COVID-19 Impact on Group Invariance Property of Equating

Paper presented at the annual meeting of
the National Council on Measurement in Education,
San Diego

April 23, 2022

Dong-In Kim, Marc Julian, and Pam Hermann
Data Recognition Corporation

COVID-19 Impact on Group Invariance Property of Equating

Introduction

Test equating is performed whenever alternate test forms are administered, and therefore the equated scores can be used interchangeably. One critical equating property is the group invariance property (Dorans & Holland, 2000). To treat scores on alternate forms as interchangeable, the equating function used to convert performance on each alternate form to the reporting scale should be the same for various subgroups, such as gender, ethnicity, or locale. If the equating functions for subpopulations are systematically different, the interchangeability of test scores on alternate forms is questioned (Kolen, 2004).

Lord (1980) indicated that true score equating relationships have subgroup independence if a unidimensional item response theory (IRT) model holds. However, unidimensional IRT models are not likely to hold in practice. That is, equating results cannot be completely group independent in practice, and it is important to examine how much the invariance property of equating is violated in each subgroup. Researchers have found that the differences in the equating conversions are small across examinee groups, especially in situations where carefully constructed test forms are equated (Angoff & Cowell; Harris & Kolen, 1986). Peterson (2008) pointed out that students with limited proficiency with the English language are potentially at a disadvantage when taking tests that measure ELA. Evaluating the invariance property in this context seems required.

The COVID-19 pandemic has led to unprecedented impacts on student learning in the last two school years. To mitigate the impact of disrupted learning on the item parameters, a pre-equated approach was frequently implemented for student scoring in the spring 2021 administration of large-scale assessments. That is, item parameters estimated before the pandemic were used to build score conversion tables wherever possible.

Because of the uneven impact of the pandemic on student learning (Betebenner and Wenning, 2021) it is important to investigate whether the group invariance property was satisfied in the application of pre-administration equating. The purpose of this study is to examine how much the group invariance property was satisfied for various subgroups when pre-equating was applied to large-scale assessments in spring 2021.

Methodology

Data

Large-scale state assessment data from spring 2019 and spring 2021, including English language arts (ELA) grades 6 and 8 and mathematics (Math) grades 6 and 8, were included in this study. Spring 2019 data were included to provide the degree of invariance under a normal situation. There were two scenarios between spring 2019 and spring 2021 equating. Scenario 1 – spring 2019 uses a traditional post-equating design, using a common-item non-equivalent group design. The anchor that links to the spring 2018 administration is a small subset sampled from the full test blueprint. Scenario 2 – 2021 uses a pre-equated approach where all items are anchored to where item parameters are currently maintained in the item bank. Some of those items are previously operational items and some items are from previous field-testing efforts. It's important to note that the student populations administered the tests in scenarios 1 and 2 have the fundamental difference of being impacted directly by the pandemic. That is, students in 2019 were administered tests under standard conditions, and students in 2021reflect post-

pandemic effect.

Each test consisted of mixed item types, including, multiple-choice, constructed-response, and various technology-enhanced types (see Tables 1 and 2). Most items on the ELA 2019 and 2021 tests were common except for 2 or 3 items, and all items on the Math 2019 and 2021 tests were the same.

All students with valid 2019 and 2021 scores were included in the study. Table 3 shows sample sizes for each spring 2019 and spring 2021. The spring 2021 tested population was found to be different from the Spring 2019 tested population regarding several demographic variables. Specifically, African American students, SES, and students from 'City' districts were underrepresented in the Spring 2021 test participants, while White and not-SES students were overrepresented. SES indicates economically disadvantaged students who were identified by National School Lunch Program, a member of a household that meets the income eligibility guidelines for free or reduced-price meals, and identified by an alternate mechanism, such as the alternate household income form.

Four different subgroups were studied: gender (Female and Male), socio-economic status classifications (SES or not), race/ethnicity groups (Hispanic, African American, White), and limited language proficiency groups (LEP or not) were included in this study.

Calibration and Equating

The three-parameter logistics (3PL) and the two-parameter partial-credit (2PPC) models were applied to calibrate multiple-choice and polytomous items (Lord & Novick, 1968; Yen & Fitzpatrick, 2006). The 2PPC model is equivalent to the generalized partial credit model (Muraki, 1992). The IRT calibrations were implemented using PARDUX software (Burket, 2002). PARDUX simultaneously estimates parameters for multiple-choice and polytomous items using marginal maximum likelihood procedures implemented via EM algorithm.

The following steps for calibration, equating, and scoring at a given subgroup were performed:

Step 1: Calibrate the 2021 assessment with subgroup student's item responses

Step 2: Perform equating by applying Stocking-Lord method to the item parameters of Step 1 using all spring 2019 item parameters as anchor item parameters

Step 3: Estimate theta points corresponding to 2021 raw-score points using the equated item parameters from Step 2 by applying the inversed TCC method

Step 4: Estimate expected raw scores on the 2019 scale by applying theta points from Step 3 to 2019 item parameters.

These steps were applied to each subgroup. The same procedure was applied to post-equating from spring 2019 to spring 2018 with 2018 anchor items, in which equating results were used as comparison criteria for invariance property.

To evaluate the results of the equating invariance – different reporting scores were used. There are three kinds of equated scores that were considered: 1) unrounded raw scores, 2) rounded raw scores, and 3) maximum likelihood estimate (MLE). The expected raw scores in Step 4 are unrounded scores that are often transformed into scaled scores using a linear transformation. Rounded raw scores were obtained by

rounding the unrounded scores with 0.5 up. Some large-scale assessments have applied a pattern scoring based on the fixed number of items. MLEs based on the pattern scoring were estimated using Step 2 item parameters.

Evaluation Criteria

The results were evaluated using the root mean square difference (RMSD) and the root expected mean square difference (REMSD) indices (Dorans & Holland, 2000). The RMSD compares equating functions between a subgroup and its population at each score point of 2021. To summarize RMSD values as a single value, Dorans & Holland (2000) also introduced REMSD:

$$\text{RMSD}(x) = \frac{\sqrt{\sum_j w_j [e_{P_j}(x) - e_P(x)]^2}}{\sigma_{Y_P}} \text{ and}$$

$$\text{REMSD} = \frac{\sqrt{\sum_j w_{jE_P} [e_{P_j}(x) - e_P(x)]^2}}{\sigma_{Y_P}},$$

where x is a raw score (RS) point; P is the 2021 population; $e_P(x)$ denotes the equating function that equates 2021 to 2019 on the population P; $e_{P_j}(x)$ denotes the equating function that equates 2021 to 2019 on the subgroup P_j of P ; σ_{Y_P} is the standard deviation of the 2019 RS for the population; w_j denotes the weight, which is the set of weights that some to unity; and E_P denotes averaging over the distribution of 2021 scores in P. The set of subgroups participations P into a set of mutually exclusive and exhaustive subgroups.

For number-correct scoring, one rounded raw score point translates often to different scale scores and potentially different performance level classifications. However, differences in equating functions for unrounded scores are not clear when a linear transformation is applied. Depending on the unit of the transformed scaled scores, differences that matter (DTM; Dorans & Feigenbaum,1994) can be different. Brennan (2008) pointed out that DTM with half a reported score unit, such as 0.5 of raw score point, applies to unrounded equivalents only. In this study, raw scores were used as scaled scores for unrounded and rounded equated scores, and each 0.5 and 1 raw score points were used as DTM.

When unrounded raw scores are applied to scoring, DTM (differences that matter; Dorans et al., 2003) can be a half point because this half-point can be one raw score point by rounding up 0.5. Davier and Wilson (2008) introduced SDTM for evaluating RMSD and REMSD. The SDTM is the DTM divided by the denominator, the standard deviation in the RMSD equation. SDTM values for unrounded scores of all four tests were about 0.05 because the standard deviation for all tests was close to 10. In a similar manner, 0.1 was the SDTM for rounded scores.

Cohen's effect size *d* (1988) was also calculated for each subgroup category (i.e., female or Hispanic) as an evaluation criterion:

$$\frac{(Mean_{catG} - Mean_{pop\ in\ catG})}{\sqrt{(SD^2_{catG} + SD^2_{pop\ in\ catG})/2}},$$

where 'pop in catG' denotes samples of the subgroup category with population equating function. RMSD and REMSD values are for each subgroup, such as gender and ethnicity, not for its' subgroup category. Cohen's *d* was applied to each group category.

Results

As one of the requirements for equating rather than linking, Dorans and Holland (2000) introduced equal reliability. Tables 4 and 5 show summary statistics, including Cronbach's alpha and standard errors of measurement across the population and subgroup categories for ELA and Math grades 6 and 8. For ELA, test reliability ranged from 0.78 to 0.90, and the reliability was similar between 2019 and 2021 across subgroup categories. The subgroup LEP showed the smallest reliability of 0.79 and 0.78 for each 2019 and 2021 ELA grade 6, and there were 0.01 differences for 2019 and 2021 reliability. For Math, test reliability ranged from 0.71 to 0.92. The reliability differences between 2019 and 2021 range from 0 to 0.06 across subgroups. As ELA, the LEP subgroup showed the smallest reliability and the largest differences of 0.06 and 0.05 for each Math grade 6 and 8 between 2019 and 2021. The ethnicity category of African American showed the next largest difference of 0.05 and 0.04 across years for each Math grade 6 and 8. The raw score means between 2019 and 2021 were similar for ELA but different for Math. The raw score means of 2019 Math were smaller than those of 2021 Math across all subgroups. Similar patterns of more changes in performance seen for Math than ELA have been reported in other large-scale assessment programs.

Figure 1 shows the raw score differences between pre- vs. unrounded post-equating for four test forms in the population. Two horizontal lines of ±0.5 in this figure were included as DTM. The differences were small across raw score points. Note that maximum raw score points for ELA and Math tests are different as can be seen in Tables 1 and 2.

Figures 2 to 5 present the conversion differences of equating functions between each category subgroup and population. For example, the conversion differences for female in Figure 2 are raw score differences between unrounded female equating function and population function based on pre-equating. A positive value indicates the category subgroup equating function value is larger than the population equating function, and a negative value indicates the category subgroup equating function value is smaller than the population equating function. Across all four sets of subgroup analyses, the differences were within ±0.5, except for African American and Hispanic categories (see Figure 4). In Math grade 6, the raw score differences for Hispanic were about 0.5 at the extreme range, and those for African American were larger than 0.5 above the raw score point of around 37. The conversion differences for LEP fluctuated across raw score points for all four tests, and some conversion differences were over 1. Figure 6 presents the plot for the rounded equated score differences for LEP. As can be expected from unrounded score plot in Figure 5, all four tests showed that there were some raw score points with conversion differences larger than ± 1.0.

Figure 7 presents RMSD for unrounded equated scores for gender, SES, ethnicity, and LEP. RMSD values were close to 0 across most raw score points for all four tests, but there were deviations from 0 for ethnicity and LEP. For ethnicity, Math grade 6 showed the largest RMSD values around high score values. For LEP, RMSD values for Math 6 fluctuated across all raw score points and all four tests showed the highest RMSD values at maximum score points. No RMSD values were larger than the SDTM of 0.05, and therefore the invariance property was satisfied for all four subgroups across four test forms.

Figure 8 presents RMSD for rounded equated scores for gender, SES, ethnicity, and LEP. The RMSD values were 0 for gender and SES for all tests. The invariance property was fully satisfied for gender and SES. Math 6 produced the largest RMSD value of about 0.05 around the high raw score points for ethnicity. For LEP, all four tests produced RMSD values larger than 0.02 around the maximum score point. No RMSD values were larger than the SDTM of 0.1, and therefore the invariance property was satisfied for all four subgroups across four test forms.

Table 6 presents REMSD of unrounded equated scores for both spring 2019 and spring 2021 subgroups. REMSD values for spring 2019 post-equating were estimated to provide guidance for evaluating the spring 2021 REMSD values. REMSD values for spring 2021 were smaller than those for spring 2019 for ELA grades 6 and 8, and REMSD values for spring 2021 were larger than those for spring 2019 for Math grades 6 and 8. Several studies have reported a larger pandemic impact on students' performance in Math than in ELA (Megan et al., 2020 and 2022). Smaller REMSD values for spring 2021 than spring 2019 may reflect the fact that the number of anchor items in 2021 is larger relative to the 2019. That is, the spring 2021 anchors were all spring 2021 items, the equating was more stable compared to spring 2019 equating. No RMSD values were larger than the SDTM of 0.05, and therefore the invariance property was satisfied for all four subgroups across four test forms.

Table 7 presents REMSD of rounded equated scores for both spring 2019 and spring 2021 subgroups. When rounded equated scores were applied, REMSD values for spring 2021 were smaller than those for spring 2019 across all four subgroups. Also, all REMSD values for spring 2021 were smaller than the SDTM value of 0.05. Only REMSD values for spring 2019 EL6 and EL6 gender were slightly larger than the SDTM value of 0.05. No RMSD values were larger than the SDTM of 0.1, and therefore the invariance property was satisfied for all four subgroups across four test forms.

To examine the impact of differences between subgroup category's equating function from population equating function, the effect size (ES) between subgroup category and population equating functions was calculated using subgroup category samples. A positive ES value indicates the subgroup category sample mean based on subgroup category equating function is larger than that based on population equating function. A negative ES value indicates the subgroup category sample mean based on subgroup category equating function is smaller than that based on population equating function. Table 8 presents the ES values for spring 2021 unrounded equated scores. Absolute ES values for Math grade 6 were larger than 0 in SES, Hispanic, African American, and LEP. Absolute ES values for Math grade 8 were larger than 0 in African American and LEP. ES value for EL 8 LEP was also larger than 0. For equating,

one tenth of standard deviation (ES=±0.1) and one twentieth of standard deviation (ES=±0.05) have been often suggested as each general and strict flag criteria values (Kolen, 2019). Only ES of -0.09 for Math grade 6 LEP was flagged with ES =±0.05.

Table 9 presents ES values for rounded equated scores between subgroup category and population for spring 2021. All ES values were 0 except for Math grade 6 African American with ES of -0.03 and LEP with ES of -0.11, which was flagged with ES=±0.1. This indicates LEP students' performances are slightly deflated if the equating function based on LEP students' samples is applied.

Table 10 presents ES values for MLE values between subgroup category and population for spring 2021. MLE values for subgroup category and population were estimated using spring 2021 items and category samples. ES values for Math grade 6 African American and LEP were flagged with ES=±0.1. When ES=±0.05 was applied, Math grade 8 African American was also flagged.

Summary and Discussion

While most studies for invariance property have been for post-equating, this study investigated the invariance property with pre-equating. Due to the potential impact on live calibration and equating, many large-scale assessment programs have opted to reuse previously administered forms and applied pre-equating rather than developing new forms and post-equating in 2021. This study included grades 6 and 8 ELA and Math from a spring 2021 large-scale assessment program and examined the invariance property for four subgroups: gender, SES, ethnicity, and LEP. Three different scoring methods were studied: unrounded equated scores, rounded equated scores, and MLE as IRT pattern scoring. As evaluation criteria, RMSD and REMSD indices were included (Dorans & Holland, 2000). ES was also used for evaluating the equating function differences between each subgroup's categories and the population.

Dorans & Holland (2000) suggested that one of equating requirements is comparable test reliability across populations and subgroups. There were some reliability differences of at least 0.05 between 2019 and 2021 in Math grade 6 African American and LEP and Math grade 8. RMSD values for unrounded and rounded LEP were larger than 0, which indicates complete invariance property satisfaction, but smaller than the SDTM value of 0.05. REMSD values were a little larger than 0 for spring 2019 and spring 2021 unrounded scores, but smaller than the SDTM value of 0.05. REMSD values for 2019 were larger than those for 2021 for ELA, and those for 2019 were smaller than 2021. This is because the pandemic impact was larger on mathematics and smaller on ELA. All RMSD values for rounded scores were smaller than the SDTM value of 0.05 except for spring 2019 ELA6 and ELA8.

When ES was applied to evaluate the equating differences between subgroup's categories and the population, unrounded scores of Math grade 6 LEP was flagged with ES=±0.05, and there were no flagged categories for rounded scores. Absolute ES values for MLE were larger than those for unrounded and rounded raw scores for most categories and tests. ES values for Math grade 6 African American and LEP were flagged with ES=±0.1. When ES=±0.05 was applied, Math grade 8 African American was also flagged.

In short, the invariance property was satisfied for four spring 2021 subgroups with respect to RMSD and REMSD when the SDTM value of 0.05 was applied. For subgroup categories, African American and LEP for Math grade 6 were flagged for rounded raw scores or MLE when ES was applied.

Invariance property has been often studied when post-equating is performed with alternate forms under common (i.e., anchor) item design. In this case, the quality and number of anchor items between alternate forms influence the equating results and invariance property. In general, equating results can be better with many anchor items if anchors are close to a minimal version of the full test. Pre-equating design fully satisfies the requirement of the maximum number of anchor items because all items can be used as anchor items, and therefore invariance property can be better achieved than post-equating with some common items under the same condition. RMSD and REMSD values for ELA, where students' performances were not much impacted by the pandemic, showed the same results.

Most studies for invariance property performed post-equating without anchor screening. However, it is often common that many large-scale assessment programs conduct anchor screening before equating especially when the IRT model is applied. Some flagged anchor items are dropped in the anchors, and the remaining anchor items are used for equating. Future invariance studies may need to include this anchor screening. There will be two options: 1) apply the same anchors from the population equating to all subgroups and 2) apply different anchors to different subgroups.

In practice, the location of DTM needs to be considered. That is, DTM across the full range of scores – is that some differences matter more because of the associated stakes of certain cut points. For example, where we have seen some differences that matter, some are more important than others. In the middle of the distribution – where we have more information available – the differences that occur there will have greater impact. Not to mention, there are likely to be cut scores at various points on the scale, with varying degrees of consequences.

Peterson (2008) mentioned, "In practice, it is very important that all operational testing programs conduct population invariance studies to determine if there are any major subgroups of interest for which the equating results may not be comparable, given the various data collection designs that could be tested." Invariance property has been often performed in post-equating, but the results for African American and LEP in this study showed that the invariance property study needs to be conducted even for pre-equating.

There is a limitation in this study. First, only one large-scale assessment program with ELA and Math grades 6 and 8 was included in this study. Second, in this study, item calibration and test equating were performed using the IRT models, 3PL/2PPC. To generalize the results, other equating methods based on other psychometric models need to be applied. Third, no anchor screening method was applied.

Reference

Angoff, W. H., & Cowell, W. R. (1986). An examination of the assumption that the equating of parallel forms is population-independent. *Journal of Educational Measurement*, 23 327-345.

Betebenner, D. W. and Wenning, R. J. (2021). Understanding pandemic learning loss and learning recovery: The role of student growth & statewide testing. National Center for Educational Assessment.

Brennan. R. L. (2008). Discussion of population invariance. *Applied psychological Measurement*, 32, 102-114.

Burket, G. R. (2002). PARDUX [Computer program]. Unpublished.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. (*ETS Research Memorandum 94-10*). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37, 281-306.

Von Davier, A., A., & Wilson, C. (2008). Investigating the population sensitivity assumption of item response theory of examinees and two test formats. *Applied Psychological Measurement*, 32, 11-26

Harris, D. J., & Kolen, M. J. (1986). Effect of examinee group on equating relationships. *Applied Psychological Measurement*, 10, 35-43.

Harris. D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education*, 6. 195-240.

Kolen, H, J. (2004). Population Invariance in Equating and Linking: Concept and History. *Journal of Educational Measurement*, 41, 3-14.

Kolen. M. J., & Brennan. R. L. (2004). *Test Equating, scaling, and linking: Methods and practices (2nd ed.)*. New York: Springer-Verlag.

Kolen. M. J. (2019). Personal discussion for equating evaluation criteria.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*, Hilladale, NJ: Lawrance Erlbaum.

Megan, K., Soland, J., Tarasawa, B., Johnson, A., Ruzek, Erik., Liu, J. (2020). Projecting the potential impact of COVID-19 school closures on academic achievement. Educational Researcher, 49, 549-565.

Megan K., Soland, J., Kewis K., & Morton, E. (2022). *The pandemic has had devastating impact on learning. What will it take to help students catch up?* Brown Center Chalkboard.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159–176.

Peterson, N. S. (2008). A discussion of population invariance of equating. *Applied psychological Measurement*, 32, 98-101.

Yen, W. M., & Fitzpatrick, R. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport, CT: American Council on Education and Praeger Publishers.

Table 1. Number of Items and Score Points by Item Type for Spring 2019

	Test	EBSR*	MC	MS	SA	TDA	TE	Total Score Point
Number of Items	EL6	2	21			1	13	37
	EL8	4	23	3		1	3	39
	MA6		31		8		7	46
	MA8		32		5		9	46
Score Points	EL6	4	21			4	23	52
	EL8	8	28	6		4	6	52
	MA6		31		8		7	46
	MA8		32		5		9	46

*EBSR: evidence based selected response; MC: multiple- choice; MS; multiple-selected; SA: short answer, TDA: text-dependent analysis; TE; technology-enhanced

Table 2. Number of Items and Score Points by Item Type for Spring 2021

	Test	EBSR	MC	MS	SA	TDA	TE	Total Score Point
Number of Items	EL6	2	23				14	39
	EL8	4	29	4			3	40
Score Points	EL6	4	23				24	51
	EL8	8	29	8			6	51

Table 3. Sample Sizes for Spring 2019 and 2021 Subgroups

			EL6		EL8		MA6		MA8	
Subgroup	Category	Ab*	2019	2021	2019	2021	2019	2021	2019	2021
Population		P	65,212	55,198	62,817	56,411	65,355	55,181	62,963	56,455
Gender	Female	F	31,803	27,007	30,705	27,354	31,867	27,006	30,771	27,397
	Male	M	33,409	28,191	32,112	29,057	33,488	28,175	32,192	29,058
SES	No	N	37,069	34,723	37,915	36,416	37,117	34,709	37,966	36,430
	YES	S	28,143	20,475	24,902	19,995	28,238	20,472	24,997	20,025
Ethnicity	Hispanic	H	8,782	7,043	8,016	7,038	8,865	7,073	8,101	7,065
	African American	B	6,849	4,105	6,226	4,256	6,866	4,073	6,246	4,261
	White	W	43,509	38,769	43,135	40,135	43,518	38,756	43,160	40,127
LEP	No	Z	61,231	52,252	59,841	53,595	61,276	52,205	59,895	53,598
	YES	L	3,981	2,946	2,976	2,816	4,079	2,976	3,068	2,857

*Abbreviation: P:population; F:Female; M:Male; N: Non-SES; S: SES; H:Hispanic; B: African American; W: White; Z: No LEP; L: LEP

Table 4. Reliability and SEM by Subgroups for ELA Grades 6 and 8

Grade	Category*	2019 Administration					2021 Administration				
		N	Mean	SD	Rel**	SEM	N	Mean	SD	Rel**	SEM
6	P	65,212	29.74	9.32	0.88	3.23	55,198	29.03	9.41	0.88	3.26
	F	31,803	30.95	9.04	0.88	3.13	27,007	29.96	9.16	0.88	3.17
	M	33,409	28.58	9.44	0.89	3.13	28,191	28.14	9.55	0.89	3.17
	N	37,069	32.61	8.52	0.87	3.07	34,723	31.32	8.87	0.87	3.20
	Y	28,143	25.96	8.98	0.87	3.24	20,475	25.13	9.00	0.87	3.24
	H	8,782	26.40	8.91	0.87	3.21	7,043	25.15	8.91	0.87	3.21
	B	6,849	22.30	8.50	0.85	3.29	4,105	21.90	8.49	0.85	3.29
	W	43,509	31.69	8.72	0.87	3.14	38,769	30.58	9.01	0.88	3.12
	Z	61,231	30.25	9.22	0.88	3.19	52,252	29.52	9.28	0.88	3.21
	L	3,981	21.82	7.05	0.79	3.23	2,946	20.28	6.94	0.78	3.26
8	P	62,817	30.24	10.21	0.90	3.23	56,411	30.19	10.11	0.90	3.20
	F	30,705	31.65	9.91	0.90	3.13	27,354	31.37	9.83	0.90	3.11
	M	32,112	28.88	10.30	0.90	3.26	29,057	29.09	10.24	0.90	3.24
	N	37,915	33.05	9.46	0.89	3.14	36,416	32.38	9.61	0.89	3.19
	Y	24,902	25.95	9.80	0.89	3.25	19,995	26.21	9.77	0.89	3.24
	H	8,016	26.72	9.83	0.89	3.26	7,038	26.66	9.82	0.89	3.26
	B	6,226	22.43	9.28	0.88	3.21	4,256	22.86	9.32	0.88	3.23
	W	43,135	32.08	9.66	0.89	3.20	40,135	31.69	9.72	0.89	3.22
	Z	59,841	30.71	10.09	0.90	3.19	53,595	30.66	10.00	0.90	3.16
	L	2,976	20.80	7.57	0.82	3.21	2,816	21.26	7.62	0.81	3.32

*P:population; F:Female; M:Male; N: Non-SES; Y: SES; H:Hispanic; B: African American; W: White; Z: No LEP; L: LEP
**Yellow highlights indicate < 0.80

Table 5. Reliability and SEM by Subgroups for Mathematics Grades 6 and 8

Grade	Category*	2019 Administration					2021 Administration				
		N	Mean	SD	Rel**	SEM	N	Mean	SD	Rel**	SEM
6	P	65,355	21.49	9.97	0.92	2.82	55,181	19.89	9.53	0.91	2.86
	F	31,867	21.51	9.68	0.91	2.90	27,006	19.59	9.24	0.90	2.92
	M	33,488	21.46	10.24	0.92	2.90	28,175	20.18	9.80	0.92	2.77
	N	37,117	24.73	9.77	0.91	2.93	34,709	22.37	9.61	0.91	2.88
	Y	28,238	17.22	8.53	0.89	2.83	20,472	15.70	7.77	0.87	2.80
	H	8,865	17.35	8.45	0.89	2.80	7,073	15.35	7.56	0.86	2.83
	B	6,866	13.75	7.01	0.84	2.80	4,073	12.53	5.98	0.79	2.74
	W	43,518	23.67	9.75	0.91	2.92	38,756	21.64	9.47	0.91	2.84
	Z	61,276	22.01	9.97	0.92	2.82	52,205	20.34	9.53	0.91	2.86
	L	4,079	13.68	5.94	0.78	2.79	2,976	12.10	5.14	0.72	2.72
8	P	62,963	20.00	9.93	0.92	2.81	56,455	18.77	9.56	0.91	2.87
	F	30,771	20.28	9.64	0.91	2.89	27,397	18.75	9.21	0.90	2.91
	M	32,192	19.72	10.20	0.92	2.88	29,058	18.80	9.88	0.92	2.79
	N	37,966	22.84	9.92	0.91	2.98	36,430	21.03	9.75	0.91	2.92
	Y	24,997	15.67	8.26	0.88	2.86	20,025	14.67	7.67	0.86	2.87
	H	8,101	15.81	8.19	0.88	2.84	7,065	14.70	7.70	0.87	2.78
	B	6,246	12.38	6.86	0.84	2.74	4,261	11.66	6.11	0.8	2.73
	W	43,160	21.90	9.78	0.91	2.93	40,127	20.34	9.54	0.91	2.86
	Z	59,895	20.40	9.94	0.92	2.81	53,598	19.16	9.60	0.91	2.88
	L	3,068	12.18	5.64	0.76	2.76	2,857	11.61	5.09	0.71	2.74

*P:population; F:Female; M:Male; N: Non-SES; Y: SES; H:Hispanic; B: African American; W: White; Z: No LEP; L: LEP
**Yellow highlights indicate < 0.80

Table 6. REMSD for Subgroups across Content Areas with Unrounded Equated Scores

Year	Test	Gender	SES	Ethnicity	LEP
2019	EL6	0.0235	0.0097	0.0116	0.0095
	EL8	0.0351	0.0069	0.0118	0.0049
	MA6	0.0105	0.0052	0.0099	0.0091
	MA8	0.0104	0.0025	0.0079	0.0067
2021	EL6	0.0060	0.0047	0.0061	0.0060
	EL8	0.0054	0.0042	0.0056	0.0058
	MA6	0.0264	0.0144	0.0296	0.0307
	MA8	0.0448	0.0262	0.044	0.0445

Table 7. REMSD for Subgroups across Content Areas with Rounded Equated Scores

Year	Test	Gender	SES	Ethnicity	LEP
2019	EL6	0.0548	0.0291	0.0328	0.0234
	EL8	0.0554	0.0144	0.0137	0.0037
	MA6	0.0234	0.0116	0.0177	0.0137
	MA8	0.0305	0.0111	0.0252	0.0173
2021	EL6	0	0	0	0.0005
	EL8	0	0	0	0.0006
	MA6	0	0	0.0127	0.0186
	MA8	0	0	0	0.0008

Table 8. Effect Size for Unrounded Equated Scores between Sub-population and Population for Spring 2021

Category	Content	Subgroup Mean	Population Mean	Subgroup SD	Population SD	Diff_ Mean	Effect Size*
F	EL6	29.96	29.96	9.16	9.16	0.00	0.00
	EL8	31.36	31.37	9.82	9.82	-0.01	0.00
	MA6	19.58	19.59	9.23	9.22	-0.01	0.00
	MA8	18.73	18.75	9.18	9.17	-0.02	0.00
M	EL6	28.15	28.15	9.55	9.55	0.00	0.00
	EL8	29.09	29.09	10.24	10.24	0.00	0.00
	MA6	20.20	20.19	9.76	9.77	0.01	0.00
	MA8	18.85	18.83	9.80	9.81	0.02	0.00
NO SES	EL6	31.33	31.33	8.88	8.88	0.00	0.00
	EL8	32.38	32.38	9.61	9.61	0.00	0.00
	MA6	22.37	22.37	9.59	9.60	0.00	0.00
	MA8	21.01	21.02	9.73	9.72	-0.01	0.00
SES	EL6	25.13	25.14	9.00	8.99	0.00	0.00
	EL8	26.22	26.22	9.76	9.76	0.00	0.00
	MA6	15.66	15.71	7.78	7.73	-0.05	-0.01
	MA8	14.69	14.73	7.57	7.55	-0.03	0.00
H	EL6	25.16	25.16	8.89	8.89	0.00	0.00
	EL8	26.67	26.66	9.81	9.81	0.00	0.00
	MA6	15.26	15.36	7.60	7.52	-0.11	-0.01
	MA8	14.70	14.74	7.65	7.61	-0.03	0.00
B	EL6	21.92	21.91	8.46	8.46	0.00	0.00
	EL8	22.87	22.87	9.32	9.30	0.00	0.00
	MA6	12.38	12.56	6.07	5.89	-0.18	-0.03
	MA8	11.79	11.82	5.93	5.86	-0.03	-0.01
W	EL6	30.59	30.59	9.03	9.02	0.00	0.00
	EL8	31.69	31.70	9.72	9.72	0.00	0.00
	MA6	21.65	21.64	9.45	9.46	0.00	0.00
	MA8	20.31	20.34	9.53	9.51	-0.02	0.00
No LEP	EL6	29.53	29.53	9.29	9.29	0.00	0.00
	EL8	30.67	30.67	10.00	10.00	0.00	0.00
	MA6	20.34	20.34	9.51	9.52	0.00	0.00
	MA8	19.16	19.17	9.54	9.54	0.00	0.00
LEP	EL6	20.31	20.30	6.87	6.89	0.02	0.00
	EL8	21.31	21.27	7.51	7.60	0.04	0.01
	MA6	11.63	12.12	5.27	5.06	-0.49	-0.09
	MA8	11.76	11.69	4.97	4.89	0.07	0.01

*Yellow highlights show |ES| > 0

Table 9. Effect Size for Rounded Equated Scores between Sub-population and Population for Spring 2021

Group	Content	Subgroup Mean	Population Mean	Subgroup SD	Population SD	Diff_ Mean	Effect Size*
F	EL6	29.96	29.96	9.16	9.16	0.00	0.00
	EL8	31.37	31.37	9.83	9.83	0.00	0.00
	MA6	19.61	19.61	9.20	9.20	0.00	0.00
	MA8	18.81	18.81	9.12	9.12	0.00	0.00
M	EL6	28.14	28.14	9.55	9.55	0.00	0.00
	EL8	29.09	29.09	10.24	10.24	0.00	0.00
	MA6	20.21	20.21	9.76	9.76	0.00	0.00
	MA8	18.89	18.89	9.76	9.76	0.00	0.00
N	EL6	31.32	31.32	8.87	8.87	0.00	0.00
	EL8	32.38	32.38	9.61	9.61	0.00	0.00
	MA6	22.38	22.38	9.59	9.59	0.00	0.00
	MA8	21.07	21.07	9.68	9.68	0.00	0.00
Y	EL6	25.13	25.13	9.00	9.00	0.00	0.00
	EL8	26.21	26.21	9.76	9.76	0.00	0.00
	MA6	15.74	15.74	7.71	7.71	0.00	0.00
	MA8	14.82	14.82	7.49	7.49	0.00	0.00
H	EL6	25.15	25.15	8.90	8.90	0.00	0.00
	EL8	26.66	26.66	9.82	9.82	0.00	0.00
	MA6	15.40	15.40	7.50	7.49	0.00	0.00
	MA8	14.83	14.83	7.55	7.55	0.00	0.00
B	EL6	21.90	21.90	8.49	8.49	0.00	0.00
	EL8	22.87	22.87	9.32	9.32	0.00	0.00
	MA6	12.41	12.60	6.01	5.88	-0.19	-0.03
	MA8	11.94	11.94	5.82	5.82	0.00	0.00
W	EL6	30.58	30.58	9.01	9.01	0.00	0.00
	EL8	31.69	31.69	9.72	9.72	0.00	0.00
	MA6	21.66	21.66	9.44	9.44	0.00	0.00
	MA8	20.39	20.39	9.46	9.46	0.00	0.00
NO LEP	EL6	29.52	29.52	9.28	9.28	0.00	0.00
	EL8	30.66	30.66	10.00	10.00	0.00	0.00
	MA6	20.36	20.36	9.50	9.50	0.00	0.00
	MA8	19.23	19.23	9.49	9.49	0.00	0.00
LEP	EL6	20.28	20.28	6.94	6.93	0.00	0.00
	EL8	21.26	21.26	7.62	7.61	0.00	0.00
	MA6	11.59	12.17	5.31	5.04	-0.58	-0.11
	MA8	11.81	11.81	4.86	4.85	0.00	0.00

*Yellow highlights show |ES| > 0

Table 10. Effect Size for MLE between Sub-population and Population for Spring 2021

Group	Content	Subgroup Mean	Population Mean	Subgroup SD	Population SD	Diff_ Mean	Effect Size*
F	EL6	-0.03	-0.02	1.09	1.09	-0.01	-0.01
	EL8	0.57	0.58	1.29	1.29	-0.01	-0.01
	MA6	-0.24	-0.24	1.22	1.22	0.00	0.00
	MA8	0.58	0.58	1.16	1.17	0.00	0.00
M	EL6	-0.24	-0.25	1.15	1.15	0.01	0.01
	EL8	0.27	0.26	1.33	1.34	0.01	0.00
	MA6	-0.20	-0.20	1.30	1.31	0.00	0.00
	MA8	0.52	0.52	1.34	1.32	0.00	0.00
N	EL6	0.14	0.14	1.06	1.06	0.00	0.00
	EL8	0.70	0.70	1.27	1.27	0.00	0.00
	MA6	0.12	0.11	1.10	1.14	0.01	0.01
	MA8	0.85	0.85	1.13	1.16	0.01	0.01
Y	EL6	-0.60	-0.60	1.09	1.09	0.00	0.00
	EL8	-0.11	-0.10	1.26	1.26	-0.01	-0.01
	MA6	-0.82	-0.78	1.35	1.28	-0.04	-0.03
	MA8	-0.02	0.01	1.28	1.23	-0.03	-0.02
H	EL6	-0.60	-0.60	1.07	1.07	0.00	0.00
	EL8	-0.04	-0.03	1.26	1.27	-0.01	-0.01
	MA6	-0.86	-0.81	1.32	1.25	-0.05	-0.04
	MA8	0.00	0.03	1.25	1.21	-0.03	-0.03
B	EL6	-1.01	-1.00	1.07	1.07	0.00	0.00
	EL8	-0.55	-0.54	1.24	1.24	-0.01	-0.01
	MA6	-1.57	-1.40	1.64	1.36	-0.17	-0.12
	MA8	-0.66	-0.55	1.46	1.27	-0.11	-0.08
W	EL6	0.05	0.05	1.07	1.07	0.00	0.00
	EL8	0.61	0.60	1.27	1.27	0.00	0.00
	MA6	0.04	0.03	1.11	1.14	0.01	0.01
	MA8	0.77	0.77	1.13	1.15	0.00	0.00
No LEP	EL6	-0.08	-0.08	1.11	1.11	0.00	0.00
	EL8	0.47	0.47	1.31	1.31	0.00	0.00
	MA6	-0.15	-0.16	1.23	1.24	0.00	0.00
	MA8	0.60	0.60	1.23	1.23	0.00	0.00
LEP	EL6	-1.17	-1.18	0.89	0.90	0.01	0.01
	EL8	-0.71	-0.71	1.00	1.01	0.00	0.00
	MA6	-1.54	-1.34	1.39	1.20	-0.20	-0.15
	MA8	-0.47	-0.44	1.18	1.14	-0.03	-0.03

*Yellow highlights show |ES| > 0

Figure 1. Differences for Pre vs. Unrounded Post Equating for Population

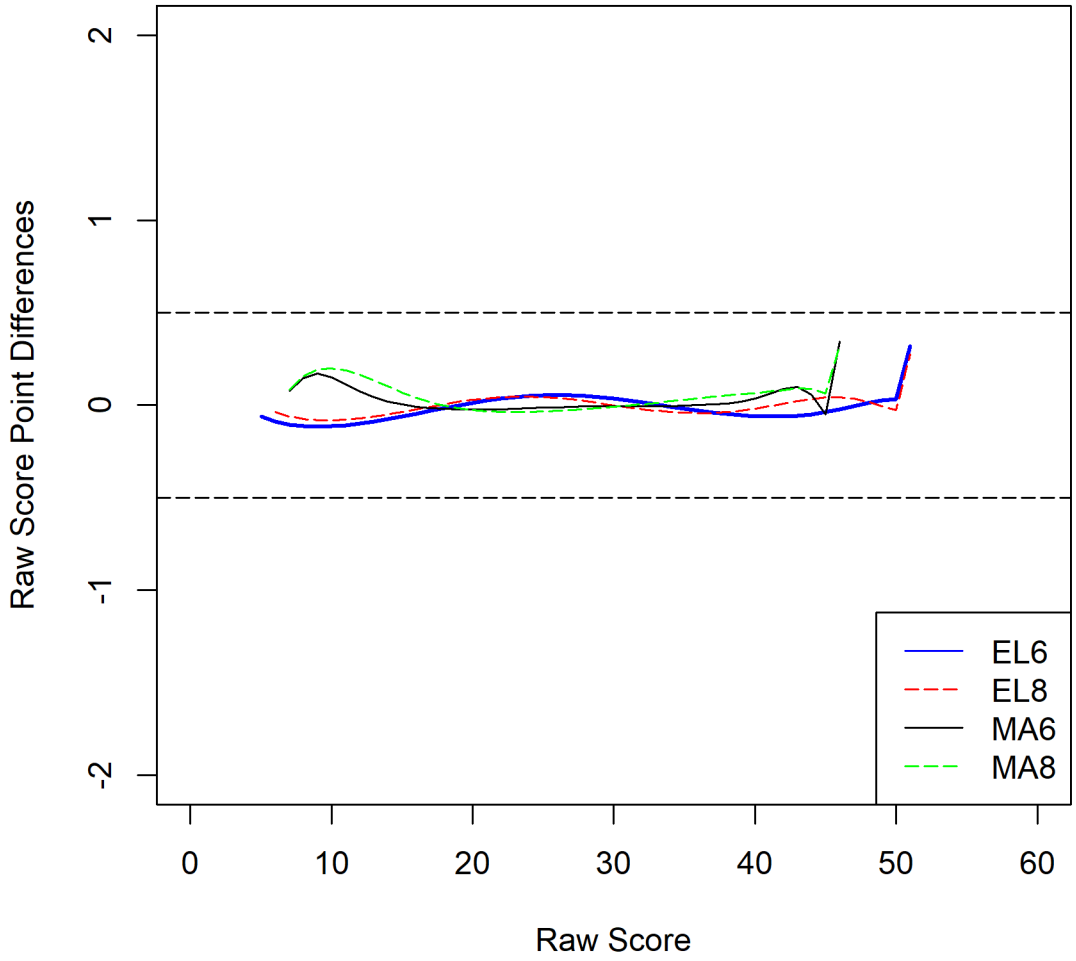
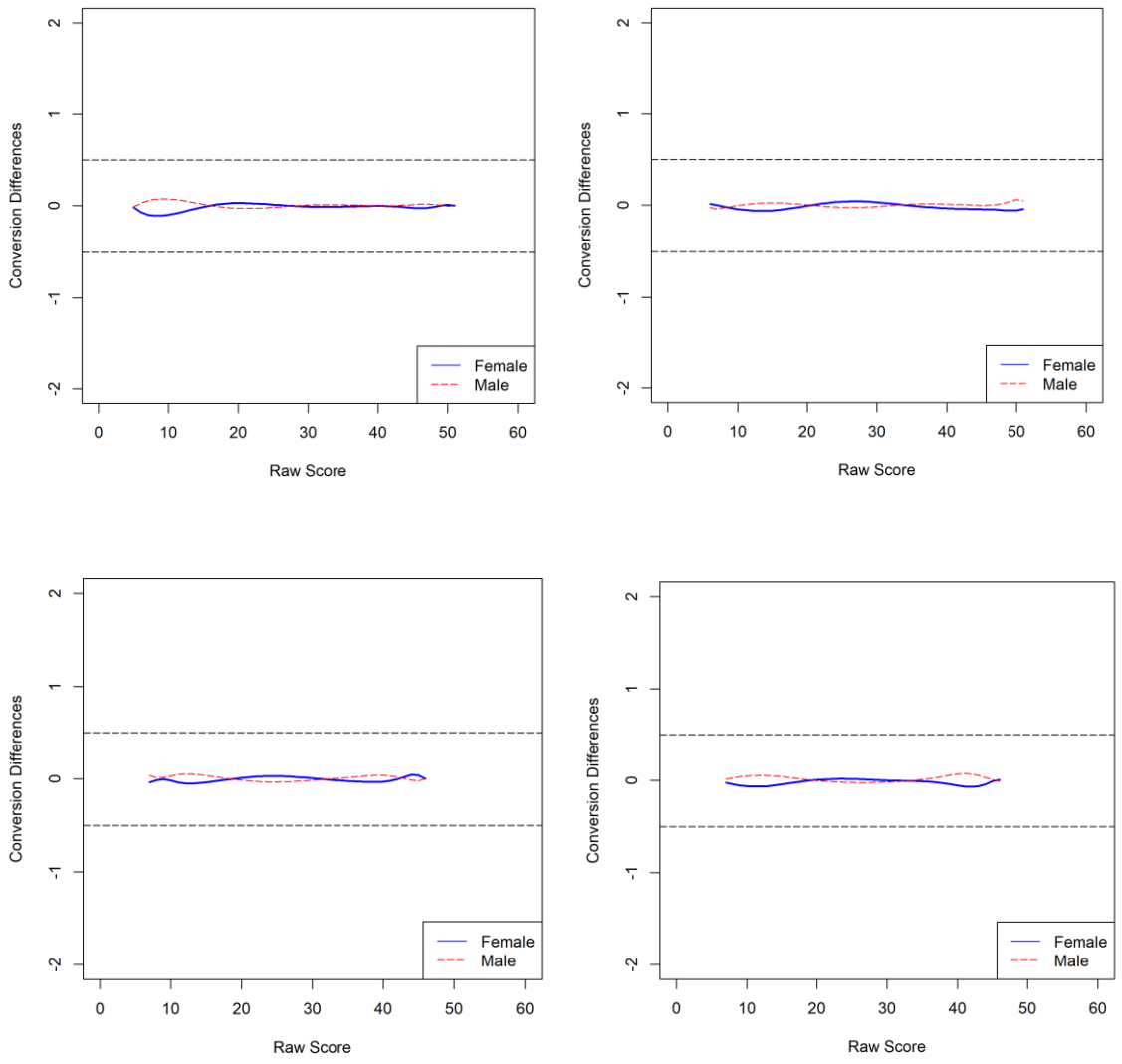
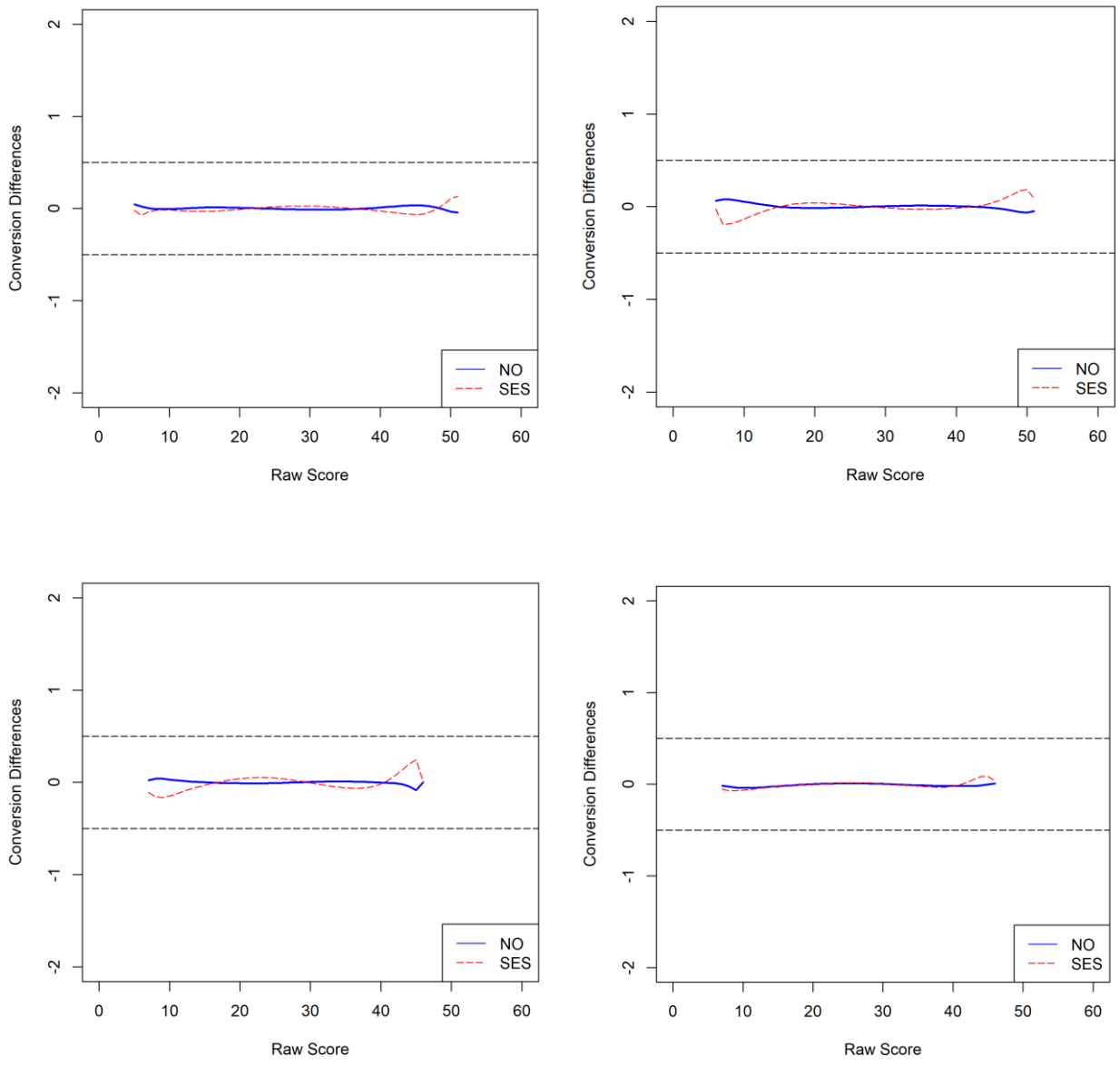


Figure 2. Unrounded Equated Score Differences by Gender Groups



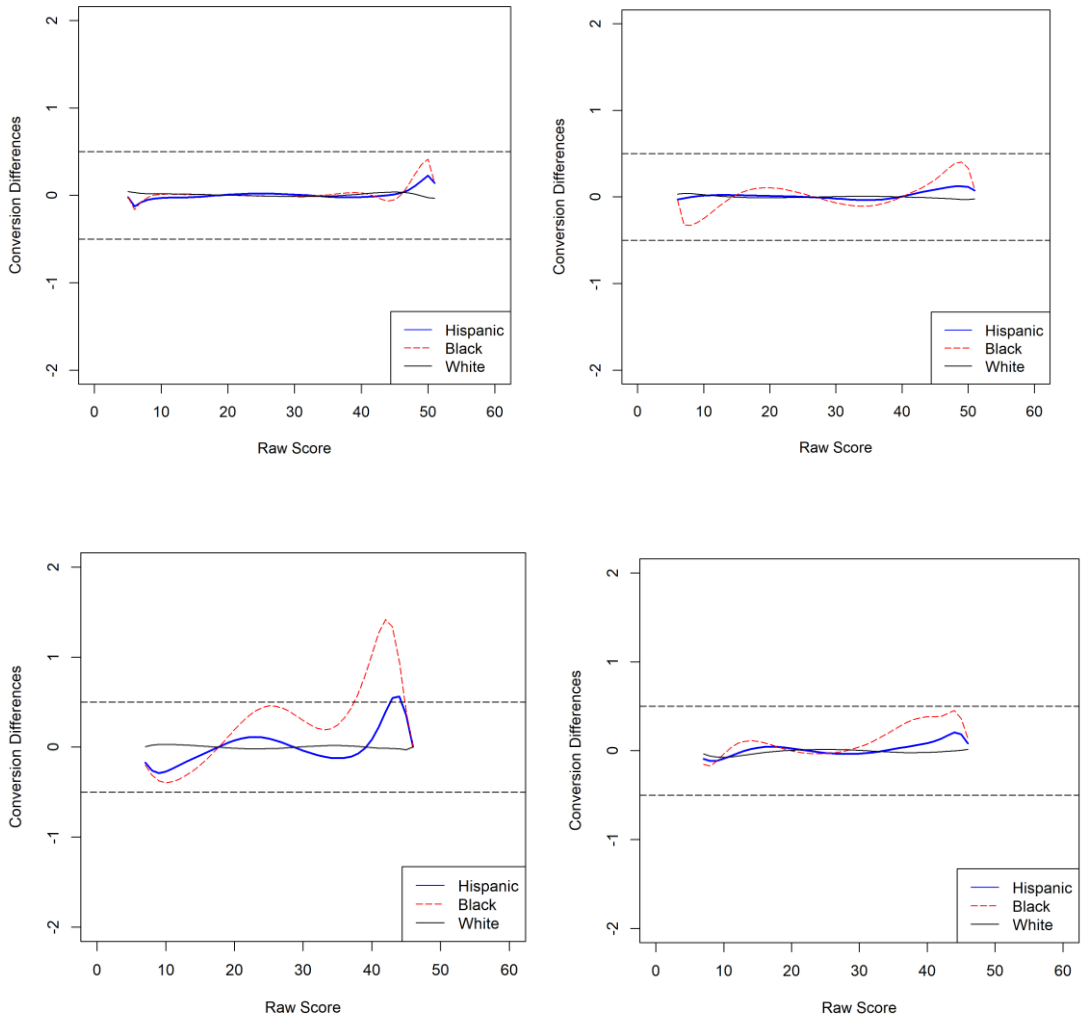
- Upper: ELA6 (left) and ELA8 (right)
- Lower: MA6 (left) and MA8 (right)

Figure 3. Unrounded Equated Score Differences by SES



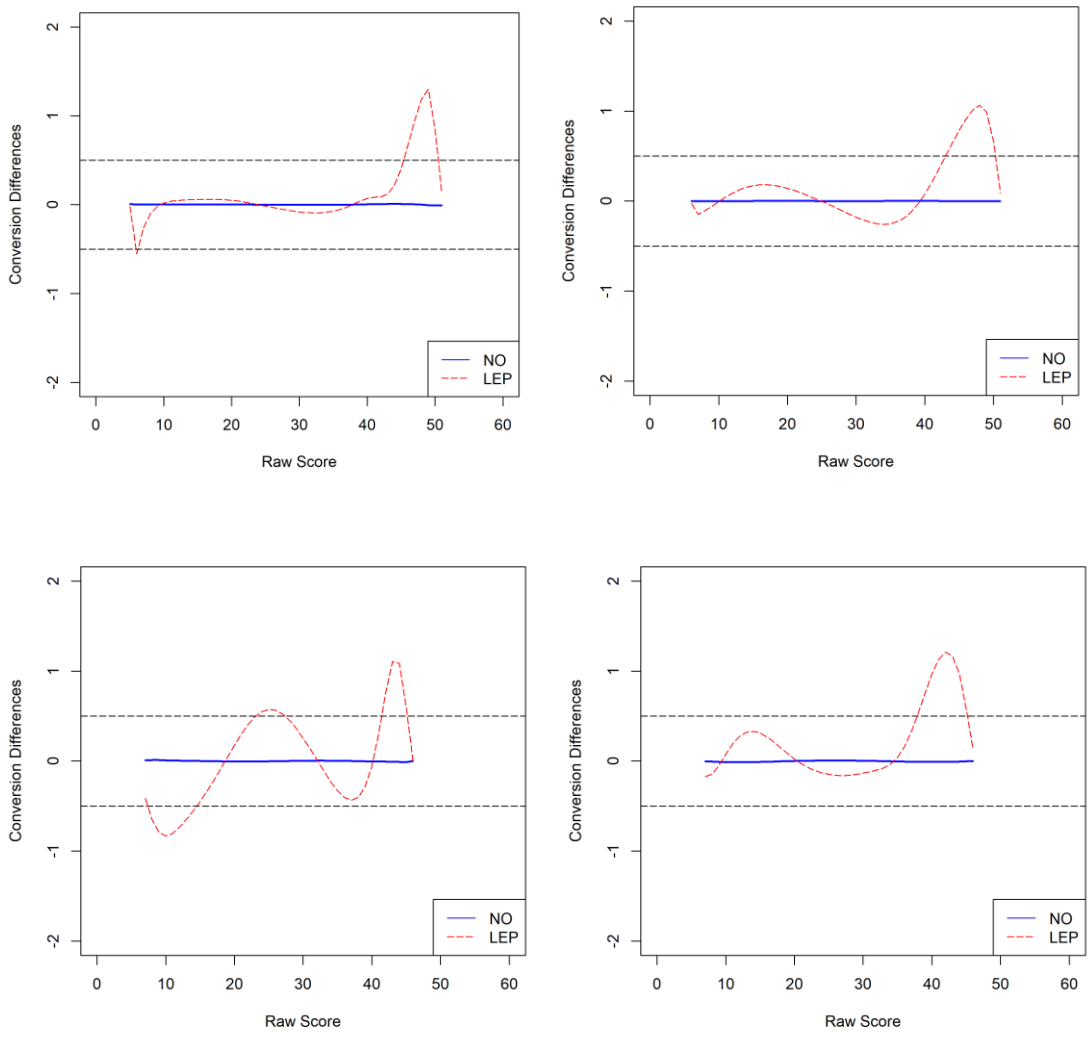
- Upper: ELA6 (left) and ELA8 (right)
- Lower: MA6 (left) and MA8 (right)

Figure 4. Unrounded Equated Score Differences by Ethnicity



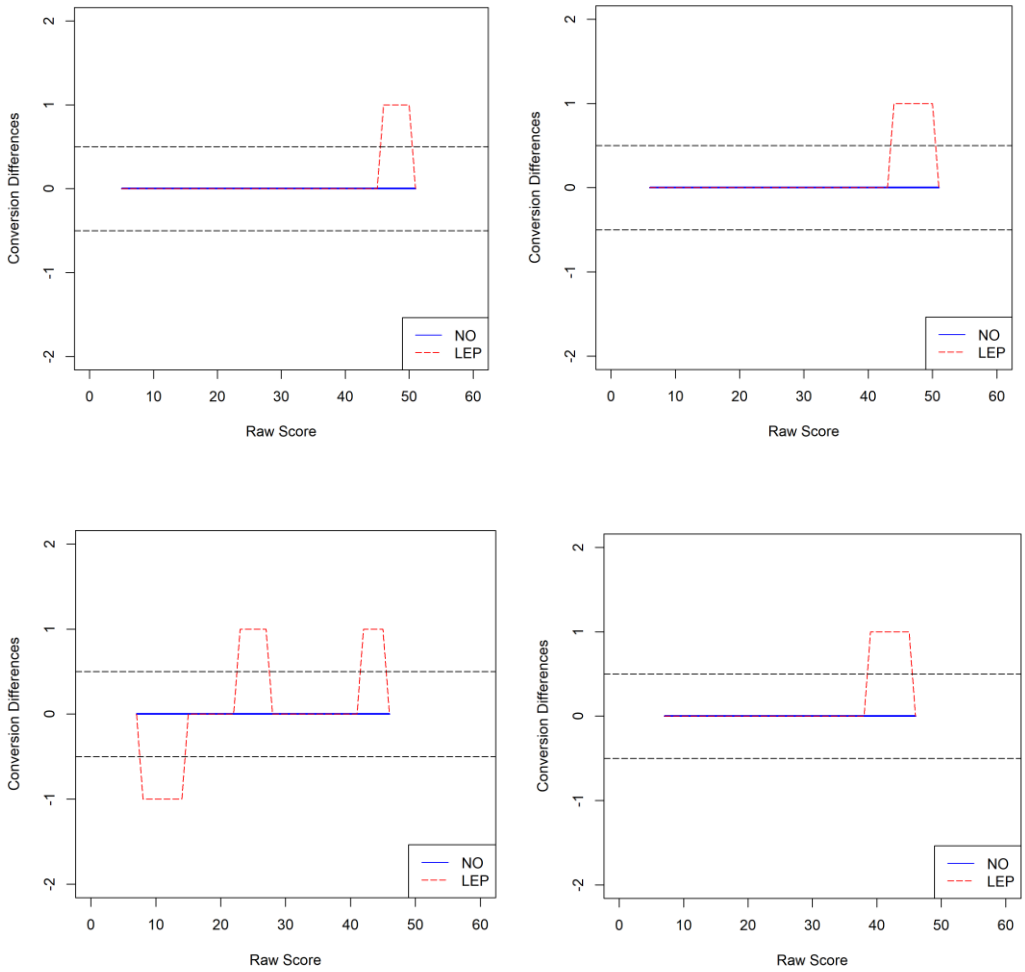
- Upper: ELA6 (left) and ELA8 (right)
- Lower: MA6 (left) and MA8 (right)

Figure 5. Unrounded Equated Score Differences by LEP



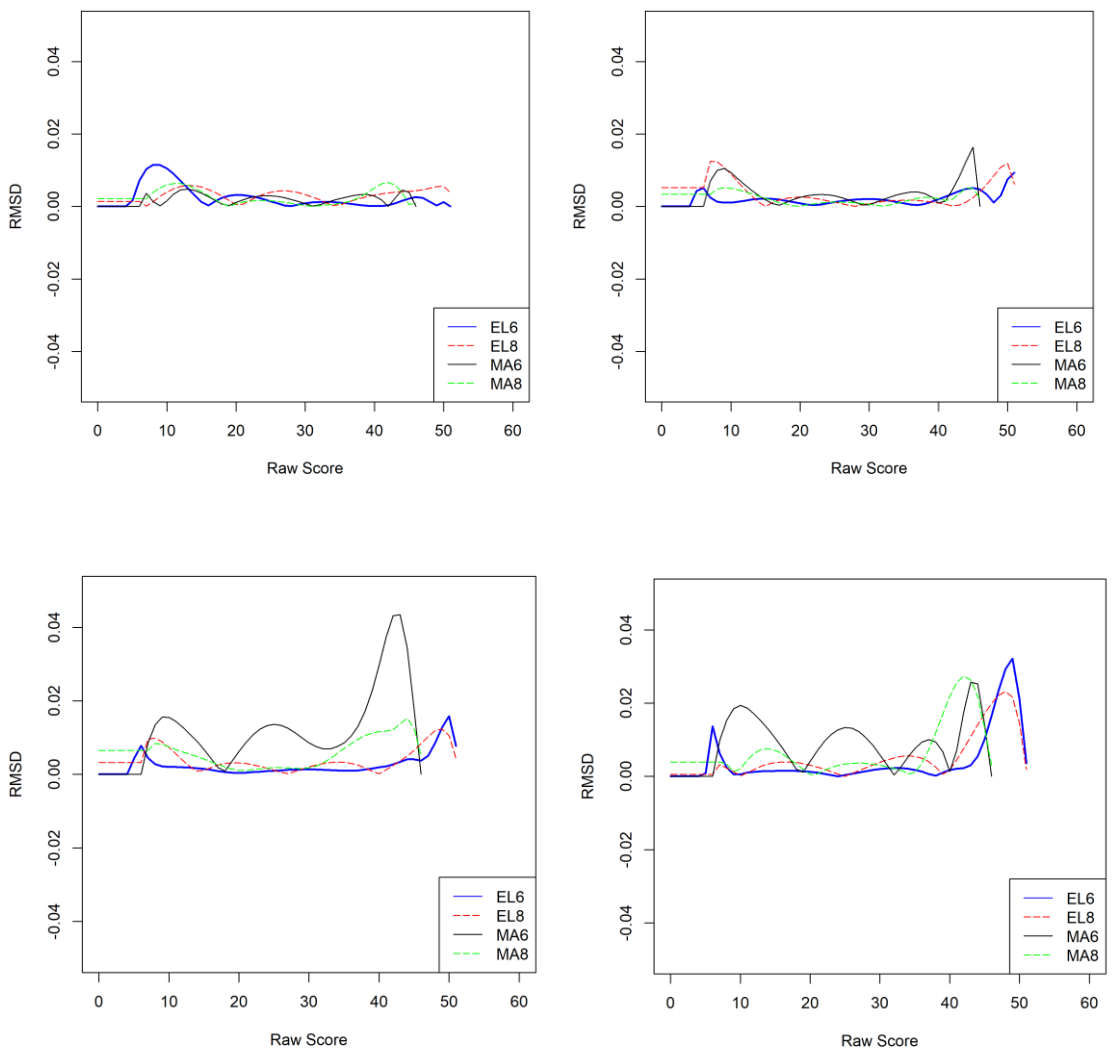
- Upper: ELA6 (left) and ELA8 (right)
- Lower: MA6 (left) and MA8 (right)

Figure 6. Rounded Equated Score Differences by LEP



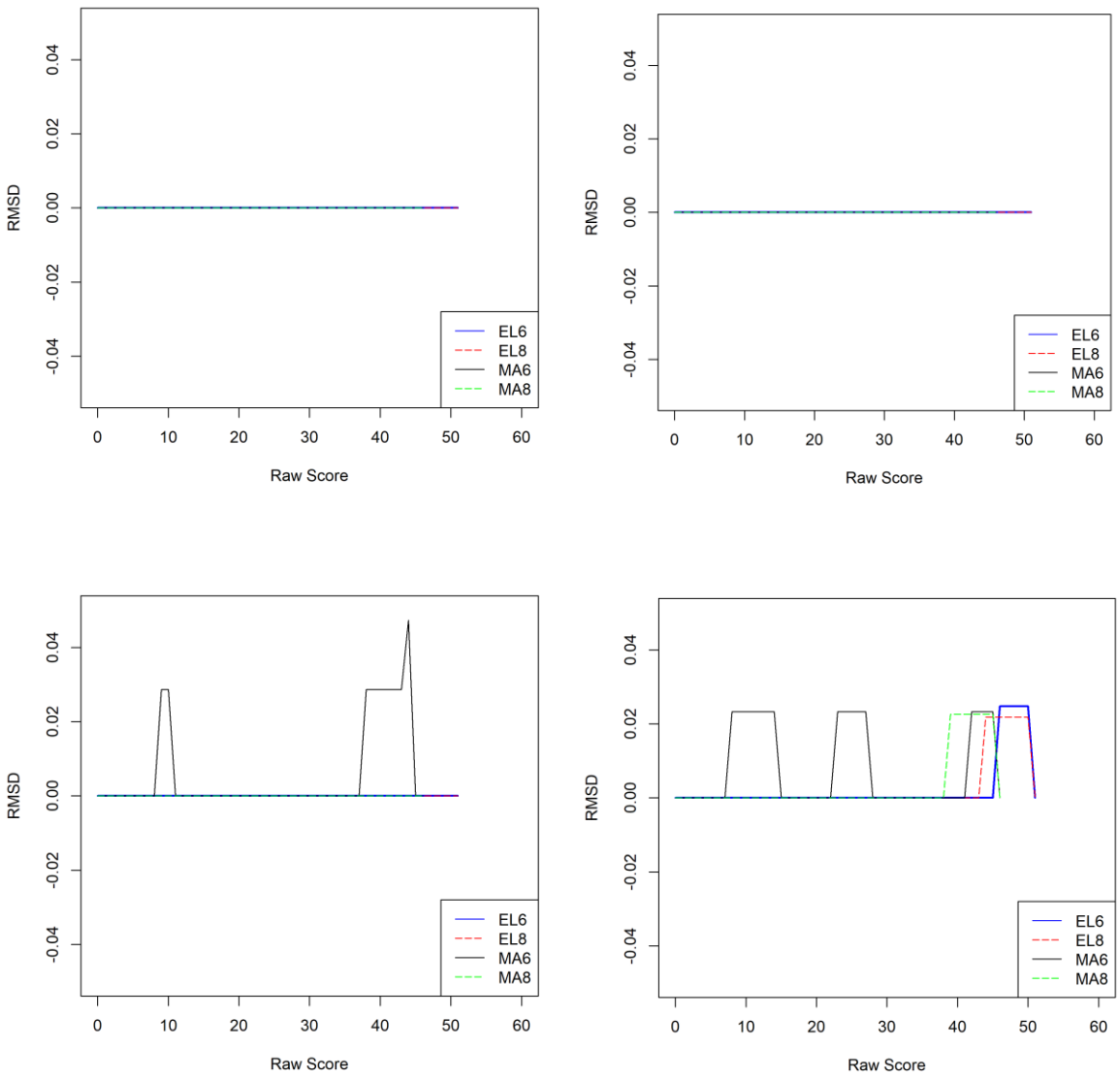
- Upper: ELA6 (left) and ELA8 (right)
- Lower: MA6 (left) and MA8 (right)

Figure 7. RMSD for Unrounded Equated Scores for Gender, SES, Ethnicity, & Location



- Upper: Gender (left) and SES (right)
- Lower: Ethnicity (left) and LEP (right)

Figure 8. RMSD for Rounded Equated Scores for Ethnicity and LEP



- Upper: Gender (left) and SES (right)
- Lower: Ethnicity (left) and LEP (right)