments to a pre-equated design are presented in this section. A comparison of the pre- vs post-equated items is presented first, followed by a comparison of the pre- vs post-equated student scores. Finally, a comparison of the results for the current study to the results for the Math pre-equating study is presented.

Comparison of Pre- vs Post-Equated Items

Differences in Parameters

The pre- and post-equated item parameters were compared to provide insight into the magnitude of change in the IRT *a* and *b* parameters. Only the subset of items that were field tested in 2017 and used operationally in 2018 were included in these analyses to focus on the differences between the current pre-equating procedure (excluding unmotivated students) and post-equating. Table 7 summarizes the differences between the pre- and post-equated IRT *a* and *b* parameters. Overall, the pre-equated parameters were slightly more discriminating (*a* parameter) in eight of the nine grades, and more difficult (*b* parameter) in six of the grades. The average difference in item discrimination were at the second or third decimal place and were generally within ±.20. The average differences in item difficulty tended to be at the second decimal place (with the exception of grade 11) and ranged from -1.89 to 1.35. Although these differences in parameters are slightly larger than might be desired, they do not show systematically large differences in parameter estimates between pre- and post-equating.

Table 7. Summary of Pre- and Post-Equated IRT A and B Parameters Differences

		Difference in A Parameters (Post-Equated – Pre-Equated)			Difference in B Parameters (Post-Equated – Pre-Equated)		
Grade	Item N	Mean	Min	Max	Mean	Min	Max
03	14	-0.002	-0.124	0.120	-0.078	-0.854	0.340
04	30	-0.029	-0.182	0.096	0.073	-0.522	0.529
05	30	-0.036	-0.128	0.004	-0.012	-0.314	0.675
06	34	-0.024	-0.191	0.143	-0.060	-1.887	0.724
07	36	-0.045	-0.233	0.147	0.036	-0.345	0.846
08	29	0.003	-0.109	0.195	-0.042	-0.409	1.349
09	28	-0.060	-0.318	0.068	-0.044	-0.593	0.724
10	34	-0.025	-0.146	0.350	0.080	-0.355	0.924
11	34	-0.014	-0.142	0.200	-0.115	-0.904	0.239

To further understand the differences in item parameters, we reviewed the ability distribution of the two samples used to estimate the parameters. Table 8 displays the average scale score and the range of the scale scores in the 2017 and 2018 student calibration samples. The scale scores in the 2017 calibration sample tended to be slightly lower than those in the 2018 sample and also tended to have slightly smaller standard deviations. These minor differences in student ability were likely introduced by the additional exclusion rules applied to the 2017 calibration sample to exclude students who were not motivated. The third motivation rule removes students who had operational scores one standard deviation greater than chance, but field test scores below chance. This rule did not identify any students who had scores close to chance, thus very low, for removal. The slight systematic differences in item discrimination could be due to the differences in the ability distribution between the two samples.

Table 8. Scale Score Distribution of 2017 and 2018 Calibration Samples

	2017 Student Sample			2018 Student Sample			Differences (2018-2017)	
Grade	N	Mean	SD	N	Mean	SD	Mean	SD
03	350,101	739.49	40.72	283,422	740.39	42.37	0.90	1.64
04	360,559	743.67	35.67	288,664	746.11	37.16	2.44	1.49
05	386,136	743.83	34.62	316,347	743.90	35.01	0.07	0.39
06	374,889	742.09	31.61	319,291	743.17	33.12	1.08	1.52
07	375,521	744.29	38.03	317,510	746.57	39.78	2.28	1.75
08	371,605	743.18	38.84	319,455	744.62	40.12	1.44	1.28
09	177,318	742.02	37.81	123,922	747.90	39.31	5.88	1.51
10	162,850	742.46	46.79	183,939	745.86	48.32	3.40	1.52
11	100,233	736.16	39.55	99,904	737.55	40.50	1.38	0.95

Note. The scale scores are on a scale from 650 to 850 for each grade.

Scatterplots of the pre- vs post-equated IRT *a* and *b* parameter estimates were reviewed to provide insight into where the differences in discrimination and difficulty are occurring and whether particular items are impacted. The scatterplots are presented in Figure 1, with the parameter estimates color coded by item type (OE = open-ended; MX = composite item; and XI = technology-enhanced item). A line showing perfect fit is imposed on each plot to assist with the comparison. Most items were very close to the line, indicating reasonable correspondence between the pre- and post-equated estimates. The plots suggest the larger differences in *b* parameters occur for more difficult items, which can be expected since such items are harder to calibrate. The plots also show no particular item type experienced consistently greater differences, just that the open-ended response items tend to be more difficult. Overall, the plots further demonstrate the differences in IRT *a* parameters are relatively small (not noticeable in these plots) and that the differences in IRT *b* parameters are not consistent and can be explained by the larger bounce that can be found among more difficult items.

Parameter Drift

Two indicators of item parameter drift were used on the PARCC assessment to identify items that should be excluded from equating: robust-z and weighted root mean squared difference (WRMSD). These measures were used in the current study to identify items where the post-equated item parameters had drifted substantially from the pre-equated item parameters. The percent of items flagged by each measure is presented in Table 9. To assist with interpretation, the table also presents the percent of items flagged by these measures during normal operational analyses. The primary difference between these two sets of results is that the current study is comparing the 2018 post-equated parameter estimates to estimates derived from the 2017 field test analyses (the pre-equated parameters) and the operational analyses are comparing post-equated parameter estimates of the 2018 anchor items to earlier estimates derived from post-equating of these items. In other words, the current study is investigating drift from field test parameters and the operational analyses investigated drift from previous operational parameters. Different item sets were examined by the two sets of analyses and we would expect the field test parameters to drift slightly more than the operational items.

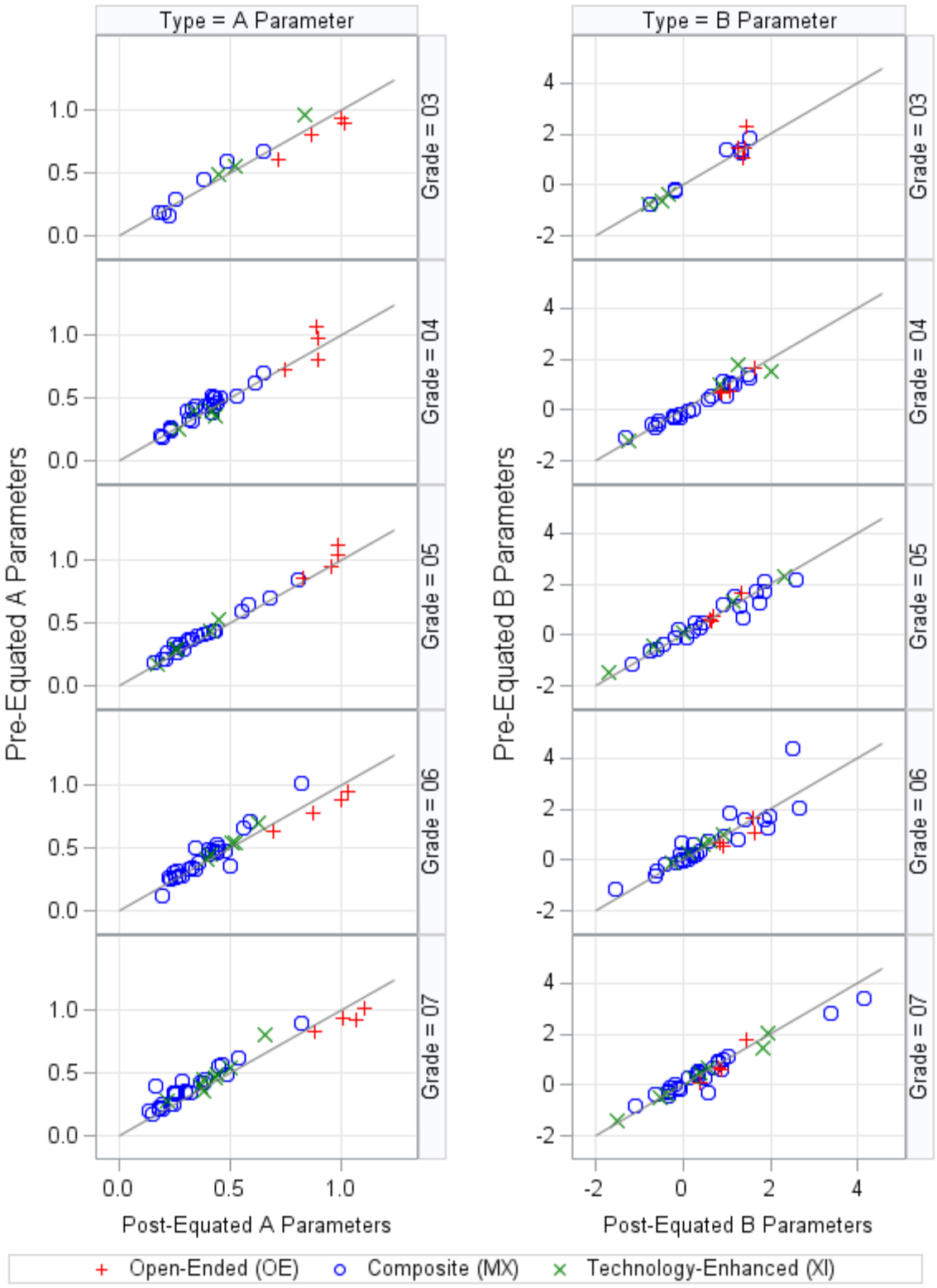

Figure 1. Scatterplots of pre- vs. post-equated A and B parameters.

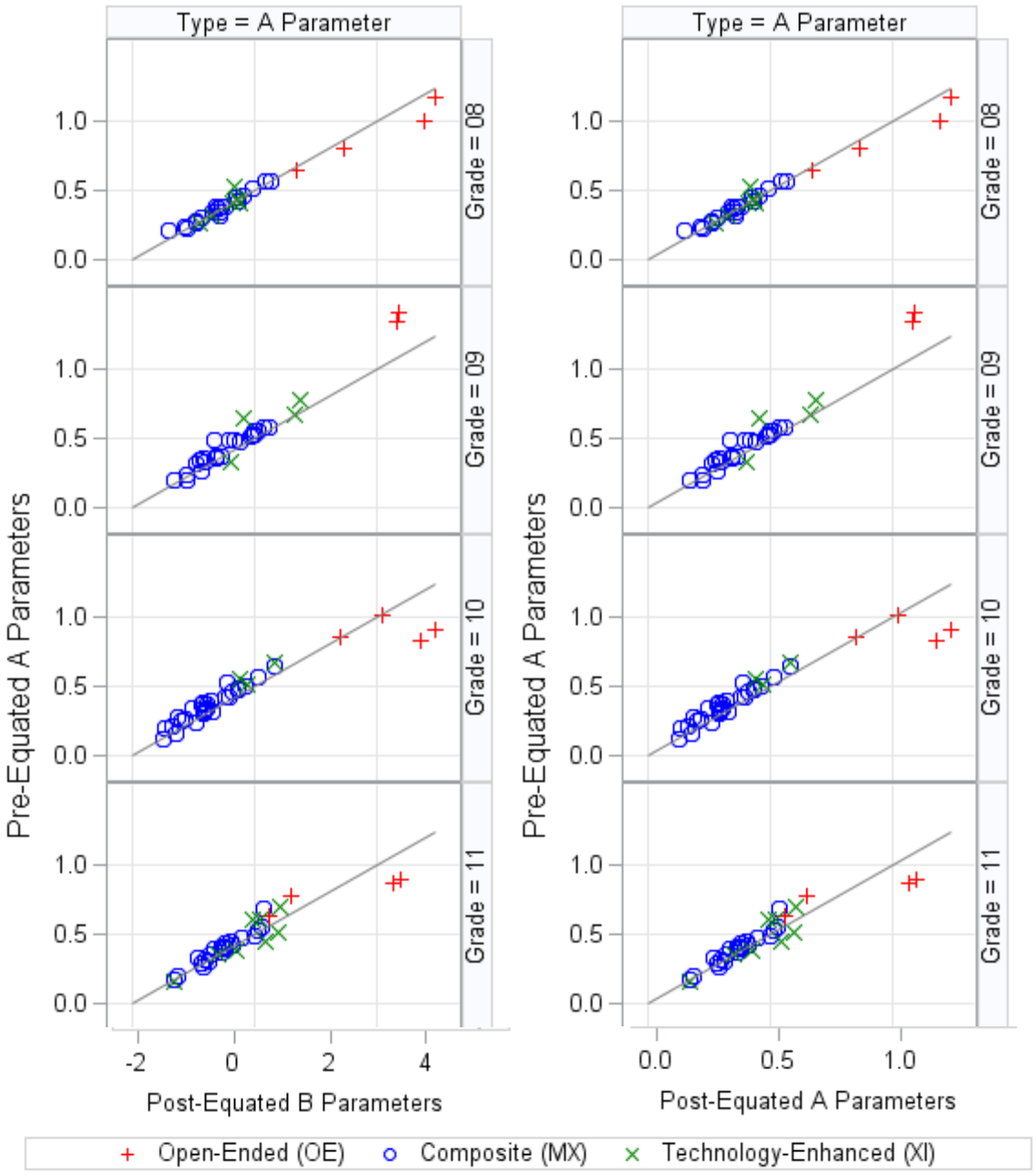

Figure 1. (Continued)

Table 9. Items Flagged for Drift: Pre-Equated Study Compared to 2018 Operational Analyses

	Current Study (Pre- vs Post-Equated Items)					2018 Operational Analyses (2018 OP items vs banked OP items)				
		% of Items Flagged by:					% of Items Flagged by:			
Grade	Common Item N	Rob.-Z: A Par.	Rob.-Z: B Par.	Rob.-Z: Total	WRMSD	Common Item N	Rob.-Z: A Par.	Rob.-Z: B Par.	Rob.-Z: Total	WRMSD
03	14	7.1	14.3	21.4	7.1	21	0.0	19.0	19.0	4.8
04	30	0.0	23.3	23.3	0.0	18	11.1	5.6	16.7	0.0
05	30	0.0	36.7	36.7	3.3	22	0.0	0.0	0.0	0.0
06	34	8.8	26.5	35.3	17.6	21	0.0	9.5	9.5	0.0
07	36	8.3	27.8	36.1	5.6	21	9.5	14.3	23.8	4.8
08	29	10.3	31.0	41.4	0.0	20	15.0	0.0	15.0	0.0
09	28	10.7	25.0	35.7	3.6	24	8.3	8.3	16.7	0.0
10	34	8.8	44.1	52.9	5.9	20	0.0	10.0	10.0	0.0
11	34	8.8	29.4	38.2	5.9	22	18.2	22.7	40.9	4.5
Mean		7.0	28.7	35.7	5.4		6.9	9.9	16.8	1.6
Minimum		0.0	14.3	21.4	0.0		0.0	0.0	0.0	0.0
Maximum		10.7	44.1	52.9	17.6		18.2	22.7	40.9	4.8

Note. All items are evaluated for robust-z differences in the *a* parameter, but only items not flagged on the *a* parameter are evaluated on the *b* parameter. Thus, the items flagged on a are mutually exclusive from the items flagged on *b*. Nearly all items that were flagged for drift on WRMSD were also flagged on robust-z. Only two items in the current study, one in grade 6 and another in grade 9, were flagged by WRMSD but not robust-z. In the 2018 operational analyses, all items that flagged for WRMSD also flagged for robust-z.

Table 9 shows that more items were flagged by robust-z than WRMSD and that the current study tended to flag more items than the 2018 operational analyses. However, there was a fair amount of variation in the results across the grade levels. The difference in the percentage of items flagged for robust-z ranges from -2.7% to 42.9% (e.g., for grade 10, 52.9%-10.0% = 42.9%). This means we flagged 2.7% fewer items to 42.9% more items than the 2018 anchor items, across the grades. The difference in the percentage of items flagged for WRMSD ranges from 0% to 17.6%. This means we flagged the same number of items to 17.6% more items, across the grades. While the difference is not large for all grades, it is quite large for some. On average, we flagged 18.9% more items on robust-z and 3.8% more items on WRMSD. This generally amounts to 5 more items flagging on robust-z in the current study and 1 more item flagging on WRMSD.

To further understand the magnitude of item drift, we examined scatterplots of the robust-z and WRMSD estimates. Figure 2 displays these plots, with green identifying drift estimates for items in the current study, blue identifying estimates for items in the 2018 operational analyses, and solid red lines identifying the critical values for flagging items. Direct comparisons should not be made between the number of items flagged in the current study and the 2018 operational analyses because different counts of items were included. These plots show that the items flagged for WRMSD in the current study were very close to the cut, with larger differences only observed for one item in grade 6 and two items in grade 7. The items flagged for robust-z exhibited a much larger range of estimates. Across the grades, between one to seven of the items, or 9.1% to 53.8% of the items flagged for robust-z, were within 0.50 units of the 2.33 cut. Overall, this suggests a number of the items that flagged for drift using the WRMSD criteria were just above the cuts which may pose less of a concern to test function.

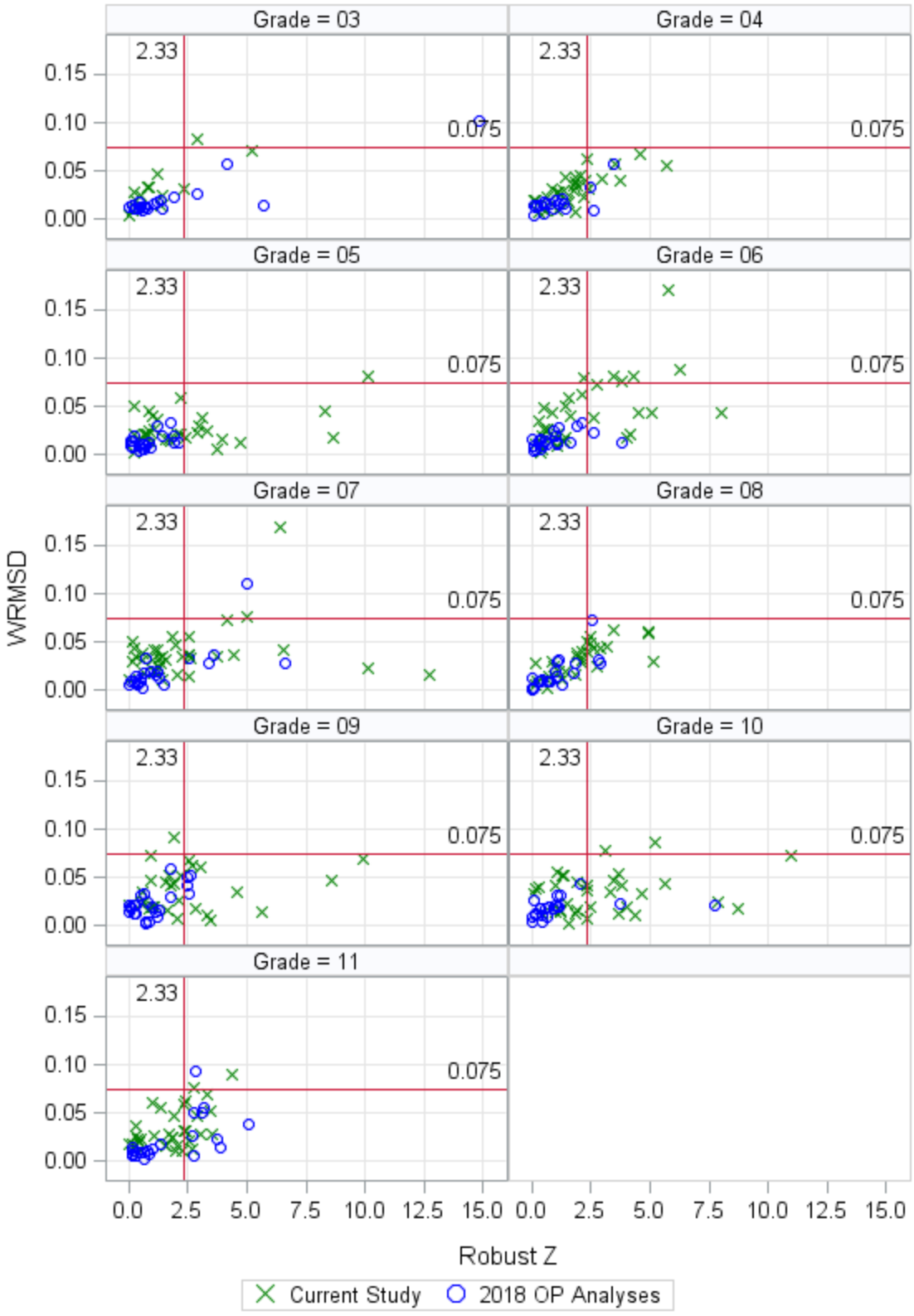

Figure 2. Scatterplots of drift estimates.

Finally, we examined estimates of drift by item type to evaluate whether certain types of items are getting flagged for drift more frequently. These results are presented in Table 10. Across all grades, 34.4% of composite items (MX), 37.0% of technology-enhanced items (XI), and 50.0% of open-ended items (OE) were flagged by robust-z. In contrast, 4.8% of MX items, 2.2% of XI items, and 14.7% of OE items were flagged by WRMSD. These results suggest slightly more OE items are being flagged for drift than the other item types. However, these results are not that surprising considering OE items tend to be more difficult and therefore had greater potential for differences in the *b* parameters.

Table 10. Percent of Items Flagged for Robust Z and WRMSD, by Item Type

			% of Items Flagged by:			
Grade	Item Type	Total Item N	Robust-Z: A Par.	Robust-Z: B Par.	Robust-Z: Total	WRMSD
03	MX	7	14.3	0.0	14.3	0.0
	XI	3	0.0	0.0	0.0	0.0
	OE	4	0.0	50.0	50.0	25.0
04	MX	22	0.0	18.2	18.2	0.0
	XI	4	0.0	50.0	50.0	0.0
	OE	4	0.0	25.0	25.0	0.0
05	MX	21	0.0	33.3	33.3	4.8
	XI	5	0.0	60.0	60.0	0.0
	OE	4	0.0	25.0	25.0	0.0
06	MX	25	12.0	24.0	36.0	16.0[a]
	XI	5	0.0	0.0	0.0	0.0
	OE	4	0.0	75.0	75.0	50.0
07	MX	24	8.3	25.0	33.3	4.2
	XI	8	0.0	25.0	25.0	0.0
	OE	4	25.0	50.0	75.0	25.0
08	MX	19	5.3	31.6	36.8	0.0
	XI	6	16.7	33.3	50.0	0.0
	OE	4	25.0	25.0	50.0	0.0
09	MX	22	4.5	31.8	36.4	0.0
	XI	4	50.0	0.0	50.0	0.0
	OE	2	0.0	0.0	0.0	50.0[a]
10	MX	27	3.7	51.9	55.6	7.4
	XI	3	0.0	0.0	0.0	0.0
	OE	4	50.0	25.0	75.0	0.0
11	MX	22	0.0	27.3	27.3	4.5
	XI	8	12.5	50.0	62.5	12.5
	OE	4	50.0	0.0	50.0	0.0
All Grades	MX	189	4.8	29.6	34.4	4.8
	XI	46	8.7	28.3	37.0	2.2
	OE	34	17.6	32.4	50.0	14.7

[a] One item flagged for WRMSD but not robust-z.

Comparison of Student Scores on Reduced Forms

Differences between the pre- vs post-equated scores are examined next. Two sets of summative conversion tables were produced to compare scores based on pre-equating to those based on post-equating. Only the subset of common items that were field tested in 2017 and then administered operationally in 2018 were used in the creation of the score conversion tables

to focus on the differences between the current pre-equating procedure (excluding unmotivated students) and post-equating. The student sample included students who were administered the ELA assessments in the spring of 2018, who met the criteria for inclusion in calibration analyses, and who were administered a form that contained items which were field tested in 2017.

The distribution of student performance levels across grades are displayed in Figure 3. The performance level distribution based on the post-equated scoring tables are presented in blue and the distribution based on the pre-equated tables are presented in red. Across the grades, the percentage of students at each performance level differed by 0% to 4% in grades 4-11. Larger differences were observed in grade 3, with differences of 2% to 6% observed across the performance levels. However, no systematic differences were observed across the pre- and post-equated results. The pre-equated results were associated with larger percentages for a particular performance level in some grades (e.g., level 5 in grades 3, 9, and 11), but lower percentages in other grades (e.g., level 5 in grades 4-7, and 10).

To further evaluate whether any consistent or systematic differences might exist in the performance level distributions, we collapsed the performance levels into proficient (levels 4 and 5) and not proficient (levels 1-3) and compared the results. Table 11 presents the percentage of proficient and not proficient students based on the pre- and post-equated scoring tables. The differences between the pre- and post-equated performance level distributions are further minimized when collapsing the results in this manner. The differences in the percent of proficient students ranged from 0% to 3.6% with the largest differences now observed in grade 7 (not grade 3). Furthermore, the direction of the change was not consistent, with the pre-equated scores associated with higher percentages of proficient students in grades 5, 6, 8, 9 and 11, and lower percentages in grades 3, 4, 7, and 10. Overall, these results further support the conclusion that there are no consistent or systematic differences between the pre- and post-equated performance distributions.

Table 11. Pre- vs Post-Equated Distributions of Student Proficiency on the Reduced Forms

	Post-Equated		Pre-Equated		Difference (Post – Pre)	
Grade	Not Prof.	Prof.	Not Prof.	Prof.	Not Prof.	Prof.
03	57.5%	42.5%	58.2%	41.8%	-0.7%	0.7%
04	54.2%	45.8%	56.0%	44.0%	-1.8%	1.8%
05	56.7%	43.3%	55.0%	45.0%	1.7%	-1.7%
06	58.0%	42.0%	57.8%	42.2%	0.2%	-0.2%
07	51.5%	48.5%	55.1%	44.9%	-3.6%	3.6%
08	53.3%	46.7%	51.6%	48.4%	1.7%	-1.7%
09	49.4%	50.6%	49.3%	50.7%	0.1%	-0.1%
10	51.0%	49.0%	51.0%	49.0%	0.0%	0.0%
11	59.0%	41.0%	57.3%	42.7%	1.7%	-1.7%

Note. Prof. = Proficient.

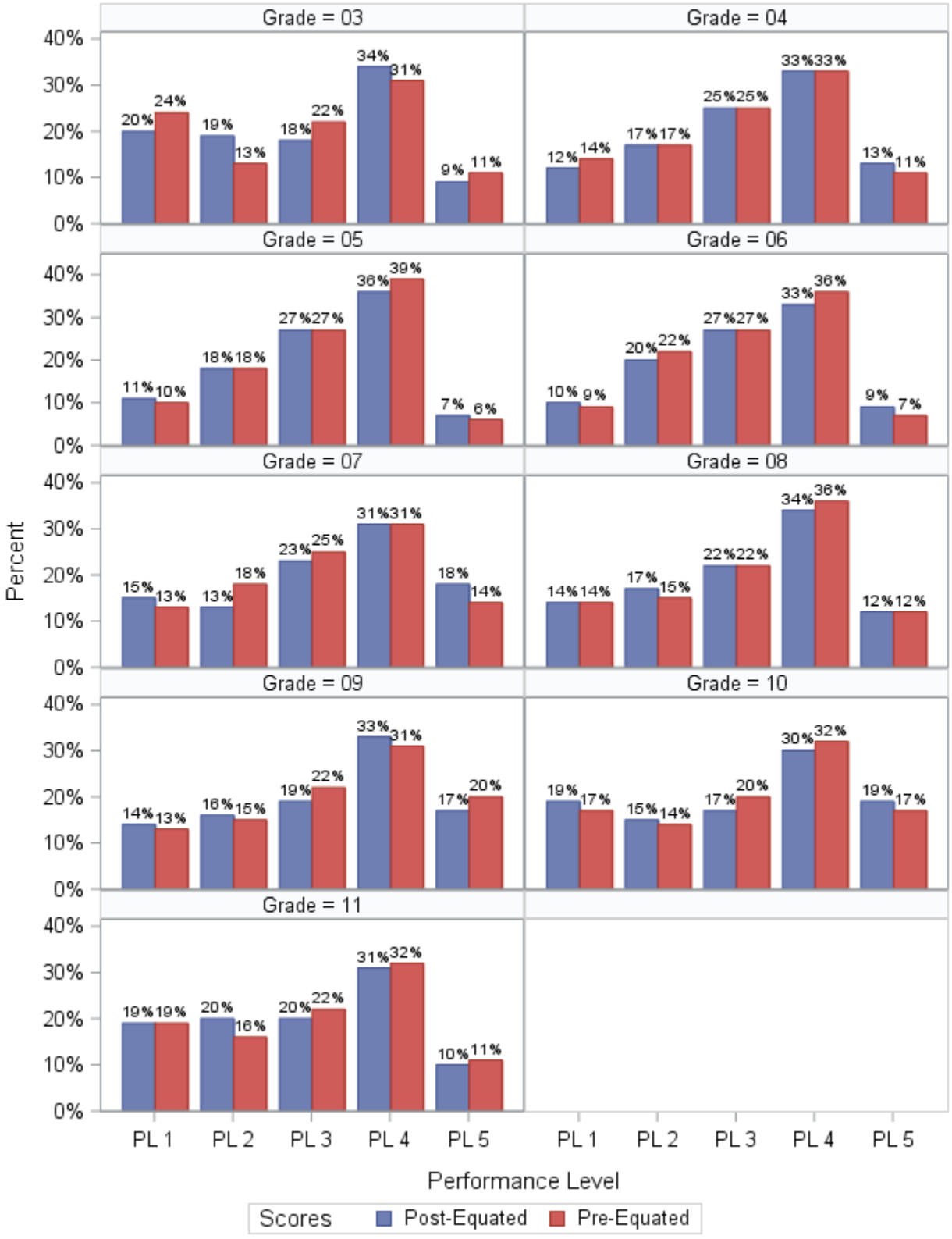

Figure 3. Pre- vs post-equated distributions of student performance levels on the reduced forms.

Differences in individual student scores are examined next. Table 12 summarizes the differences in individual student scores that were observed when using the pre- vs post-equated conversion tables. It is important to note, a 200-point scale is used to score students (650 to 850) and the current set of scoring tables are based on a subset of the operational items (no more than 51 items). This means there are going to be some score points that are not used, which could lead to bigger shifts in scale score differences than we might otherwise be comfortable with. To assist with understanding the magnitude of the score differences, we computed Cohen's *d* effect sizes for each of the point differences summarized in the table (e.g., 3-point difference, 6-point difference). According to Cohen (1988), *d*'s of .20, .50, and .80 correspond to small, medium and large effects, respectively. Using this metric, 95.4% to 100% of the students within each grade experienced pre- and post-equated score differences (≤ 9-points) of a small magnitude or less ($d \leq 0.22$). The remaining students experienced score differences (10-24 points) of a small to medium magnitude ($0.22 < d \leq 0.58$).

Table 12. Pre- vs Post-Equated Student Score Differences on the Reduced Forms

		Students with Pre- vs Post- Equated Scale Score Diff. of:						
Grade	Max Raw Score on Reduced Forms [a]	0 Points	≤ 3 Points	≤ 6 Points	≤ 9 Points	≤ 12 Points	≤ 15 Points	≤ 24 Points
03	20-27	8.2%	21.0%	47.1%	95.4%	96.6%	97.7%	100.0%
04	40-43	12.9%	50.6%	99.7%	100.0%			
05	40-51	17.3%	74.8%	96.9%	100.0%			
06	43-51	14.5%	66.7%	95.5%	99.8%	100.0%		
07	43-55	6.2%	39.4%	78.0%	97.1%	99.3%	100.0%	
08	35-49	3.7%	58.6%	99.2%	99.8%	100.0%		
09	20-51	6.6%	27.9%	59.5%	97.8%	100.0%		
10	43-51	13.8%	64.5%	91.4%	98.2%	100.0%		
11	43-51	22.0%	41.9%	94.6%	100.0%			
	Effect Size (Cohen's D) for Corresponding Point Differences	*0.00*	*0.07*	*0.14*	*0.22*	*0.29*	*0.36*	*0.58*

[a] These max scores reflect the weighted raw scores where PCR items are weighted to be worth more points than the number of categories they are scored on. With the exception of one core form in grade 9, all forms include at least 1 PCR item. The full operational forms included max raw scores between 80-109.

Table 13 summarizes the differences based on the pre- vs post-equated score tables. Across the grades, 84.6% to 96.2% of students were assigned to the same performance level by the two sets of conversion tables. Between 3.8% to 15.4% of students were assigned to different performance levels. Since the scoring tables reflect a subset of items, we are likely to see greater bounce in students across the performance levels than we would see on a full operational form. The cut scores for the performance levels were generally within the standard error of measurement for one or two scale score points below the cut. This means that these estimates can be viewed as a lower bound of the types of differences that could be observed, and estimates based on the full form would likely be higher.

Table 13. Pre- vs Post-Equated Student Performance Level Differences on the Reduced Forms

	Percent of Students Assigned Pre-Equated Scores that are:		
Grade	1 Level Below Post-Equated	Same Level as Post-Equated	1 Level Above Post-Equated
03	7.5%	84.6%	7.9%
04	7.7%	92.3%	0.0%
05	1.0%	94.4%	4.7%
06	5.8%	89.6%	4.6%
07	8.9%	88.2%	2.9%
08	0.0%	96.2%	3.8%
09	4.1%	85.4%	10.5%
10	2.5%	92.9%	4.6%
11	0.0%	93.9%	6.1%

To further understand where the differences in the scale were occurring, we compared the test characteristic curves (TCCs) between the pre- and post-equated scoring tables. These results are presented in Figure 4. There were two core forms (forms with the same set of operational items) for each grade and each form summed to a different number of raw score points due to the different number of common items on each form. Information about the number of items on each form that were flagged for drift and the mean drift among those items is also provided to offer context about item characteristics that could contribute to differences in TCCs. The largest difference in TCCs was observed on grade 3 form 1 and smaller differences were observed on grade 6 form 1, grade 7 form 1 and 2, grade 8 form 1, and grade 9 form 1. These trends generally correspond to the student score differences observed in Table 12—with grade 3 exhibiting the largest differences between pre- and post-equated scores and grades 6 through 10 exhibiting differences greater than 9 scores points. The remaining 12 forms had TCCs that were very close or overlapping.

Examination of the drift information suggests that the TCCs with differences did not have more items flagged for drift than other forms, but they did tend to have items flagged for robust-z with larger robust-z values. However, there were also some forms with higher robust-z values that did not exhibit different TCCs (grade 5 form 2, grade 6 form 2, and grade 9 form 2, grade 10 form 1); and the grade with the largest percentage of items flagged for drift (52.9% in grade 10, see Table 9) had pre- and post-equated TCCs that were very closely aligned. These results suggest that the differences in TCCs, and thus some amount of difference in student scores, are being driven by item drift, but not all drift contributes to differences in TCCs and student scores.

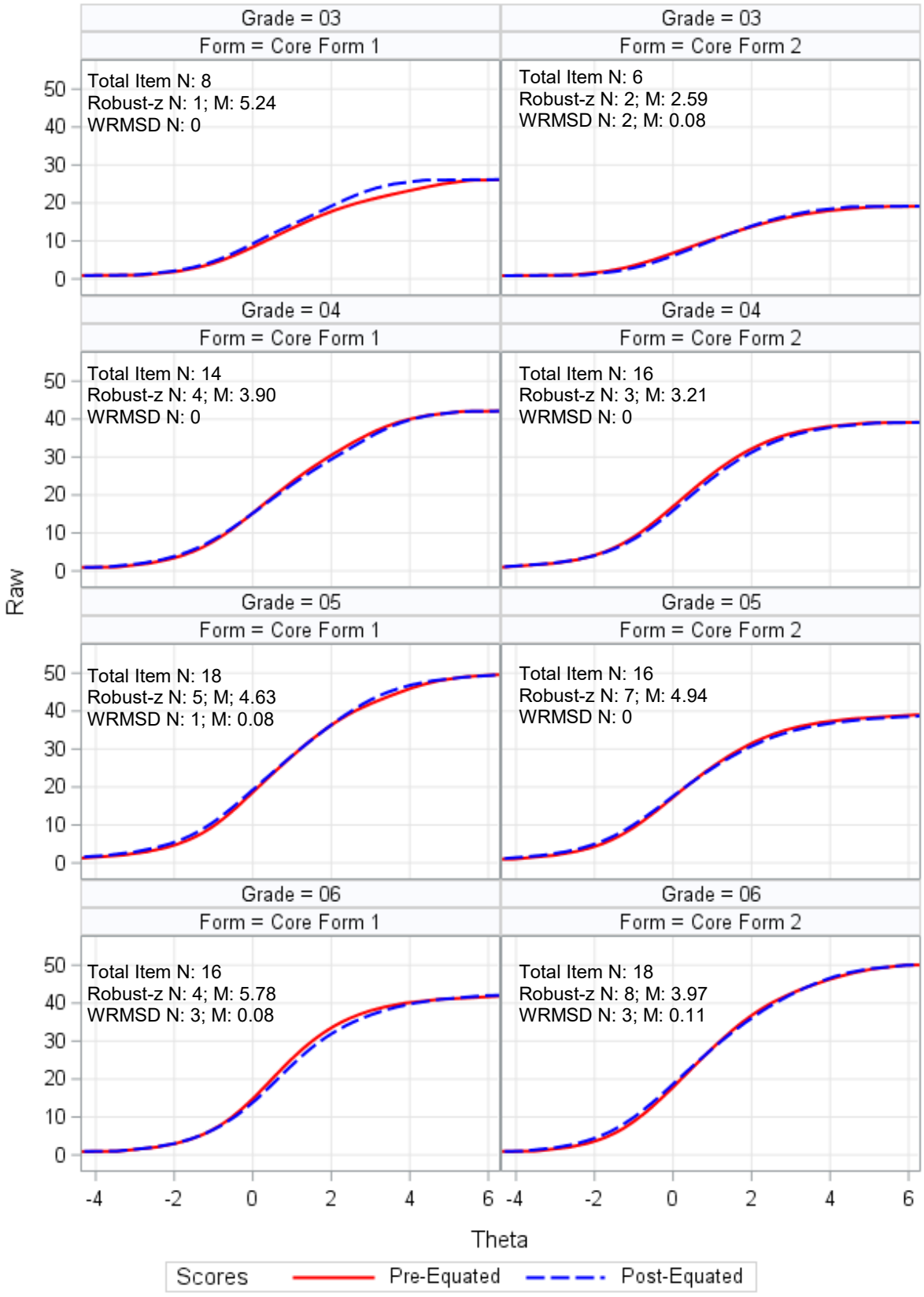

Figure 4. Pre- vs post-equated TCCs on the reduced forms.

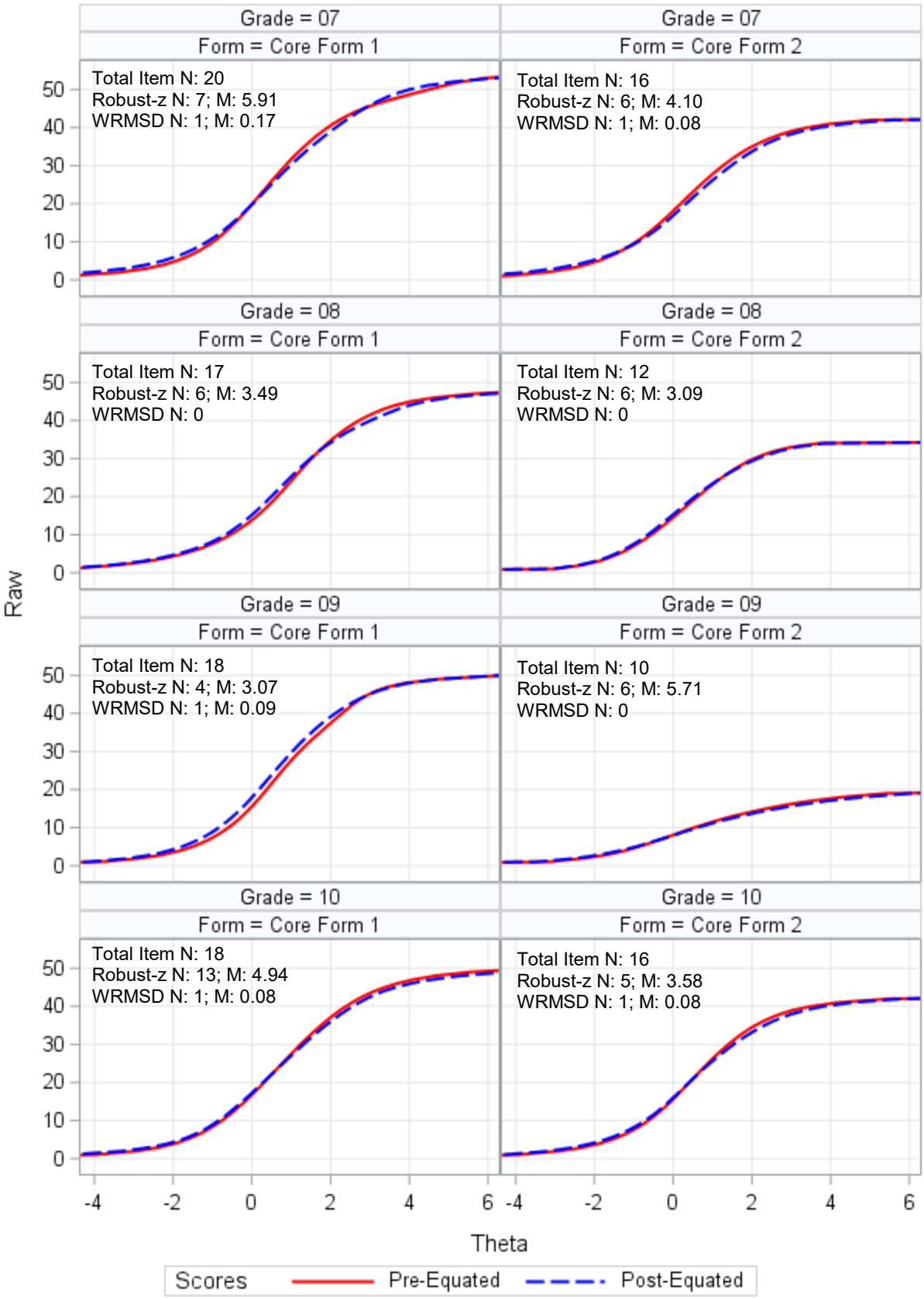

Figure 4. (Continued)

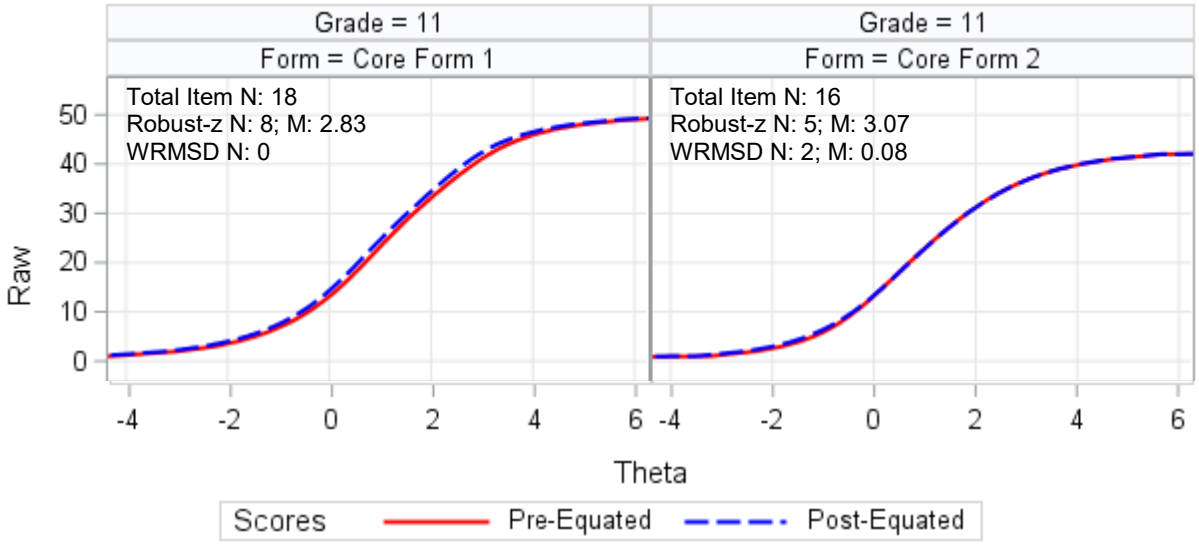

Figure 4. (Continued)

Comparison of Student Scores on Full Forms

Finally, differences between the pre- and post-equated scores based on the full 2018 operational forms are examined. This set of analyses differs from the previous because it compares student scores based on *all* 2018 operational items, not just the subset that were field tested in 2017 and then administered operationally. The pre-equated parameter estimates included the field test estimates where available, but also included estimates from the item bank for all other items. The post-equated parameter estimates were available for all items from the spring 2018 operational analyses. The inclusion of banked item-parameters (from post-equating of previous operational administrations) introduced greater stability into the item set and thereby masked some of the instability introduced by the pre-equated field test estimates, but it allowed comparisons of student scores based on intact operational forms that met the test blueprint. Similar to the previous set of analyses, two sets of scoring tables were computed based on the pre- and post-equated parameters and the tables were applied to a sample of students.

The distribution of student performance levels across grades are displayed in Figure 5. The performance level distribution based on the post-equated scoring tables are presented in blue and the distribution based on the pre-equated tables are presented in red. Across the grades, the percentage of students at each performance level differed by 0% to 3%, with 3% differences only observed in grades 8 and 9. As expected, these differences were smaller than those observed in the previous set of analyses using the reduced forms (see Figure 3; differences of 0% to 6%). No systematic differences were observed across the pre- and post-equated results.

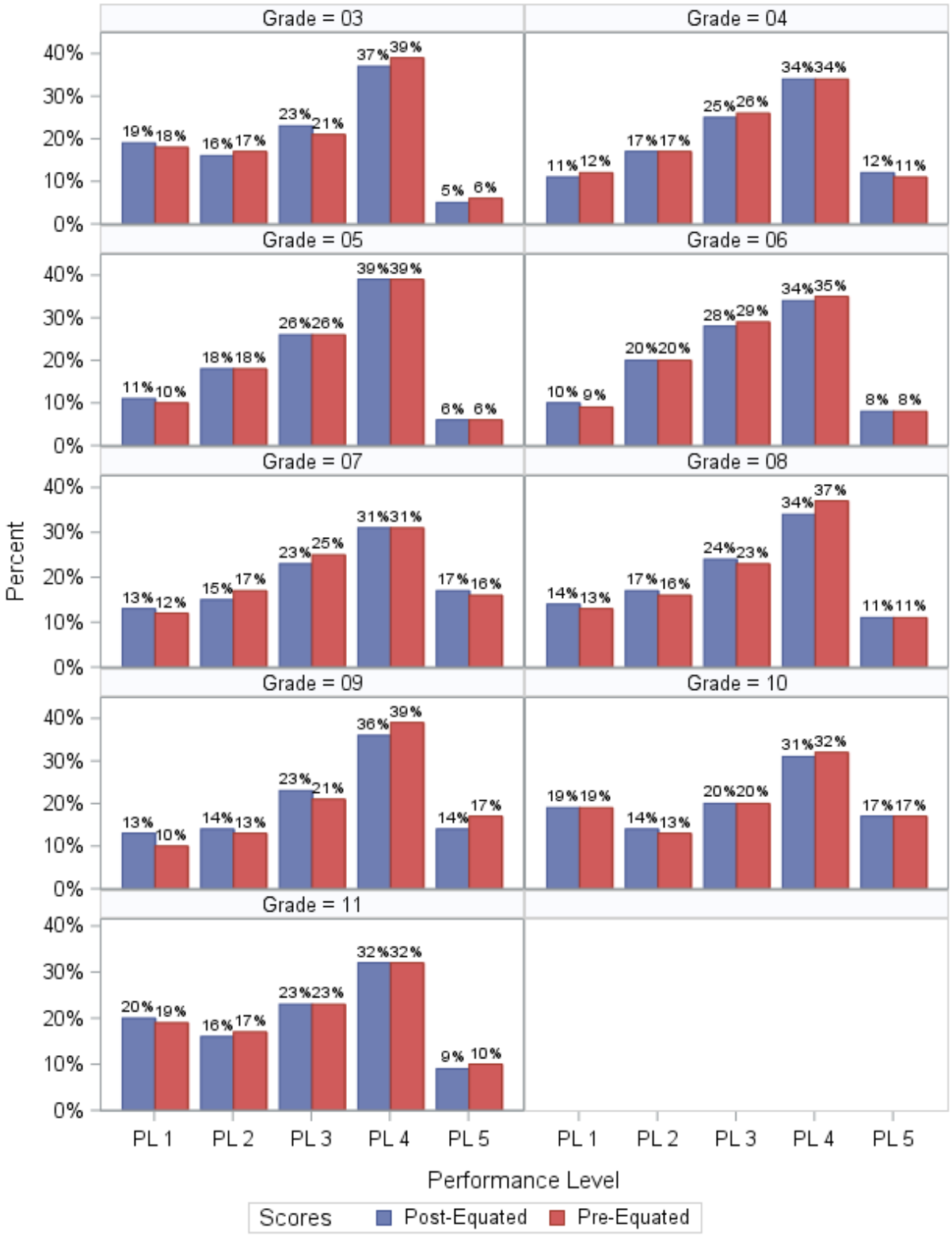

Figure 5. Pre- vs post-equated distributions of student performance levels on the full forms.

To further understand the pre- vs post-equated performance distributions, we collapsed the performance levels into proficient (levels 4 and 5) and not proficient (levels 1-3) and compared the results. Table 14 presents the percentage of proficient and not proficient students based on the two sets of scoring tables. The differences in the percent of proficient students ranged from 0.1% to 5.8%, with higher levels of student proficiency found among the post-equated score tables in grades 4 and 7. These results offer further support of the conclusion that there are no consistent or systematic differences between the pre- and post-equated performance distributions.

Table 14. Pre- vs Post-Equated Distributions of Student Proficiency on the Full Forms

	Post-Equated		Pre-Equated		Difference (Post – Pre)	
Grade	Not Prof.	Prof.	Not Prof.	Prof.	Not Prof.	Prof.
03	58.1%	41.9%	55.7%	44.3%	2.4%	-2.4%
04	53.4%	46.6%	55.2%	44.8%	-1.8%	1.8%
05	55.7%	44.3%	54.8%	45.2%	0.9%	-0.9%
06	57.7%	42.3%	57.6%	42.4%	0.1%	-0.1%
07	51.6%	48.4%	53.1%	46.9%	-1.5%	1.5%
08	54.7%	45.3%	51.8%	48.2%	2.9%	-2.9%
09	49.9%	50.1%	44.1%	55.9%	5.8%	-5.8%
10	52.1%	47.9%	51.3%	48.7%	0.8%	-0.8%
11	59.6%	40.4%	58.1%	41.9%	1.5%	-1.5%

Note. Prof. = Proficient.

Differences in individual student scores are examined next. Table 15 summarizes the differences in individual student scores that were observed when using the pre- vs post-equated conversion tables. To assist with understanding the magnitude of the score differences, we computed Cohen's *d* effect sizes for each of the point differences summarized in the table (e.g., 3-point difference, 6-point difference). According to Cohen (1988), *d*'s of .20, .50, and .80 correspond to small, medium and large effects, respectively. Using this metric, 99.3% to 100% of the students within each grade experienced pre- and post-equated score differences (≤ 9-points) of a small magnitude or less ($d \leq 0.23$).

Table 15. Pre- vs Post-Equated Student Score Differences on the Full Forms

	Max Raw Score on Full	Students with Pre- vs Post- Equated Scale Score Diff. of:				
Grade	Forms [a]	0 Points	≤ 3 Points	≤ 6 Points	≤ 9 Points	≤ 12 Points
03	80-82	7.8%	64.1%	98.8%	99.8%	100.0%
04	106	18.0%	95.4%	100.0%	100.0% [b]	
05	106	44.2%	96.4%	99.8%	100.0%	
06	109	17.0%	94.3%	100.0%		
07	109	9.7%	89.2%	97.4%	99.3%	100.0%
08	107-109	7.2%	98.2%	99.7%	100.0%	
09	109	0.8%	5.1%	93.6%	100.0%	
10	109	40.2%	100.0%			
11	109	10.6%	81.4%	100.0%		
Effect Size (Cohen's D) for Corresponding Point Differences		*0.00*	*0.08*	*0.15*	*0.23*	*0.31*

[a] These max scores reflect the weighted raw scores where PCR items are weighted to be worth more points than the number of categories they are scored on.

[b] 114 of the 2888,664 grade 4 students had score point differences of 7-9 points. Rounding

Table 16 summarizes the differences based on the pre- vs post-equated score tables. Across grades 3-8 and 10-11, 93.7% to 98.5% of students were assigned to the same performance level by the two sets of conversion tables. In grade 9, a smaller portion (83.8%) were assigned the same performance level. The remaining students (1.5% to 16.2%) were assigned pre-equated performance levels that were above or below the post-equated in grades 3, 6, and 7, below in grade 4, and above in grades 5, 8, 9 10, and 11. Overall, a larger proportion of students are assigned the same performance level in the current set of analyses using the full forms than the previous analyses based on the reduced forms (see Table 13). This would be expected because the full scoring tables are less coarse (have more scores represented) than the reduced scoring tables. However, there is a slight trend in the current results that was not observed on the reduced forms. The pre-equated score tables assign more students to higher proficiency levels than the post-equated tables in the higher grades (8-11). This suggests that there could be some systematic differences between the pre-equated parameters from the item bank (from post-equating of earlier administrations) and the post-equated parameters. Specifically, the pre-equated parameters from the item bank may be slightly more difficult than the post-equated parameters for the same items. The current study does not investigate item-level differences for items that were not field tested in 2017, but such differences could also impact student scores. Given that the differences were very small and did not occur across all grade levels, normally conducted checks on item parameter drift should be sufficient to monitor and account for these small differences before they substantially impact student scores for future administrations.

Table 16. Pre- vs Post-Equated Student Performance Level Differences on the Full Forms

	Percent of Students Assigned Pre-Equated Scores that are:		
Grade	1 Level Below Post-Equated	Same Level as Post-Equated	1 Level Above Post-Equated
03	1.0%	93.9%	5.1%
04	4.4%	95.6%	0.0%
05	0.0%	97.4%	2.6%
06	2.2%	94.7%	3.1%
07	3.8%	94.0%	2.2%
08	0.0%	93.7%	6.3%
09	0.0%	83.8%	16.2%
10	0.0%	98.5%	1.5%
11	0.0%	95.5%	4.5%

Comparison to Math Pre-Equating Results

Pearson conducted a study in the summer of 2016 to evaluate the impact of moving the PARCC Math assessment from a post-equating model to a pre-equating model. Pearson's study examined the impact of such a transition on student scores, using the same approach as the analyses in this section and comparing scores based on full operational forms. The results of the study were briefly summarized in a presentation for the PARCC Technical Advisory Committee (TAC). The presentation included Figure 6, which presents the scale score difference for each potential raw score for three Math grade 7 forms (each color represents a distinct form).

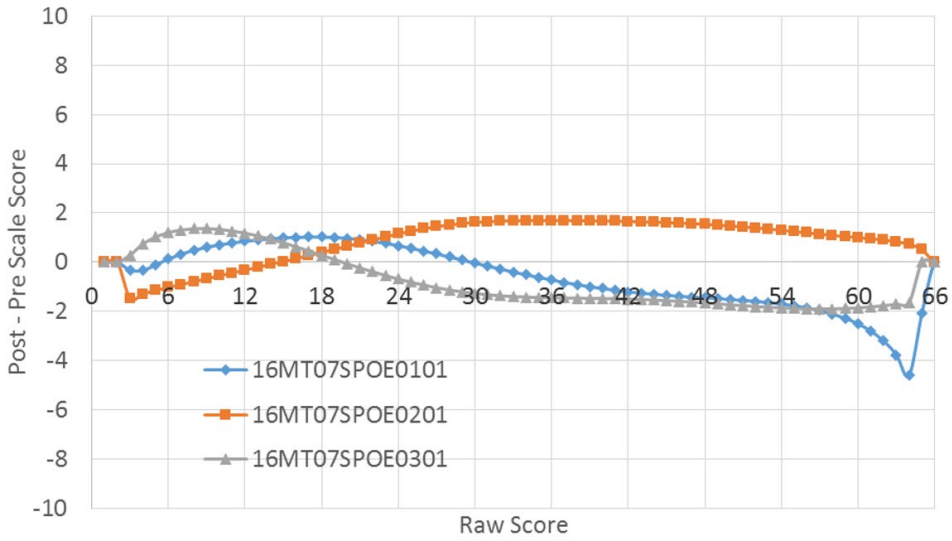

Figure 6. Math G7 differences in pre- vs post-equated scale scores on the full forms.

We created comparable plots for the ELA assessments in order to compare Pearson's Math results to the current ELA results. The plots are presented in Figure 7. To further assist with interpretation, we changed the x axis to represent the pre-equated scale score and the y axis to then represent the post-equated score difference. Additionally, the scale score cuts associated with the five proficiency levels are provided as vertical grey lines and a density plot capturing the distribution of the 2018 operational student scores is presented in blue. These plots are directly comparable to the Math plot above.

The maximum difference observed in the grade 7 math plot was approximately 5 scale score points. The maximum difference observed across the ELA grades was 11 scale score points, but most ELA scale score differences were within 5 score points. Across the ELA grades, larger differences in scores tended to occur at places along the scale where fewer students were observed. This means the majority of students had relatively small differences in their pre- and post-equated scores, and only a small group experienced the larger differences. Overall, the example math grade exhibited smaller scale score differences between the pre- and post-equated scores than the ELA grades, but the ELA grades were reasonably close.

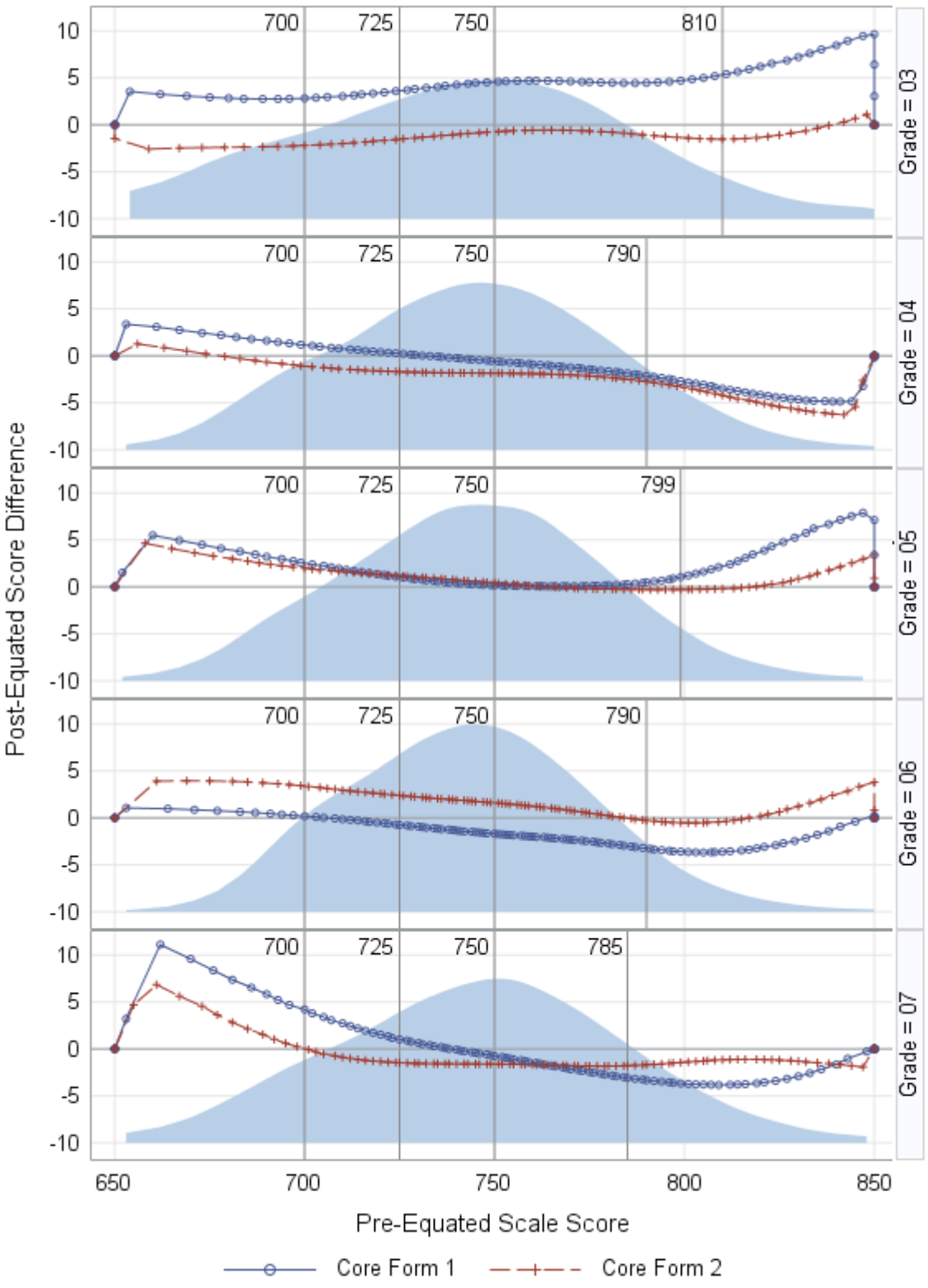

Figure 7. Differences in ELA pre- vs post-equated scale scores on the full forms.

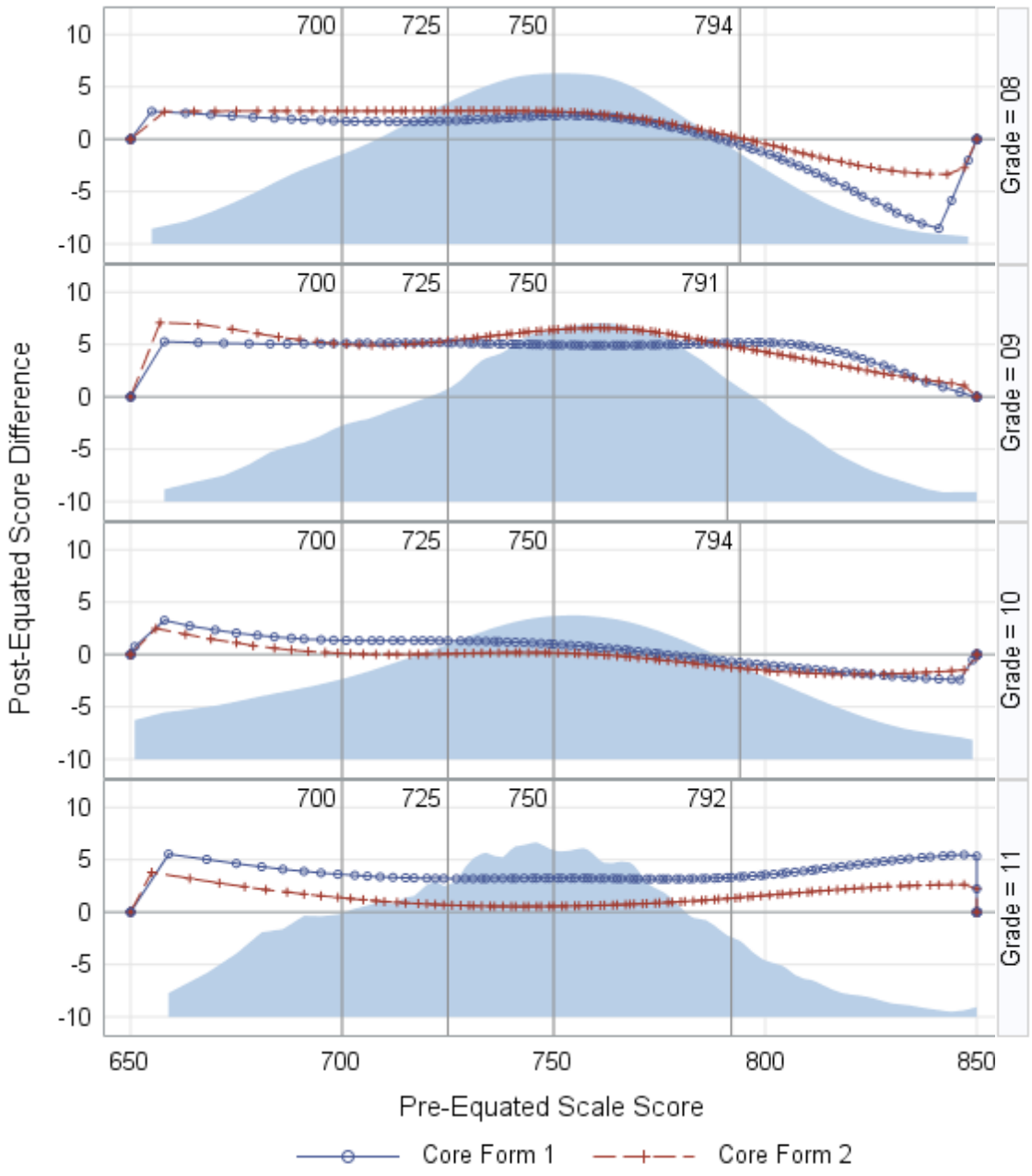

Figure 7. (Contined)

Summary and Conclusions

Three sets of analyses were conducted in the current study to examine the impact of transitioning the PARCC ELA assessments to a pre-equated scoring model. The first set of analyses examined item parameter differences, the second examined score differences, and the third compared the results of the current study to the Math pre-equating study.

Overall, these analyses did not identify any substantial systematic or consistent differences between the pre- and post-equated results across the grades. The pre-equated item parameters were found to be slightly more discriminating, but no more or less difficult than the post-equated parameters. Differences in individual student scores and performance levels were observed between the pre- and post-equated scoring tables, but overall student performance level distributions were similar. These findings generally support the conclusion that the pre-equated parameters and scoring tables are not consistently different from the post-equated parameters and tables. Additionally, these findings suggest that the rules for excluding unmotivated students helped to address motivation issues and account for differences between the pre- and post-equated parameters that could otherwise result.

However, item drift was found to drive differences in test function for some grades. The current study flagged a relatively large proportion of the post-equated item parameters for drift from their pre-equated parameter estimates (21.4%-52.9% of the common items across the grades, see Table 9). These proportions were larger than those commonly observed during post-equating, when post-equated operational estimates are compared to post-equated estimates from earlier administrations. Further examination of test characteristic curves (TCCs) and student score differences suggested that item drift of a large magnitude can contribute to differences in scores, but it does not always. Across the 18 core forms in grades 3 through 11, eleven of the forms had mean robust-z values for the flagged items that exceeded 3.5 (1.20 units above the robust-z critical value). Only four of these forms were found to have slight differences in their pre- and post- equated TCC curves, the other seven forms exhibited very closely aligned TCCs. Furthermore, two additional forms were found to have slight differences in their TCC curves, but they had lower mean robust-z estimates.

These findings concerning item drift are not surprising. We know item drift occurs and it impacts both item and test function. In a pre-equated context, drift will also impact student scores. Thus, the current study aimed to examine the level of drift that occurs and the impact on student scores. Although a large number of items were flagged for drift, the majority of pre- and post-equated TCCs showed minimal differences. Furthermore, the difference between the pre-equated and post-equated scores for almost all students were of a small magnitude (within 0.22 scale score standard deviations) or less. We know post-equating provides the *best* estimates of current student ability, but the pre-equated estimates were reasonably close. Together, these findings suggest item drift from pre-equated field test parameters does not have a detrimental impact on PARCC ELA student scores and that the pre-equated scores are comparable to post-equated scores.

A transition to pre-equating is generally supported based on comparisons to the Math pre-equating study conducted by Pearson. Although limited results were available from that study, to make comparisons against, we were able to compare pre- vs post-equated score table differences (based on the complete operational forms) between the two studies. The score differences were generally within 5 score points in the current study of ELA and were within 2 score points for grade 7 in the Math pre-equating study (the only grade with results available for

comparison). Although the example math grade exhibited smaller scale score differences, the ELA grades were reasonably close.

If the PARCC ELA assessments are transitioned to a pre-equated scoring design, item drift analyses will play an important role in maintaining the item bank. Item drift of all operational items will need to be examined after test administration and items flagged for drift should be reviewed by a content specialist to evaluate whether there may be a reason for item drift. Recalibration of the item parameters should also be considered using student data from the operational calibration. The drifted items would not be updated for students who were administered the item, but they could be updated for students who will see the item in the future.

References

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, N.J: L. Erlbaum Associates.