An Investigation of Item Parameter Invariance Using Focused Calibration Samples for MAP Growth

November 2021

Wei He, NWEA Psychometric Solutions

nwea

Document History

Date	Version	Description
2021-09-23	0.1	Initial draft created by Wei He.
2021-11-09	1.0	Finalized by Patrick Meyer; published.

This report benefited from the editorial assistance of Kelly Rivard.

Table of Contents

List of Tables

List of Figures

Executive Summary

New MAP® Growth™ assessments are being developed that administer items more closely matched to the grade level of the student. However, MAP Growth items are calibrated with samples that typically consist of students from a variety of grades, including the target grade to which an item is aligned. While this choice of calibration sample is reasonable given that the current MAP Growth tests are not limited to assessing grade-specific content, a question arose about the appropriateness of using the current MAP Growth item parameter estimates in the new MAP Growth assessments. In other words, are the existing MAP Growth item parameter estimates invariant across different calibration samples?

To evaluate the MAP Growth item parameter invariance, two studies were conducted using focused calibration samples (i.e., a limited sample of examinees). Study 1 explored the degree to which on-grade only item parameter estimates were comparable to their original all-grade counterparts. Study 2 explored the degree to which item parameter estimates using responses from the target grade plus the adjacent grades were comparable to their on-grade and all-grade counterparts.

While the item difficulty estimates of a few items from the more focused calibration samples had different parameter estimates from their original counterparts, the parameter estimates of most items were comparable regardless of the calibration sample. These findings, in conjunction with the result of a recent study by Wan and Thum (2021) that used differential item functioning (DIF) analyses to reveal that MAP Growth items perform comparably across states and grades, provide the quantitative evidence of MAP Growth item parameter estimate invariance to support the use of the existing parameter estimates in the new MAP Growth assessments.

1. Introduction

Current MAP® Growth™ tests are not limited to assessing grade-specific content. Cross-grade goal structures allow the assessments to adapt to the ability levels of all students and provide precise measurement of students who are performing on, above, and below their grade levels. The existing item calibration approach supports this cross-grade use of items. It includes all students administered an item of interest, rather than limiting to students from the grade for which the item was designed. NWEA® has been developing new assessments that prioritize items that match a student's grade and have a difficulty close to the student's ability level. This prioritization of item grade level has raised questions about the suitability of existing item calibrations for tests that more closely align to a student's grade. Therefore, the purpose of this study is to evaluate the invariance of item difficulty estimates by comparing them from the current all-grade calibration samples to those from the more focused calibration samples.

MAP Growth field test items must be calibrated to the Rasch Unit (RIT) scale and pass item review criteria to become operational. They are embedded in fixed positions on operational tests and adaptively administered. Responses are continuously collected until the calibration sample size requirement is met to allow the field test items to be included in the calibration process. Although MAP Growth items are typically aligned to a target grade, they are exposed to students beyond the target grade. Consequently, the item parameter estimates are derived from samples of students at the item's target grade level and students beyond the target grade of an item. Item parameter estimates are then used to score students who are administered the item once it becomes operational, regardless of whether the target grade of an item matches the student's grade.

Invariance is a property of item response model parameters that in practice may or may not hold for parameter estimates derived from subgroups of a calibration sample. Evidence supporting item parameter invariance exists when items exhibit the same parameter estimates across subgroups under the same item response theory (IRT) model. In the studied context, item parameter invariance is the degree to which MAP Growth item parameter estimates derived from more focused calibration samples are comparable to each other and to the original item parameter estimates derived from a more broadly defined calibration sample that includes students from all grades. To evaluate item parameter invariance for MAP Growth items, two studies were conducted:

- Study 1 explored the degree to which the on-grade item parameter estimates are comparable to their original all-grade counterparts. On-grade item parameter estimates were derived only using responses from students in the target grade to which an item is aligned, whereas the original all-grade item parameter estimates were derived using responses from all students in a variety of grades who responded to the item.
- Study 2 explored the degree to which item parameter estimates using responses from three grades (i.e., the target grade plus the two adjacent grades, one below and one above) are comparable to their on-grade and all-grade counterparts.

2. Data and Methods

2.1. Study Sample

Table 2.1 presents the number of items included in the study across subjects and target grades. Items of interest included 2,960 unique items successfully calibrated in June 2021. The target grade alignment was assigned by NWEA content experts based on the Common Core State Standards (CCSS; National Governors Association Center for Best Practices & Council of Chief State School Officers [CCSSO], 2010). Because some items were aligned to multiple target grades and were therefore included more than once, the total number of items used in this study was 3,236 instead of 2,960. While Study 1 used all items in the table, Study 2 only used the items in Grades 4–7. Language Usage is not included because the number of Language Usage items available for this study was too small to yield meaningful results.

Items with calibration sample sizes less than 300 were excluded from both studies, as shown in Table 2.1. Calibration sample size has been found to play a key role in item parameter estimate accuracy. The use of test events from the target grade or the target grade + the adjacent grades entails a smaller calibration sample size than that used to derive the all-grade item parameter estimates. To mitigate the effects of calibration sample size on item parameter estimates, the minimum size of a calibration sample was set at 300 based on the Rasch item calibration literature that suggests item calibration estimates would be similar to each other as long as the sample size reaches 250 (Hambleton et al., 1991; He, 2015).

Table 2.1. Number of Items in the Study Sample

Target Grade	N_{item}							
	Total				Calibration Size ≥ 300			
	Math	Reading	Science	Total	Math	Reading	Science	Total
1	64	–	–	64	1	–	–	1
2	247	6	4	257	208	6	–	214
3	113	249	18	380	104	249	15	368
4	105	286	8	399	96	280	7	383
5	190	164	13	367	175	160	12	347
6	103	245	49	397	98	208	46	352
7	194	257	49	500	185	212	47	444
8	163	194	49	406	163	190	45	398
9	11	147	29	187	10	90	22	122
10	1	147	29	177	–	43	20	63
11	–	22	29	51	–	5	16	21
12	–	22	29	51	–	–	3	3
Total	1,191	1,739	306	3,236	1,040	1,443	233	2,716

Note. Study 1 used all items in this table, whereas Study 2 only used the items in Grades 4–7.

2.2. Study Procedures

Items in both studies were recalibrated with standard MAP Growth item calibration procedure but different calibration samples. Table 2.2 summarizes the notations and descriptions of the item parameter estimates in the two studies.

Table 2.2. Item Parameter Estimates Compared in Both Studies

Item Difficulty Estimate	Description
$\hat{b}_{base}$	The original all-grade item difficulty estimate derived from the current all-grade calibration sample
$\hat{b}_{ongrade}$	The on-grade item difficulty estimate derived from responses in the target grade to which an item is aligned
$\hat{b}_{3grades}$	The item difficulty estimate derived from responses in the target grade to which an item is aligned + the two adjacent grades, one below and one above
$\hat{b}_{lowgrade}$	The item difficulty estimate derived from the lower target grade for items aligned with multiple target grades
$\hat{b}_{upgrade}$	The item difficulty estimate derived from the upper target grade for items aligned with multiple target grades

2.2.1. Study 1: On-Grade Item Calibration

Items in Study 1 underwent the standard MAP Growth item calibration procedure but with on-grade samples.[1] For example, if the target grade of an item was Grade 4, its item calibration used responses from Grade 4 students only. To explore how the on-grade parameter estimates (denoted as $\hat{b}_{ongrade}$) of items aligned to only one target grade compared to their original all-grade counterparts (denoted as $\hat{b}_{base}$), the following procedures were conducted:

1. Compute the differences between $\hat{b}_{ongrade}$ and $\hat{b}_{base}$.
2. Allocate items into one of the following difficulty parameter difference categories (in logit) based on $\hat{b}_{ongrade} - \hat{b}_{base}$ and aggregate the results by subject and grade. The square bracket "[" or "]" indicates inclusive, whereas the bracket "(" or ")" indicates exclusive.
    - (0,0.3]
    - [-0,3,0]
    - (0.3,0.6]
    - [-0.6, -0.3)
    - (0.6,1]
    - [-1, -0.6)
    - >1
    - <-1
3. Compute a series of correlation coefficients, including the on-grade and the original item parameter estimates.
4. Evaluate item parameter drift by applying the Robust Z method (Huynh & Rawls, 2011), a statistical hypothesis test used in large-scale assessments to detect items that have significantly drifted from the underlying scale. For this study, the Robust Z test was two-sided with 0.1 level of significance.

Items aligned to multiple target grades were calibrated using each on-grade calibration sample and therefore had multiple on-grade item difficulties. The number of these items was small and only found in the reading and science assessments. The analyses of these items were focused

[1] For a description of the standard MAP Growth item calibration procedure, please refer to the MAP Growth technical report (NWEA, 2019).

on how item parameter estimates (denoted as $\hat{b}_{lowgrade}$ and $\hat{b}_{upgrade}$) from the different target grade calibration samples were comparable with each other and with their original all-grade item difficulties.

2.2.2. Study 2: Item Calibration with Target + Adjacent Grades

Items in Study 2 also underwent the standard MAP Growth item calibration procedure but with responses from students in the target grade and the two adjacent grades, one below and one above. For example, the item difficulty estimate (denoted as $\hat{b}_{3grades}$) for an item aligned to Grade 4 was obtained using responses from students in Grades 3–5. The analysis was focused on comparing $\hat{b}_{3grades}$ with their on-grade and original all-grade counterparts.

3. Results

3.1. Study 1: On-Grade Item Calibration

Table 3.1 presents the summary statistics of the on-grade item parameter estimates and the differences between the on-grade and the all-grade item parameter estimates across subjects and grades. The average differences between the on-grade and all-grade item parameter estimates ($\hat{b}_{on\_grade} - \hat{b}_{base}$) are less than 0.1 logit (i.e., 1 RIT). The average on-grade item response counts ranged between 393 for reading and 643 for math, which accounted for the 19%–27% of the responses used to derive the original item difficulties.

Table 3.1. Summary Statistics of Item Parameter Estimates—Study 1

		$\hat{b}_{on\_grade}$		$\hat{b}_{base}$		$\hat{b}_{on\_grade} - \hat{b}_{base}$		*Prop*(Nongrade/Ntotal)		$N_{ongrade}$	
Grade	N_{item}	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Math											
1	1	-5.86	–	-6.00	–	0.14	–	0.54	–	1,277	–
2	208	-2.10	1.41	-2.02	1.38	-0.09	0.34	0.30	0.14	850	448
3	104	-0.34	2.01	-0.29	2.01	-0.05	0.32	0.22	0.08	540	236
4	96	1.45	1.78	1.52	1.79	-0.07	0.28	0.24	0.08	604	236
5	175	3.23	1.77	3.32	1.87	-0.09	0.26	0.29	0.15	631	380
6	98	3.16	2.01	3.18	2.16	-0.02	0.43	0.27	0.10	583	262
7	185	4.13	2.24	4.20	2.32	-0.08	0.23	0.25	0.05	595	209
8	163	4.46	2.02	4.52	2.00	-0.06	0.18	0.28	0.06	586	183
9	10	5.45	2.94	5.58	2.72	-0.13	0.29	0.20	0.06	329	121
Overall	1,040	2.00	3.14	2.07	3.16	-0.07	0.29	0.27	0.11	643	326
Reading											
2	6	-0.19	1.39	-0.23	1.35	0.04	0.08	0.27	0.03	659	462
3	249	0.66	1.07	0.70	1.07	-0.04	0.20	0.23	0.03	451	78
4	280	1.46	1.38	1.46	1.30	0.00	0.22	0.19	0.02	396	71
5	160	2.04	1.31	2.06	1.29	-0.02	0.14	0.20	0.04	415	160
6	208	2.17	1.15	2.17	1.09	0.00	0.18	0.15	0.06	313	125
7	212	2.77	1.23	2.77	1.21	0.00	0.16	0.19	0.07	425	423
8	190	3.17	1.17	3.15	1.14	0.02	0.14	0.22	0.06	441	130
9	90	3.65	0.97	3.59	0.94	0.05	0.15	0.12	0.02	248	36
10	43	3.94	0.81	3.84	0.79	0.10	0.20	0.10	0.01	228	24
11	5	4.00	1.27	3.94	1.31	0.06	0.26	0.07	0.01	471	197
Overall	1,443	2.12	1.54	2.12	1.50	0.00	0.18	0.19	0.06	393	202

Grade	N_{item}	$\hat{b}_{on\_grade}$		$\hat{b}_{base}$		$\hat{b}_{on\_grade} - \hat{b}_{base}$		$Prop_{(Nongrade/Ntotal)}$		$N_{ongrade}$	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Science											
3	15	0.54	1.29	0.57	1.19	-0.02	0.22	0.16	0.05	376	144
4	7	0.19	1.05	0.19	1.03	0.01	0.11	0.24	0.03	530	152
5	12	1.09	1.71	1.15	1.76	-0.06	0.13	0.27	0.05	656	121
6	46	1.52	1.25	1.62	1.29	-0.10	0.18	0.20	0.06	500	186
7	47	1.68	1.36	1.68	1.34	0.00	0.16	0.24	0.06	577	180
8	45	1.69	1.39	1.68	1.37	0.00	0.12	0.23	0.06	566	180
9	22	3.70	2.39	3.67	2.41	0.02	0.10	0.28	0.09	785	322
10	20	3.88	2.42	3.87	2.44	0.01	0.13	0.25	0.08	768	358
11	16	4.46	2.31	4.46	2.35	0.00	0.15	0.14	0.05	439	206
12	3	6.70	0.91	6.63	0.92	0.06	0.13	0.06	0.02	285	51
Overall	233	2.14	2.07	2.16	2.06	-0.02	0.15	0.22	0.08	572	241

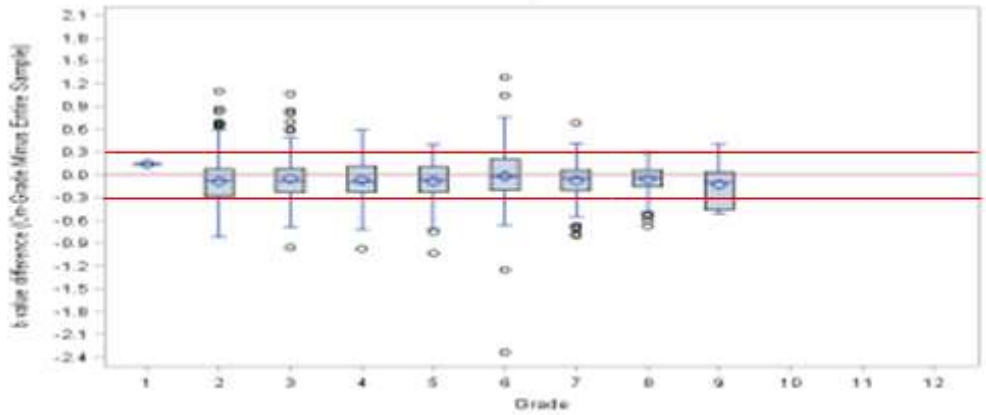
Note. $\hat{b}_{on\_grade} - \hat{b}_{base}$ indicates the difference between the on-grade and the original all-grade item parameter estimates. $Prop_{(Nongrade/Ntotal)}$ indicates the proportion of the on-grade response count over the total response count. $N_{ongrade}$ indicates the on-grade response count. "Overall" indicates the statistics for all items in a subject.

3.1.1. On-Grade vs. Original Item Parameter Estimates

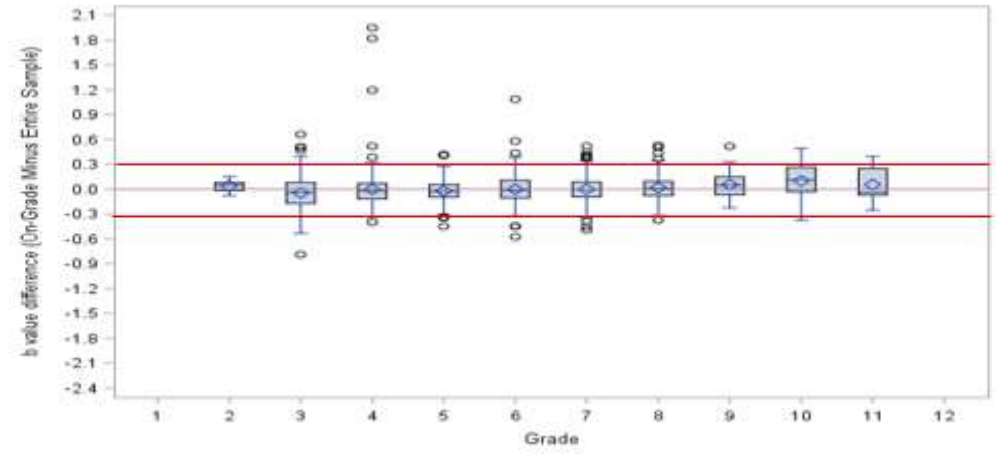
Figure 3.1 presents box plots of the differences between the on-grade and the original all-grade item parameter estimates across subjects and grades. The differences corresponding to the 25th and 75th percentiles were within the "0.3 Logit Difference" (Miller et al., 2004) band for all grades except Grade 9 math, which only contains 10 items. In large-scale assessment programs using the Rasch model, the "0.3 Logit Difference" rule is widely used to identify items that have significant item difficulty estimate differences (Huynh & Rawls, 2011). If the difference is beyond 0.3 logits, that item can be viewed as potentially unstable. In comparison with science, both math and reading had a few items with substantial item parameter estimate differences exceeding 1 logit (i.e., 10 RITs).

Figure 3.1. Box Plots of Differences between the On-Grade and Original Item Difficulty Parameters

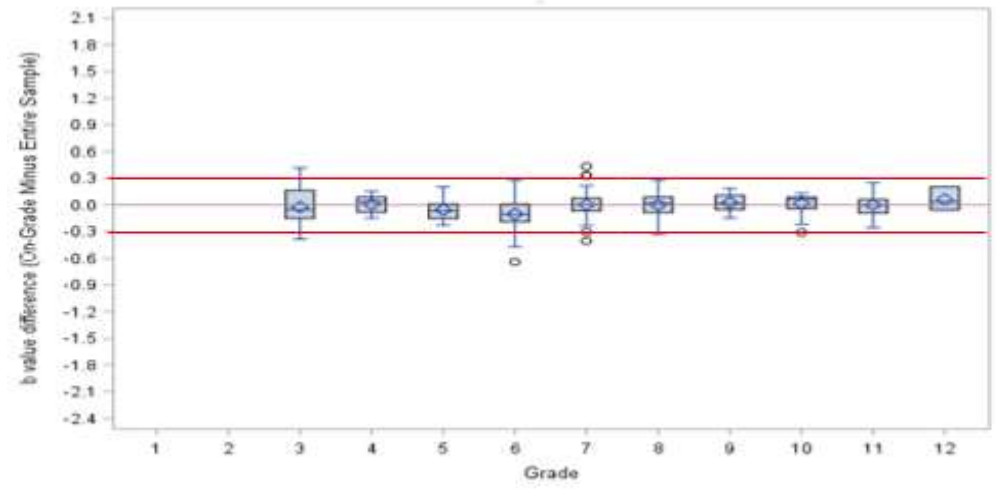
Math



Reading



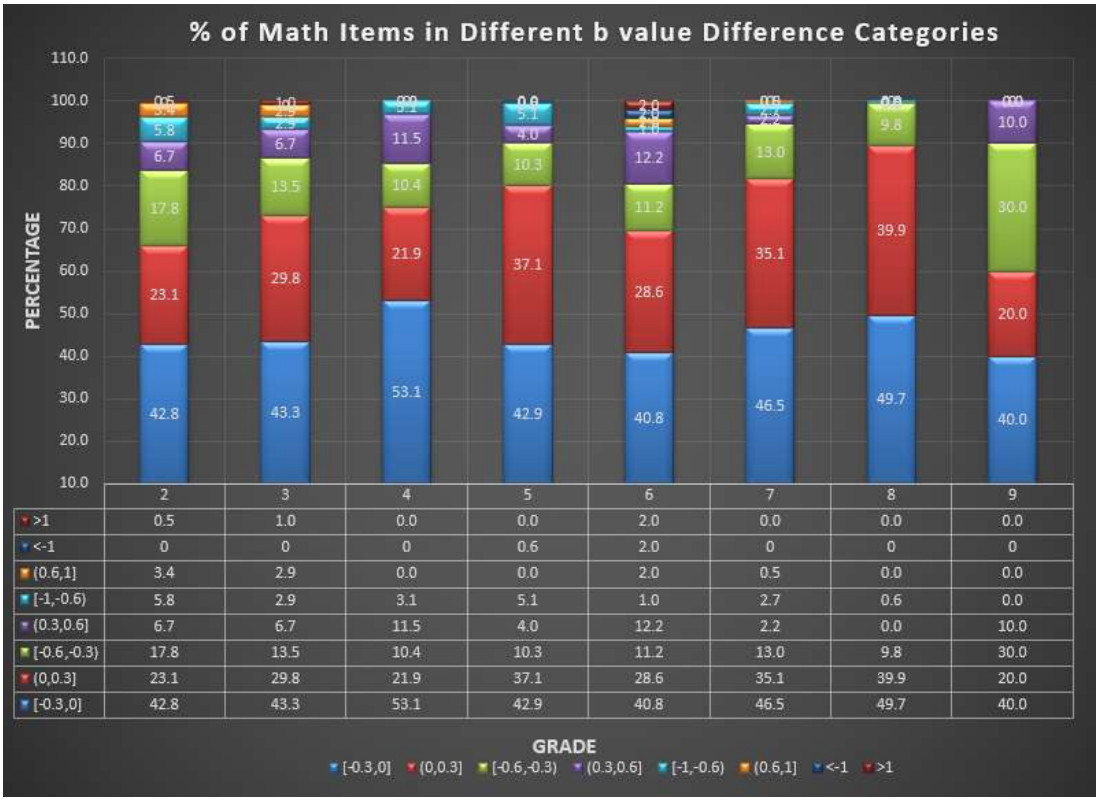
Science



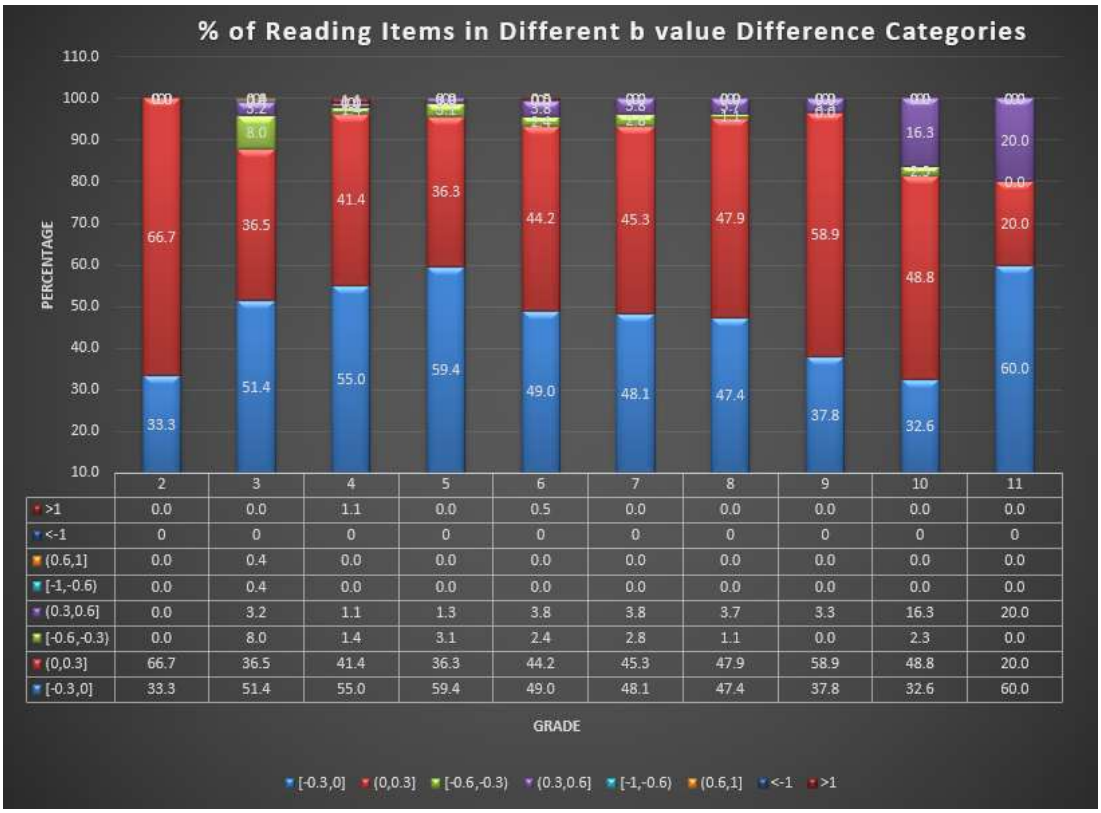
3.1.2. Difficulty Parameter Difference Categories

Figure 3.2 presents the percentage of items falling into the different difficulty parameter difference categories by subject and grade based on $\hat{b}_{on\_grade} - \hat{b}_{base}$. The red and blue bars, which indicate the 0.3 logit difference categories, show that at least 90% of reading and science items had differences within 0.3 logits for almost all grades. Math tended to see fewer items, about 80% of items for most grades, with differences within 0.3 logits. Overall, these results demonstrate that the item parameter estimates from both the on-grade and original all-grade samples are comparable.

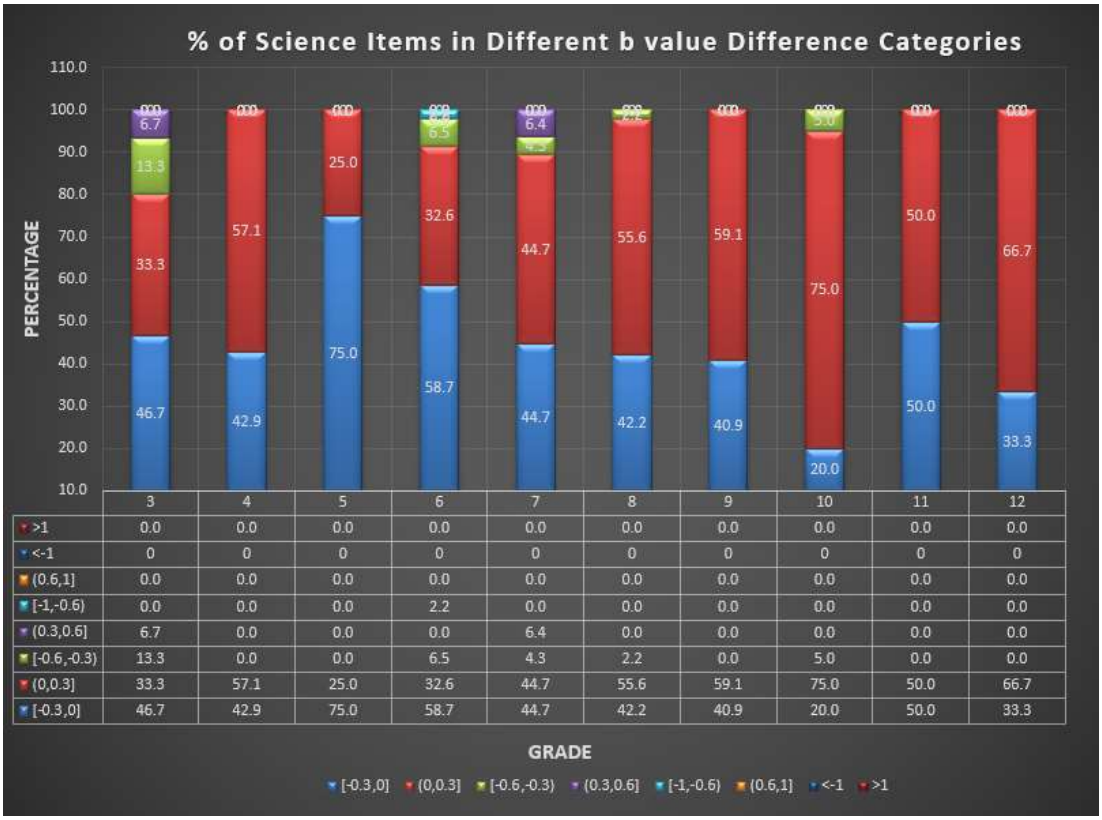
Figure 3.2. Percentage of Items in Different Difficulty Parameter Difference Categories



	2	3	4	5	6	7	8	9
>1	0.5	1.0	0.0	0.0	2.0	0.0	0.0	0.0
<-1	0	0	0	0.6	2.0	0	0	0
(0.6,1]	3.4	2.9	0.0	0.0	2.0	0.5	0.0	0.0
[-1,-0.6)	5.8	2.9	3.1	5.1	1.0	2.7	0.6	0.0
(0.3,0.6]	6.7	6.7	11.5	4.0	12.2	2.2	0.0	10.0
[-0.6,-0.3)	17.8	13.5	10.4	10.3	11.2	13.0	9.8	30.0
(0,0.3]	23.1	29.8	21.9	37.1	28.6	35.1	39.9	20.0
[-0.3,0]	42.8	43.3	53.1	42.9	40.8	46.5	49.7	40.0



	2	3	4	5	6	7	8	9	10	11
>1	0.0	0.0	1.1	0.0	0.5	0.0	0.0	0.0	0.0	0.0
<-1	0	0	0	0	0	0	0	0	0	0
(0.6,1]	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
[-1,-0.6)	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(0.3,0.6]	0.0	3.2	1.1	1.3	3.8	3.8	3.7	3.3	16.3	20.0
[-0.6,-0.3)	0.0	8.0	1.4	3.1	2.4	2.8	1.1	0.0	2.3	0.0
(0,0.3]	66.7	36.5	41.4	36.3	44.2	45.3	47.9	58.9	48.8	20.0
[-0.3,0]	33.3	51.4	55.0	59.4	49.0	48.1	47.4	37.8	32.6	60.0



3.1.3. Correlations

As shown in Table 3.2, the on-grade item parameter estimates are highly correlated with their original counterparts, with correlation coefficients ($r_{b\_base,\ b\_ongrade}$) close to 1 for most grades. The correlations between the calibration sample size difference and item parameter estimate difference ($r_{Ndiff,\ bdiff}$) are small for most grades, implying that the calibration sample sizes used to derive the on-grade and the original item difficulties did not affect the difference between the on-grade and the original item difficulty estimates.

Table 3.2. Correlation Coefficients Related to Parameter Estimates

Grade	N_{item}	$r_{b\_base,\ b\_ongrade}$	$r_{Ndiff,\ bdiff}$	$r_{(prop,\ bdiff)}$
Math				
2	208	0.97	0.27	0.02
3	104	0.99	-0.07	-0.08
4	96	0.99	0.15	0.37
5	175	0.99	-0.01	0.16
6	98	0.98	0.46	0.29
7	185	1.00	0.38	0.31
8	163	1.00	0.11	0.13
9	10	1.00	0.86	0.20
Overall	1,039	1.00	0.23	0.12

Grade	N_{item}	$r_{b\_base,\ b\_ongrade}$	$r_{Ndiff,\ bdiff}$	$r_{(prop,\ bdiff)}$
Reading				
2	6	1.00	0.37	-0.21
3	249	0.98	0.20	0.15
4	280	0.99	0.07	-0.03
5	160	0.99	0.01	0.00
6	208	0.99	-0.05	-0.16
7	212	0.99	-0.04	-0.11
8	190	0.99	-0.11	-0.20
9	90	0.99	-0.05	0.04
10	43	0.97	-0.14	-0.06
11	5	0.98	0.55	-0.48
Overall	1,443	0.99	-0.03	-0.12
Science				
3	15	0.99	0.62	0.10
4	7	0.99	0.04	-0.17
5	12	1.00	-0.34	-0.08
6	46	0.99	0.19	0.31
7	47	0.99	0.11	0.02
8	45	1.00	-0.17	-0.04
9	22	1.00	0.15	-0.42
10	20	1.00	0.36	0.32
11	16	1.00	0.02	0.08
12	3	0.99	-0.44	-0.41
Overall	233	1.00	0.05	0.08

Note. $r_{b\_base,\ b\_ongrade}$ indicates the correlation between the on-grade and original all-grade item parameter estimates. $r_{Ndiff,\ bdiff}$ indicates the correlation between the calibration sample size difference and the item parameter estimate difference. "Overall" indicates the statistics for all items in a subject.

3.1.4. Robust Z Statistics

As shown in Table 3.3, 13%–15% of items in each subject were found by the Robust Z statistics to be significantly drifted away from their underlying scales.[2] To explore which items tended to be flagged as unstable, descriptive statistics of both the flagged and unflagged items show that the average differences in item difficulty estimates for the flagged group were not that different from those of the unflagged group. For example, math had the largest difference. The magnitude of the average difference was 0.15 logit for the flagged math items but was 0.05 logit for the unflagged items. Figure 3.3 also suggests that there was no clear pattern as to which items tended to be flagged, as flagged items in each subject spread out over their entire underlying scale in a similar manner as the unflagged items.

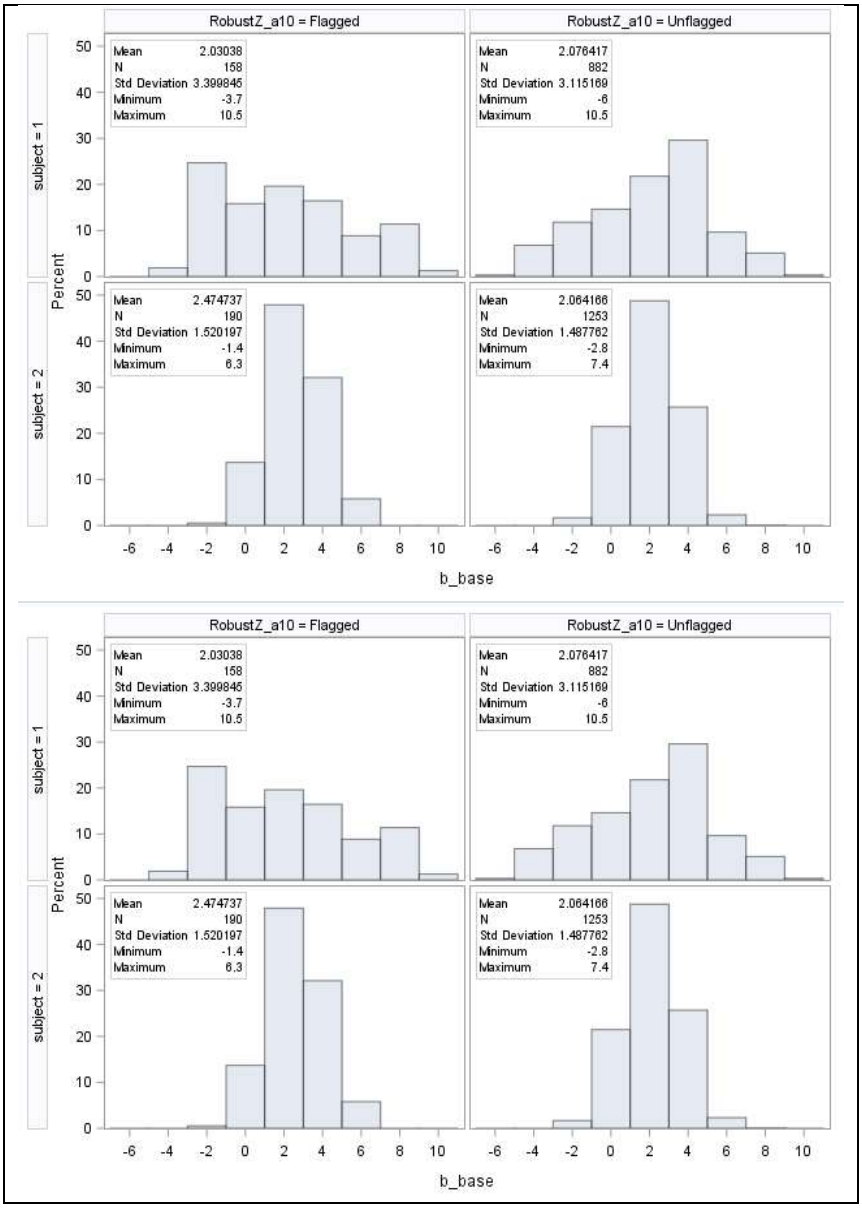
[2] These percentages are similar to what was reported in He et al. (2016), which found that 19%, 15%, and 25% of MAP Growth items in math, reading, and science, respectively, were flagged by the Robust Z procedure as unstable.

Table 3.3. Number of Flagged and Unflagged Items by Robust Z Procedure and their Summary Descriptive Statistics

	Robust Z			$\hat{b}_{on\_grade} - \hat{b}_{base}$				*Prop*(Nongrade/Ntotal)				$N_{on\text{-}grade}$			
Subject	**Status**	N_{item}	$\%_{item}$	**Mean**	**SD**	**Min.**	**Max.**	**Mean**	**SD**	**Min.**	**Max.**	**Mean**	**SD**	**Min.**	**Max.**
Math	Unflagged	882	84.8	-0.05	0.18	-0.44	0.32	0.28	0.11	0.06	0.75	666	332	205	2089
	Flagged	158	15.2	-0.15	0.61	-2.34	1.28	0.21	0.08	0.07	0.45	516	263	200	1636
Reading	Unflagged	1,253	86.8	-0.01	0.11	-0.25	0.23	0.19	0.06	0.07	0.36	400	211	200	6201
	Flagged	190	13.2	0.08	0.40	-0.79	1.95	0.17	0.05	0.05	0.33	347	119	201	1058
Science	Unflagged	201	86.3	-0.01	0.10	-0.23	0.20	0.23	0.07	0.04	0.44	598	240	200	1628
	Flagged	32	13.7	-0.05	0.32	-0.64	0.43	0.15	0.05	0.09	0.25	411	186	202	867

Note. Prop(Nongrade/Ntotal) indicates the proportion of on-grade item responses over the overall item responses. $N_{on\text{-}grade}$ indicates the on-grade calibration sample size. The % of items is out of the total number of items in the calibration sample for each subject across all grades.

Figure 3.3. Item Difficulty Distributions of Flagged and Unflagged Items by Robust Z Procedure



3.1.5. Items with Multiple Target Grades

Items with multiple target grades included 43 reading items aligned to Grades 9 and 10 and 67 science items distributed in different grade spans including Grades 6–7, 6–8, 7–8, 9–10, 9–11, and 9–12. As shown in Table 3.4, while the average lower and upper on-grade item difficulty parameters are slightly different from their original item difficulty parameters for both subjects, the differences are considered negligible, with 0.1 logit at most.

The histograms in Figure 3.4 and Figure 3.5, with the x-axis indicating the magnitude of different difficulty parameter differences, further suggest that, for most items, the item difficulty estimates using different on-grade calibration samples are comparable with each other and with their original item difficulty estimates. The correlation coefficients between different parameter estimates were high as well, as shown in Table 3.5.

Table 3.4. Summary Statistics of Difficulty Estimates for Items Aligned to Multiple Target Grades

	N_{item}	Mean	SD	Min.	Max.
Reading					
$\hat{b}_{lowgrade}$	43	3.92	0.82	1.68	5.97
$\hat{b}_{upgrade}$	43	3.94	0.81	1.84	5.66
$\hat{b}_{base}$	43	3.84	0.79	1.60	5.70
$\hat{b}_{lowgrade} - \hat{b}_{base}$	43	0.07	0.12	-0.23	0.33
$\hat{b}_{upgrade} - \hat{b}_{base}$	43	0.10	0.20	-0.37	0.49
$\hat{b}_{upgrade} - \hat{b}_{lowgrade}$	43	0.03	0.23	-0.39	0.54
Science					
$\hat{b}_{lowgrade}$	67	2.27	2.01	-0.66	7.86
$\hat{b}_{upgrade}$	67	2.34	1.99	-0.76	7.74
$\hat{b}_{base}$	67	2.33	2.00	-0.60	7.70
$\hat{b}_{lowgrade} - \hat{b}_{base}$	67	-0.06	0.17	-0.64	0.43
$\hat{b}_{upgrade} - \hat{b}_{base}$	67	0.00	0.13	-0.33	0.33
$\hat{b}_{upgrade} - \hat{b}_{lowgrade}$	67	0.07	0.22	-0.44	0.56

Table 3.5. Correlation Coefficients between Different Difficulty Estimates for Items Aligned to Multiple Target Grades

Subject	N_{item}	$r_{(b_{lower},b_{base})}$	$r_{(b_{lower},b_{up})}$	$r_{(b_{up},b_{base})}$
Reading	43	0.99	0.96	0.97
Science	67	1.00	0.99	1.00

Figure 3.4. Histograms of Difficulty Differences for Items Aligned to Multiple Target Grades—Reading

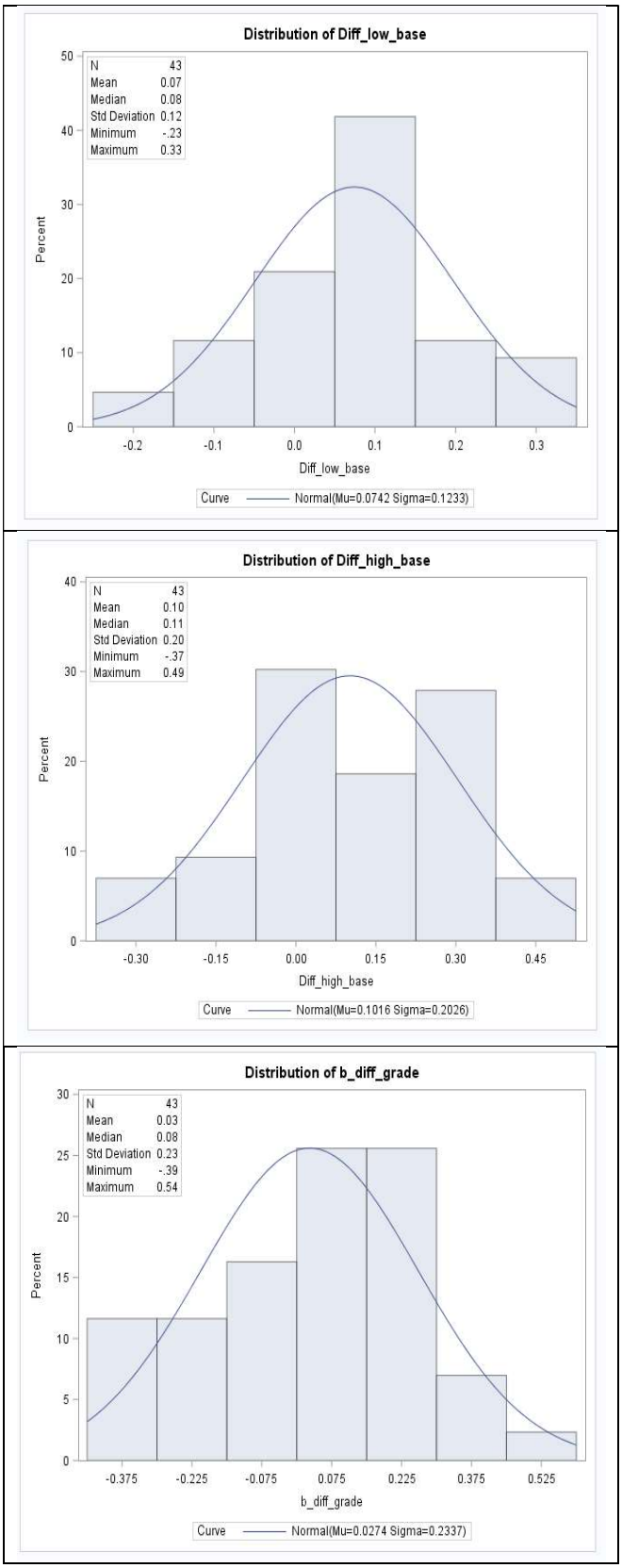
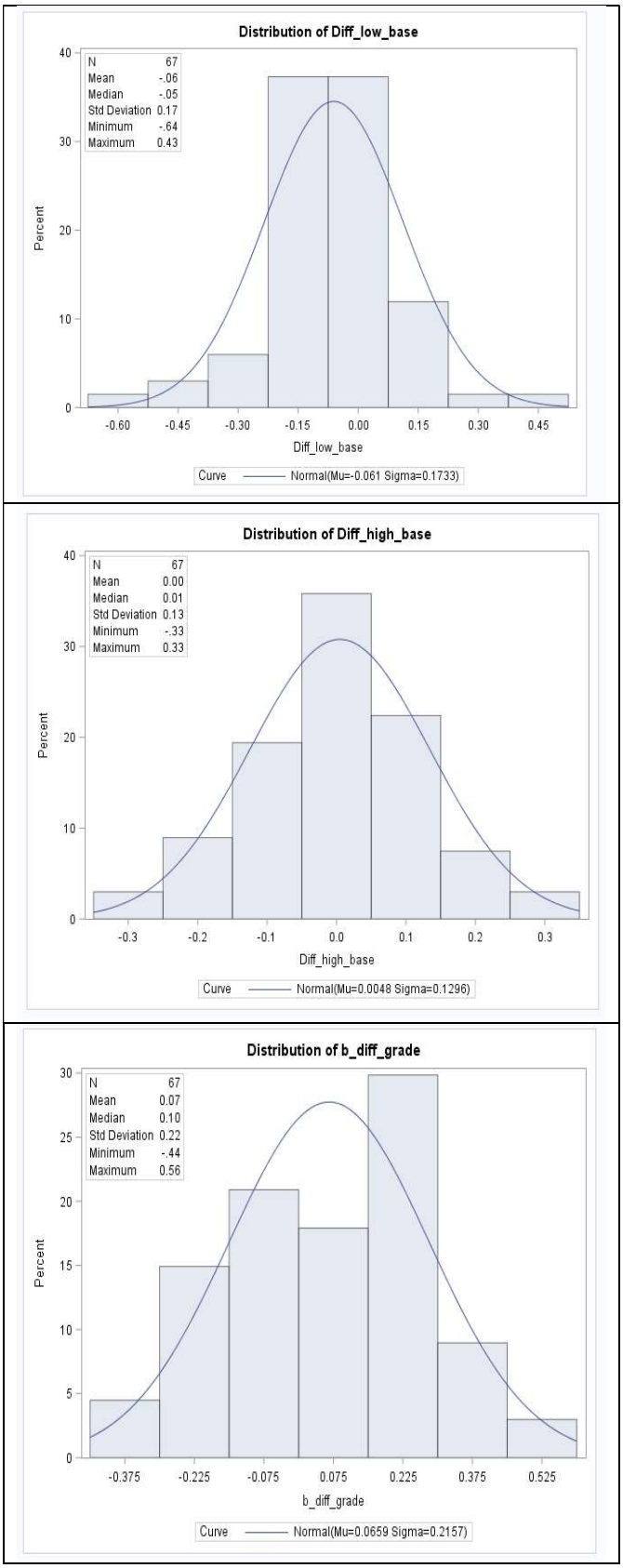


Figure 3.5. Histograms of Difficulty Differences for Items Aligned to Multiple Target Grades—Science



3.2. Study 2: Item Calibration with Target + Adjacent Grades

Table 3.6 presents a series of summary statistics, including the differences between the original parameter estimates with those from the Study 2 samples and with the on-grade samples, as well as the differences in calibration samples in the form of proportion. The mean difference columns indicate that, in general, item parameter estimate differences between $\hat{b}_{3grades}$ and $\hat{b}_{base}$ are no different from the difference between $\hat{b}_{ongrade}$ and $\hat{b}_{base}$. The magnitude of the largest average difference is 0.05 logit in Grade 7 math. As expected, using responses from three grades tended to increase the calibration sample size by at least two times for almost all grades.

Table 3.6. Summary Statistics of Calibration Samples—Study 2

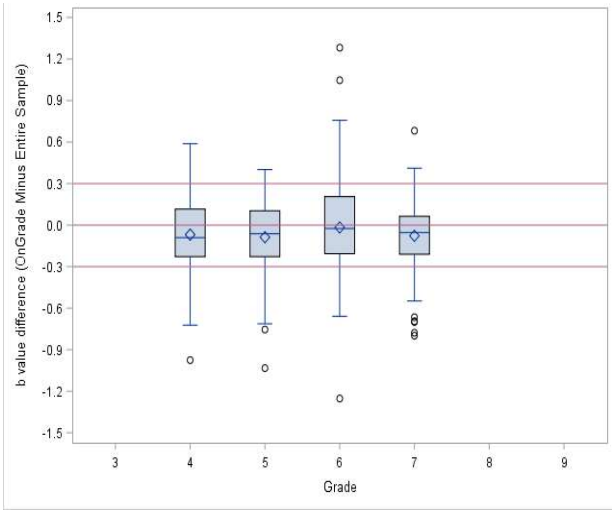
Grade	N_{item}	Mean Difference		Proportion		N_{all}
		$Mean(\hat{b}_{3grades} - \hat{b}_{base})$	$Mean(\hat{b}_{ongrade} - \hat{b}_{base})$	$N_{ongrade}/N_{all}$	N_{3grade}/N_{all}	
Math						
4	96	-0.05	-0.07	0.24	0.71	2,471
5	175	-0.11	-0.09	0.29	0.55	2,098
6	98	-0.06	-0.02	0.27	0.53	2,157
7	185	-0.03	-0.08	0.25	0.75	2,402
Overall	554	-0.06	-0.07	0.27	0.64	2,274
Reading						
4	280	-0.04	0.00	0.19	0.56	2,033
5	160	-0.02	-0.02	0.20	0.51	2,089
6	208	-0.02	0.00	0.15	0.47	2,134
7	212	0.01	0.00	0.19	0.58	2,248
Overall	860	-0.02	0.00	0.18	0.53	2,121
Science						
4	7	0.01	0.01	0.24	0.70	2,235
5	12	-0.04	-0.06	0.27	0.54	2,474
6	46	-0.05	-0.10	0.20	0.48	2,472
7	47	-0.02	0.00	0.24	0.66	2,475
Overall	112	-0.03	-0.05	0.23	0.58	2,458

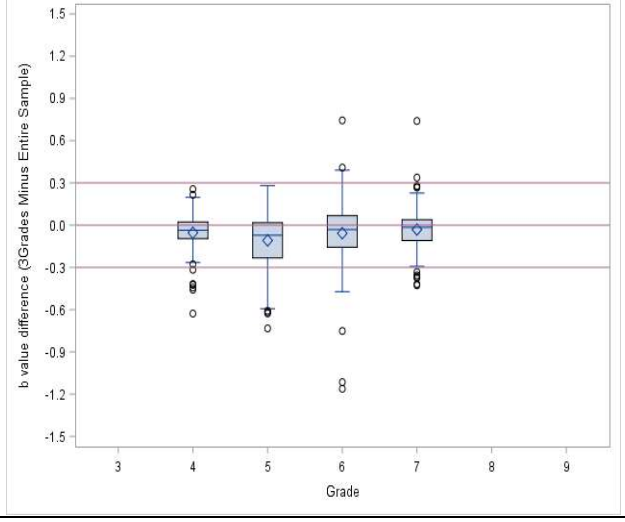
Note. "Overall" indicates the statistics for all items in a subject.

Figure 3.6 presents box plots of the differences between the original difficulty estimates and the on-grade difficulty estimates (i.e., figures on the left panel) and of the differences between the original difficulty estimates and the difficulty estimates from three grades (i.e., figures on the right panel) by subject and grade. The differences of item difficulty estimates for most items were within 0.3 logit. A few math and reading items had parameter estimate differences larger than 1 logit.

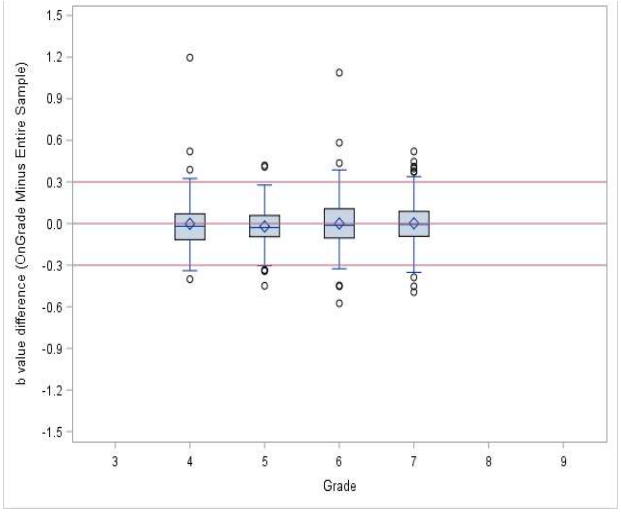
Figure 3.6. Box Plots of Item Difficulty Differences between b_{base} and $b_{ongrade}$/$b_{3grades}$

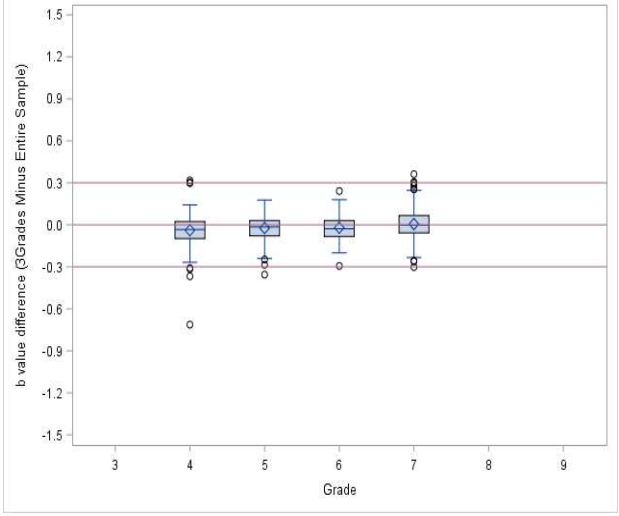
Mathematics



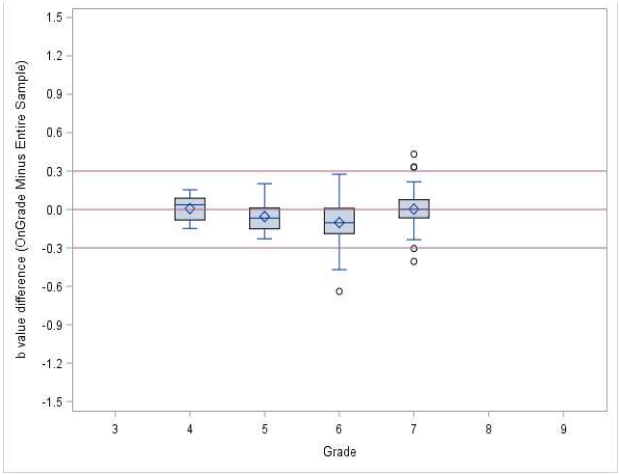


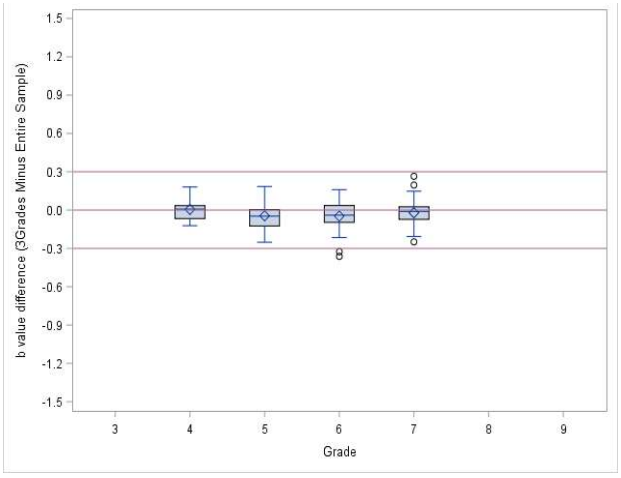
Reading





Science





4. Conclusion and Discussion

Parameter invariance is a fundamental assumption underlying the application of IRT models. It refers to population rather than sample quantities and occurs when items exhibit the same parameter estimates across subgroups under the same IRT model. For MAP Growth items used for students in different grades, the parameter invariance requires that items perform comparably for students across grades, which can be viewed as a broader context under which this research study was conducted.

The study compared the item parameter estimates obtained from the original all-grade samples with those from more focused samples, including on-grade only and on-grade plus two adjacent grades. Results of this study indicate that the average parameter estimates across all calibration samples (i.e., original all-grade, target grade, and target + two adjacent grades) are almost no different from each other. The parameter estimates of most items are comparable as well. This result is resonant with the finding by Wan and Thum (2021) who used differential item functioning (DIF) analyses to reveal that MAP Growth items perform comparably across states and grades. Both findings provide quantitative evidence of the invariance of MAP Growth item parameter estimates.

Item parameter estimate accuracy can be affected by various factors, and calibration sample size is just one of them. By exploring the effects of mixture distribution of calibration samples, Wang (2011) suggests that item parameter estimate accuracy can be affected by various factors, some of which may not always be known to practitioners. Wang (2011) further recommended that, to mitigate the effects of those unknown factors on item parameter estimate accuracy, it is always preferred to use a large sample for item calibration if possible. In the standard MAP Growth item calibration procedure, the minimum sample size is set to be 1,000 and the average calibration sample size for an item tends to be at least 2,000. This is expected to have contributed to what the study has found about the invariance of MAP Growth item parameter estimates from the different calibration samples.

Overall, the study finding of the invariance of item parameter estimates suggests that the existing item parameter estimates are still appropriate if used in the new MAP Growth assessments that intend to administer items more closely aligned to the grade level of the student. Items with different parameter estimates using different calibration samples will go through further content and psychometric review to understand the possible reasons for the differences. More items are also planned to be included in the future to see whether the same findings still hold.

5. References

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.

He, W., Li, S., & Kingsbury, G. (2016). *A large-scale, long-term study of scale drift: The micro view and the macro view*. Paper presented at the IMEKO conference, Berkeley, CA.

He, W. (2015). CAT field-test item calibration sample size: How large is large under the Rasch model? *Global Journal of Human-Social Science, 15(*1). https://globaljournals.org/GJHSS_Volume15/8-CAT-Field-Test-Item-Calibration.pdf.

Huynh, H., & Rawls, A. (2011). A comparison between Robust Z and 0.3-Logit Difference procedures in assessing stability of linking items for the Rasch model. *Journal of Applied Measurement, 12*(2), 96–105.

Miller, G. G., Rotou, O., & Twing, J. S. (2004). Evaluation of the .3 logits screening criterion in common item equating. *Journal of Applied Measurement*, *5*(2), 172–177.

National Governors Association Center for Best Practices & Council of Chief State School Officers (CCSSO). (2010). *Common core state standards*. http://www.corestandards.org/read-the-standards/

NWEA. (2019). *MAP Growth technical report*.

Wan, S., & Thum, Y. (2021). *Investigating states and grades DIF of items in a CAT test*. Paper presented at the 2021 NCME conference.

Wang, S. (2011). *The effects of mixture distribution of calibration sample on accuracy of Rasch item parameter estimation*. NWEA.