



To: Michael J. Muenks, Shaun Bates

**Dept: Missouri Department of
Elementary and Secondary Education**

Location: Missouri

**Subject: Missouri Assessment Program
Item Comparability Study for Math and
ELA (Version 2)**

From DRC: Joanna Tomkowicz

Dept: Psychometric Services

Location: Maple Grove, Minnesota

Date: July 30, 2015

Missouri Assessment Program Item Comparability Study for Math and ELA

Version 2

Version History

This document is intended to replace the previous document with the same title and dated June 19, 2015. The update includes results from classical item analysis, calibration and equating of ELA Grade 8 and Mathematics Grade 8 assessments. The aforementioned analyses were completed on the MAP 2015 census data. The previous document included sample-based results of item analysis, calibration and equating of ELA Grade 8 and Mathematics Grade 8 core (CA) forms and sample-based results of item analysis of ELA Grade 8 and Mathematics Grade 8 performance assessment (PA) forms as at the time of preparing the original report we did not acquire sufficient representative samples of student who took both core and performance assessments forms to include PA forms in calibration and equating. The same methodology as implemented for the original data analysis was used in the final ELA Grade 8 and Mathematics Grade 8 analysis.

Purpose

The goal of these analyses was to examine the item level performance to verify that the items were performing as expected. Specifically, this study was limited to the grade 3 through 8 Mathematics and ELA items included in core (CA) and performance assessment (PA) forms on 2015 MAP test administration.

Sampling and census data

Because the scored census data were not available during the time of initial data analysis in June 2015, CTB Research selected samples from each of the test forms that mirrored the 2014 demographic distribution of examinees using gender and ethnicity. Additionally, we selected samples that represented a range across the performance level distribution.

Generally samples were selected with a minimum N count of 1,000 to 1,500 students per test form. ELA Grade 8 and Mathematics Grade 8 data analyses, as explained in the Version History paragraph, were repeated on the census data available in July 1015.

Characteristics Examined

To examine the item level performance, CTB Research reviewed classical item statistics, calibration results, and equating results.

Item-Level Statistics

The following statistics were examined for each item. The following criteria were used to flag items for review by Research and Content experts. Items with:

1. Classical item difficulty, p-values ≤ 0.20 or ≥ 0.90 the item was flagged for review for being too difficult or too easy.
2. Omit rates $\geq 5\%$
3. Low point biserial correlations, $p_{bis} \leq 0.10$
4. Positive point biserial correlations on a distractor
5. Poor fit using Q1 statistic

Evaluation of Equating Results and IRT Regression Curves

Much like in anchor item evaluation, CTB Research compared the item parameter estimates from SBAC to the transformed estimates from the current administration using Stocking and Lord's (1983) test characteristic curve (TCC) method. The TCC method determines the scaling constants by minimizing the quadratic loss function (F). Using this method, outlying items were identified by plotting the input and estimated item parameters along with the line of best fit. Items with an absolute difference of parameters greater than two times the root mean squared difference were flagged.

Additionally, we examined the differences between the item characteristic regression curves using the parameter estimates SBAC and those from the current calibration. The differences between the curves are evaluated using the following statistics:

- UnWtd Mean = Average signed difference in estimated probability.
- UnWtd Mean Abs Dif = Average Absolute (unsigned) difference in estimated probability.
- UnWtd RMSD = Root mean squared difference.
- Wtd Mean = Weighted average signed difference in estimated probability.
- Wtd Mean Abs = Weighted average Absolute (unsigned) difference in estimated probability.
- WtdRMSD = Weighted Root mean squared difference

For the six statistics listed above, differences greater than $+0.10$ are considered large, and differences between $+0.07$ and $.10$ are considered moderate. Additionally, the Maximum

Absolute difference (MaxAbsDifPC) will be identified. For MaxAbsDifPC, large differences are those greater than +.15, and moderate differences are all differences between +.125 and .15.

Items flagged for large differences on four of the seven statistics (listed above) considered when examining the differences between the IRT regression curves were examined for further review.

For items flagged on multiple statistical criterion, the Content team was asked to review the item for content coverage and to also verify that the key was correctly identified and the scoring rules were also correct.

High-Level Summary of Findings

Note that the results for grades 3 through 7, presented in this document, are based on representative samples of students. Results for grade 8 are census-based.

ELA

All ELA grades and forms have been reviewed for classical statistics. ELA core forms for grades 3, 4, 6, 7, and 8 were calibrated and equated. In addition ELA grade 5 and 8 calibration and equating was conducted on combined core and PA forms. All calibrations were conducted using concurrent calibration method within a grade. The equating was performed using the Stocking and Lord method.

Form reliabilities are well within the accepted range for high-stakes assessments and range from 0.86 to 0.92 for core forms and from 0.66 to 0.74 for PA forms.

Table 1 summarizes the equating results for ELA including number of iterations, value of F function, correlations between a-parameter input and estimates, correlations between b-parameter input and estimates, and the number of outliers.

Table 1. ELA equating results

Content	Forms	Grade	TCC results		Parameters Comparison Statistics			
					A Parameter		B Parameter	
			# of Iterations	F Value	Corr	# of RMSD Outliers	Corr	# of RMSD Outliers
ELA	Core	3	5	0.11202	0.75	3	0.78	5
ELA	Core	4	5	0.284726	0.80	2	0.93	3
ELA	Core and PA	5	8	0.442081	0.81	1	0.87	2
ELA	Core	6	4	0.379465	0.88	2	0.89	4
ELA	Core	7	7	0.16139	0.74	1	0.87	1
ELA	Core and PA	8	3	0.193209	0.90	4	0.88	2

Table 2 summarizes the number of items flagged using classical statistics criteria, model fit, and the IRT regression method. Flag indicators by item are provided in a separate file called: MAP Comp Study Summary V2.073015.xlsx.

Table 2: Item Flag Counts, ELA

Content	Form	Grade	# Items	# Flagged				
				LPval	Hpval	Low Pbis	Model Fit	Regression Curve
ELA	Core	3	101	3	0	0	13	6
ELA	Core	4	97	4	1	0	9	9
ELA	Core and PA	5	115	4	4	0	10	7
ELA	Core	6	102	7	0	0	4	4
ELA	Core	7	103	0	1	0	9	14
ELA	Core and PA	8	117	8	1	0	5	6

Math

All Mathematics grades and forms have been reviewed for classical statistics. The core forms for grades 3, 4, 6, 7, and 8 were calibrated and equated. Math grade 5 and 8 calibration and equating was conducted on combined core and PA forms. All calibrations were conducted using concurrent calibration method within a grade. The equating was performed using the Stocking and Lord method.

Form reliabilities range from 0.87 to 0.92 for core forms and from 0.65 to 0.74 for PA forms. These values are within the accepted range for high-stakes tests.

Table 3 summarizes the equating results for Math including number of iterations, value of F function, correlations between a-parameter input and estimates, correlations between b-parameter input and estimates, and the number of outliers.

Table 3. Mathematics equating results

Content	Forms	Grade	TCC results		Parameters Comparison Statistics			
					A Parameter		B Parameter	
			# of Iterations	F Value	Corr	# of RMSD Outliers	Corr	# of RMSD Outliers
Math	Core	3	4	0.080611	0.96	0	0.99	0
Math	Core	4	7	0.11883	0.85	1	0.98	0
Math	Core and PA	5	7	0.501404	0.91	1	0.94	0
Math	Core	6	7	0.384639	0.91	1	0.95	1
Math	Core	7	10	0.10997	0.93	0	0.97	1
Math	Core and PA	8	25	0.639081	0.95	1	0.98	1

Table 4 summarizes the number of items flagged using classical statistics criteria, model fit, and the IRT regression method. Flag indicators by item are provided in a separate file called: MAP Comp Study Summary V2.073015.xlsx.

Table 4: Item Flag Counts, Mathematics

Content	Form	Grade	# Items	# Flagged				
				LPval	Hpval	Low Pbis	Model Fit	Regression Curve
Math	Core	3	65	3	2	0	8	3
Math	Core	4	67	8	3	0	5	1
Math	Core and PA	5	81	11	1	0	8	1
Math	Core	6	66	15	1	1	5	1
Math	Core	7	62	8	0	0	9	2
Math	Core and PA	8	76	20	0	0	6	4

Summary and Recommendations

Typically, for both ELA and Math, items flagged for p-values had similar values to what we saw during form selection. For model fit issues, there was generally a misfit in the areas where there were few students at a certain part of the ability distribution. None of these issues are uncommon and were expected given the depth of the item pool used for selection.

The DRC/CTB Content team spot checked items of most concern and found no reason for their suppression. We uphold our recommendation to not suppress any items in 2015 MAP assessments.